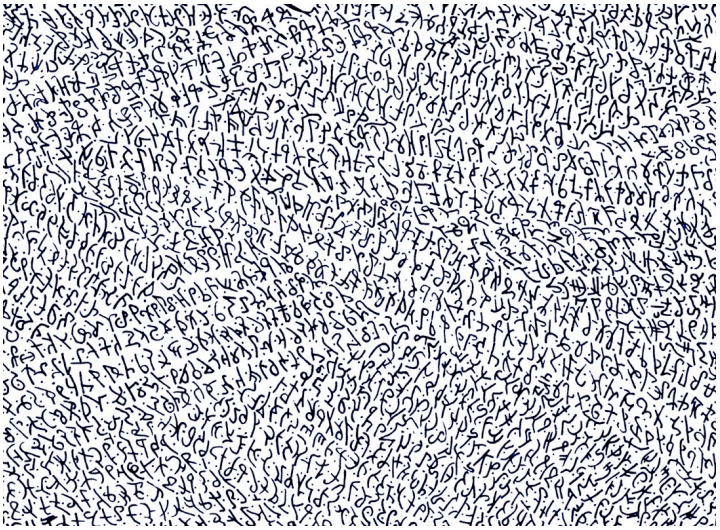
COLLOQUE SHESL-HTL 2015
http://shesl-htl2015.sciencesconf.org/

Corpus et constitution des savoirs linguistiques

RÉSUMÉS DES COMMUNICATIONS



E. Demont-Suisse apozepoa@gmail.com

LLL

HTL

S.H.E.S.L.

30 janvier 2015 :
INALCO/BULAC, amphithéâtre : 65, rue des Grands Moulins, 75013 Paris
31 janvier 2015 :
BNF, salle 70: Quai François Mauriac, 75013 Paris

LA CONSTITUTION D'UN CORPUS PARLE DU GEORDIE : CHOIX EPISTEMOLOGIQUES ET REALISATIONS EMPIRIQUES. RETOUR SUR UN DEMI-SIECLE DE SOCIOPHONETIQUE ANGLAISE

Maelle AMAND
Université Paris Diderot CLILLAC-ARP

« *En hommage au travail des dialectologues, à leur magnifique moisson de données, à leur expérience tout simplement humaine…* » (Meunier-Crespo, p. 12).

Cette communication présente d'un point de vue épistémologique et empirique le travail de constitution du corpus parlé de l'anglais du Tyneside ou *Diachronic Electronic Corpus of Tyneside English* dit DECTE (Corrigan *et al*. 2001) qui consigne près de cinquante ans de recherche en linguistique de corpus. Nous traiterons de sa genèse puis de son étoffement en corpus diachronique ainsi que de son exploitabilité grâce aux outils informatiques actuels. Cette analyse sera replacée dans le contexte de la sociophonétique de corpus de l'école anglaise (Milroy 1992 & Firth 1946).

À la fin des années soixante, lorsque l'enquête linguistique de la région du Tyneside (Tyneside Linguistique Survey ou TLS) fut mise en place, deux très grandes enquêtes venaient de s'achever : la première en Grande-Bretagne, et la seconde, de l'autre côté de l'Atlantique, concernant l'anglais américain. Or si Labov recommandait aux sociophonéticiens de se limiter aux variables répondant à des exigences théoriques spécifiques, le traitement de la TLS avait pour but de proposer le plus de variables possibles (de l'ordre du millier par locuteur) afin de proposer des données exploitables de la façon la plus empirique possible par le biais des outils informatiques naissant (Pellowe *et al*. 1972, Jones-Sargent 1983). Ce sont tout d'abord les outils et méthodes de la TLS, très novatrices et peut-être trop ambitieuses pour un parc informatique de première génération, que nous allons étudier ici. En somme, dans quelle mesure les limites technologiques ont orienté la TLS vers un système de transcription peu conventionnel et quel est le degré de transférabilité des données vers des outils de mesure contemporains afin de poursuivre une recherche en adéquation avec les conventions des années 2000 ? Nous éclaircirons la réflexion ayant guidé le recueil des données par la première équipe de chercheurs en vue d'analyses relevant de domaines diversifiés – prosodie, phonétique, sociophonétique et interaction, analyses en cluster et multimodales (Moisl 2005). Nous aborderons les problèmes liés à l'enregistrement l'archivage et à la numérisation de ce corpus à l'aide d'exemples concrets. La question de la transcription est également centrale au niveau du TLS et suscitera une réflexion sur le niveau de transcription nécessaire à une description fine du texte sans rendre le nombre de variables impossibles à gérer (Autesserre 1989, Kerswil 1990). Nous verrons enfin comment exploiter tous les atouts de la linguistique de corpus et de la phonétique actuelle, afin de mettre en lumière des résultats du point de vue segmental. Nous montrerons ainsi l'évolution des techniques et des méthodes pour l'observation du système vocalique du Geordie, qui monophtongue et centralise grand nombre de ses voyelles. Nous confronterons ce demi-siècle d'analyse sociophonétique aux logiciels linguistiques et statistiques contemporains, tels que Praat, WinPitch, SPPAS (Bigi 2012), R et de NORM (Thomas & Kendall 2007).

Mots-clés
LINGUISTIQUE DE CORPUS, CORPUS PARLE, DECTE, DIALECTOLOGIE, SPPAS, EVOLUTION DES METHODOLOGIES EN SOCIOPHONETIQUE

Bibliographie

AUTESSERRE, Denis, PERENNOU, Guy et ROSSI, Mario. 1989. « Methodology for the Transcription and Labelling of a Speech Corpus », *Journal of the International Phonetic Association*, 19/01, p. 2.

BEAL, Joan, CORRIGAN, Karren et MOISL, Herman, éd. 2007. *Creating and Digitizing Language Corpora*, Palgrave Macmillan éd., vol. I, Basingstoke.

BIGI, Brigitte. 2012. « SPPAS: a tool for the phonetic segmentation of speech », *in* Proceedings of the Language Resource and Evaluation Conference, Istanbul, Turquie, p. 1748-1755.

CAMERON, Deborah. 2001. *Working with Spoken Discourse*, London, SAGE.

FIRTH, John Rupert. 1946. « The English School of Phonetics », *Transactions of the Philological Society*, 45/01, p. 92-132.

JONES-SARGENT, Val. 1983. *Tyne Bytes. A Computerised Sociolinguistic Study of Tyneside*, p. 368.

KERSWILL, Paul et WRIGHT, Susan. 1990. « The Validity of Phonetic Transcription: Limitations of a Sociolinguistic Research Tool », *Language Variation and Change 2*, p. 255-275.

MILROY, James. 1992. *Linguistic Variation and Change: On the Historical Sociolinguistics of English*, Oxford, Blackwell.

MEUNIER-CRESPO, Mariette. 2009. « La Constitution D'un Corpus Oral, Parcours Initiatique En Linguistique », halshs-00359903.

MOISL, Herman et JONES, Val. 2005. « Cluster Analysis of the Newcastle Electronic Corpus of Tyneside English: A Comparison of Methods », *Literary and Linguistic Computing*, 20, p. 125-146.

PELLOWE, John *et al.* 1972. « A Dynamic Modelling of Linguistic Variation: The Urban (Tyneside) Linguistic Survey », *Lingua*, 30, p. 1-30.

THOMAS, Erik R. and KENDALL, Tyler. 2007. *NORM: The vowel normalization and plotting suite*. [Online Resource: http://ncslaap.lib.ncsu.edu/tools/norm/]

LE GRAND CORPUS DES GRAMMAIRES FRANÇAISES, DES REMARQUES ET DES TRAITES SUR LA LANGUE

Wendy AYRES-BENNETT
Université de Cambridge
Bernard COLOMBAT
UMR 7597, Université Paris Diderot / CNRS

En 2011 la première tranche du *Grand corpus des grammaires françaises, des remarques et des traités sur la langue (XIV^e^-XVII^e^ s.)* a paru chez Classiques Garnier Numérique ; elle comprend trois volets : Grammaires française de la Renaissance, Grammaires françaises du XVII^e^ siècle, et Remarques sur la langue française (XVII^e^ siècle). Cette première tranche contient 48 textes, soit 3,36 millions de mots et/ou environ 15000 pages. Fortement outillée et balisée selon des critères spécifiques, elle permet des recherches complexes dans des domaines comme l'histoire des théories linguistiques, l'histoire de la langue ou, plus généralement, l'histoire culturelle. Grâce à une licence nationale, le corpus est disponible (depuis avril 2014) dans toutes les bibliothèques publiques et universitaires en France.

Une deuxième tranche de 49 textes est actuellement en préparation qui ajoutera les grammaires du XVIII^e^ siècle et étoffera le corpus des volumes de remarques en ajoutant d'autres volumes publiés au XVII^e^ siècle et en étendant le corpus au XVIII^e^ siècle. Nous voudrions présenter les textes choisis pour cette deuxième tranche et considérer plusieurs questions soulevées par la sélection de textes à inclure :

- Dans quelle mesure un tel corpus peut-il (ou doit-il) être « représentatif » ? Que veut dire « représentativité » ? quels sont les critères de « représentativité » ? que veut dire « texte de référence » ?
- Dans le cas étudié, quels critères spécifiques sont les plus pertinents pour choisir les textes ? Par exemple, où sont les limites du genre des remarques sur la langue française ? Une grammaire générale d'expression française peut-elle être classée dans les grammaires françaises ? Comment choisir parmi les nombreuses grammaires de la tradition scolaire ce qui est le plus représentatif de ce type de texte métalinguistique ? En quoi le format de publication des textes (par ex. articles séparés publiés dans une encyclopédie *vs* texte autonome) peut-il influencer la perception qu'on peut avoir d'une œuvre ?
- Quels sont les effets de la *création* d'un tel corpus ? Il semble évident que la première tranche du *Grand corpus* permettra de renouveler la recherche dans l'histoire des idées linguistiques en France, en permettant aux chercheurs de faire des recherches quantitatives et beaucoup plus exhaustives qu'auparavant. Mais ne risque-t-on pas de créer une sorte de canon de textes qui seront beaucoup étudiés à cause des outils de recherche disponibles ? Et, à l'opposé, ne risque-t-on pas l'éviction ou la marginalisation de textes qui seront peu ou pas étudiés, malgré leur intérêt, parce qu'ils ne sont pas inclus dans le corpus de référence ?
- Quels sont les effets du choix de l'*outillage* d'un tel corpus ? Les choix de requêtes sont nécessairement faits en amont, avant l'exploration des textes : leur utilisation effective dans l'exploitation du corpus a-t-elle des effets sur la perception que l'on offre des textes ?

Mots-clés

CORPUS, GRAMMAIRES FRANÇAISES, REMARQUES SUR LA LANGUE FRANÇAISE, REPRESENTATIVITE, CANON

Éléments bibliographiques

COLOMBAT, Bernard, FOURNIER, Jean-Marie, AYRES-BENNETT, Wendy dir. 2011. *Grand Corpus des grammaires françaises, des remarques et des traités sur la langue (XIV^e^-XVII^e^ s.)*, Paris, Classiques Garnier Numérique.

DU DICTIONNAIRE LEXICO-PHONETISE AUX CORPUS ORAUX, QUELQUES PROBLEMES EPISTEMOLOGIQUES POUR L'ECOLE DE GUIERRE

Nicolas BALLIER
Univ. Paris Diderot Sorbonne Paris Cité, CLILLAC-ARP (EA3967)

Lionel Guierre a été, dans les années soixante-dix, l'un des pionniers d'une certaine forme de phonologie de corpus, en constituant une version électronique du dictionnaire de prononciation de l'anglais de Jones (dans sa douzième édition, celle de 1968). Il a systématisé et théorisé (Guierre 1979) l'interrogation des régularités accentuelles à partir des séquences graphématiques. Cette contribution se propose de revenir sur près de quarante ans de cette tradition d'analyse du placement accentuel de l'anglais, en exposant le déplacement de certaines problématiques, d'un questionnement de l'institutionnalisation de la variation à partir de sa consignation dans les dictionnaires de prononciation à son exploration dans les corpus oraux.

Dans un premier temps, on exposera quelques-uns des résultats de cette école théorique, en montrant l'importance quantitative des données lexicales étudiées. On examinera ensuite les conséquences théoriques d'une analyse qui procède d'une lecture tabulaire des données, telle qu'on peut l'établir aujourd'hui à partir de ces dictionnaires lexico-phonétisés, en expliquant les possibilités ouvertes par la comparaison avec des bases de données type CELEX (Baayen *et al.* 1995). En particulier, l'indication de la fréquence lexicale peut servir à esquisser des tropismes accentuels et des phénomènes de diffusion lexicale des variations, qui complètent certaines analyses des disciples de Guierre (Deschamps 2001, Fournier 2007, Trevian 2007). Au plan épistémologique, on montrera que ce type de représentation tabulaire des données milite en faveur d'une « phonologie plate ».

Dans un deuxième temps, on examinera les problèmes théoriques rencontrés par la deuxième génération de disciples de Guierre (Martin 2011, Videau 2013), qui s'efforcent de confronter les variantes institutionnalisées répertoriées dans les dictionnaires de prononciation à des analyses sous praat de réalisations phonétiques collectées. Deux ordres de problématiques seront envisagés. La question de l'échantillonnage du « corpus oral » phonétique à constituer pour observer la variabilité et surtout des phénomènes discursifs à considérer seront examinés. La prise en compte du cotexte et de la sémantique met en tension certains traits retenus par la lexicographie (par exemple, la possibilité d'un *stress shift,* notée pour certaines unités lexicales). Au-delà des problèmes posés par les modélisations statistiques de la fréquence (en particulier dans les modèles LNRE, cf. Baayen 2008), la question de la représentativité des *tokens* à considérer dans les corpus écrits est peut-être réglée par les giga-corpus (Bernardini, Baroni & Evert 2006), mais elle est balbutiante pour la représentation des variantes phonétiques (au-delà de nos rassurantes formes lemmatisées) dans les corpus oraux. A partir de l'exemple des préfixes (Videau 2013), on exposera ce deuxième aspect problématique, ce que l'on pourrait appeler la « lemmatisation du sublexical », la reconstitution *a priori* de l'ensemble des formes alternantes des unités inférieures au lexème et les problèmes que pose leur examen en corpus oral.

Mots-clés
PHONOLOGIE, TROISIEME REVOLUTION DE LA GRAMMATISATION, MODELES LNRE, CONSTITUANTS, ECHANTILLONNAGE

Bibliographie

BAAYEN, Harald, PIEPENBROCK, Richard, & GULIKERS, Leon. 1995. *The CELEX lexical database (release 2)*, distributed by the Linguistic Data Consortium, University of Pennsylvania.

BAAYEN, Harald. 2008. *Analyzing Linguistic Data: A practical introduction to statistics using R,* Cambridge, Cambridge University Press.

BERNARDINI, Silvia, BARONI Marco & EVERT, Stefan. 2006. « A WaCky Introduction". *Wacky! Working papers on the Web as Corpus*, Bologne, Gedit, p. 9-40.

DESCHAMPS, Alain. 2001. "Stress-patterns, rules and variants: Can stress variation be accounted for?", *Anglophonia 9/2001: Langues et littératures*, (9), p. 41-57.

FOURNIER, Jean- Michel. 2007. "From a Latin syllable-driven stress system to a Romance versus Germanic morphology-driven dynamics: in honour of Lionel Guierre", *Language Sciences*, 29(2), p. 218-236.

GUIERRE, Lionel. 1979. *Essai sur l'accentuation en anglais contemporain*, thèse d'État, Université de Paris VII.

MARTIN, Marjolaine. 2011. *De l'accentuation lexicale en anglais australien standard*, thèse non publiée, sous la direction de Jean-Michel Fournier, Université de Tours.

TREVIAN, Ives. 2007. "Stress-neutral endings in contemporary British English: an updated overview", *Language Sciences*, 29 (2), p. 426-450.

VIDEAU, Nicolas. 2013. *Préfixation et phonologie de l'anglais : analyse lexicographique, phonétique et acoustique*, thèse non publiée, sous la direction de Jean-Louis Duchet et de Sylvie Hanote, Université de Poitiers.

CORPUS ET SAVOIRS : DES LIAISONS HEUREUSES ?

Olivier BAUDE, Céline DUGUA
Laboratoire Ligérien de Linguistique UMR 7270, Université d'Orléans
Olivier.Baude@univ-orleans.fr ; Celine.Dugua@univ-orleans.fr

Dans sa présentation du livre de William Labov *Sociolinguistique*, Pierre Encrevé reprend les propositions de fondement d'une sociolinguistique qui se propose d'être *la linguistique, toute la linguistique mais la linguistique remise sur ses pieds*. Ces propositions accordaient une place centrale aux variations linguistiques et à la nécessité de les observer, de les décrire et de les analyser au sein d'une masse de données issues d'enquêtes sociolinguistiques.
Quarante ans après, on peut constater que les données organisées en corpus ont acquis un statut majeur en linguistique sans toutefois qu'elles correspondent toujours à la méthodologie et aux cadres théoriques énoncés par la sociolinguistique (notamment en développant uniquement l'aspect masse de données traitée par des outils informatiques). Plus récemment encore nous assistons au tournant épistémologique des « humanités numériques » qui créent de nouveaux objets et de nouvelles démarches scientifiques autour de l'interopérabilité des données numériques et des approches trans/interdisciplinaires.
Cette communication souhaite porter un regard sur un exemple concret de relations entre corpus et analyses linguistiques. Ainsi, la liaison est un phénomène linguistique dont une partie importante des analyses récentes repose sur l'étude de corpus. Ce fut le cas pour des travaux fameux : notamment Encrevé (1988), Laks (2005, 2012), Dejong (1994) mais aussi plus récemment dans le cadre de projets de grand corpus, tels que *Phonologie du Français Contemporain*. Ces différents travaux réalisés sur plusieurs décennies, permettent de poser la question de l'impact du recours aux corpus sur la constitution des savoirs linguistiques. Quelles sont les avancées produites et quels sont les échecs ? Quelles conséquences en tirer, tant du point de vue des données et de leur recueil que des connaissances théoriques sur la liaison ?
Ce sont ces questions que nous souhaitons aborder à partir de l'expérience du corpus des Enquêtes SocioLinguistiques à Orléans. Les ESLO se concrétisent par un double corpus de français parlé réalisé en 1968-1971 (ESLO1) et depuis 2008 (ESLO2). Ces deux corpus ont pour vocation de permettre des études linguistiques à partir de données situées. Les choix méthodologiques opérés à quarante années d'intervalle par les auteurs offrent un point de vue privilégié pour une étude épistémologique sur la constitution et l'exploitation de grands corpus. En revenant sur certains de ces choix, et en les soumettant à des données sur la liaison, nous nous proposons de questionner justement les liens entre les choix méthodologiques et les données qui en résultent. Nous prendrons notamment l'exemple d'une échelle de classification des locuteurs utilisée dans ESLO1, échelle élaborée par Alix Mullineaux, qui constitue un facteur sociologique important pour l'étude de phénomènes variables tels que la liaison.
L'exemple des analyses sur la liaison dans ces corpus illustrera les réponses aux questions posées mais aussi les limites du recours au corpus. Ce sera également l'occasion de présenter des perspectives concrètes qui ne sont finalement qu'un prolongement de réponses à la définition d'une science du corpus qui se doit aussi de délimiter son périmètre au sein de la discipline tout en s'ouvrant à de nouveaux horizons.

Mots-clés
CORPUS ORAL, VARIATIONS, LIAISONS, SOCIOLINGUISTIQUE, HUMANITES NUMERIQUES

Bibliographie

ASHBY, William. 1981. « French liaison as a sociolinguistic phenomenon », Cressey, William & Napoli, Diana éd., *Linguistics Symposium on Romance Languages (9th)*, Washington, DC, Georgetown University Press, p. 46-57.

BAUDE, Olivier & DUGUA, Céline. 2011. « (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf linguiste ? », *Corpus*, 10, *Varia*, p. 99-118.

BAUDE, Olivier & DUGUA, Céline. A paraitre. Usages de la liaison dans le corpus des ESLOs : vers de nouveaux (z)ouvrages de référence ? actes du colloque international DIA du français actuel, *La dia-variation en français actuel des corpus aux ouvrages de référence (dictionnaire, grammaire),* 29-31 mai 2013 Université de Sherbrooke, Québec, Canada, Peter Lang.

CHEVROT, Jean-Pierre, DUGUA, Céline & FAYOL, Michel. 2009. « Liaison, word segmentation and construction in French : a usage-based account », *Journal of Child Language*, 36 (3), p. 557-596.

DE JONG, Daan. 1990. « The syntax-phonology interface and French liaison », *Linguistics*, 28, (1), p. 57-88.

DE JONG, Daan. 1991. « La liaison à Orléans (France) et à Montréal (Québec) », *Actes du XII^e Congrès International des Sciences Phonétiques,* Aix-en Provence, France, p. 198-201.

DE JONG, Daan. 1994. « La sociophonologie de la liaison orléanaise », Lyche, Chantal ed., *French Generative Phonology : Retrospective and Perspectives*, Salford, ESRI, p. 95-129.

DURAND, Jacques, LAKS, Bernard & LYCHE, Chantal. 2002. « La phonologie du français contemporain : usages, variétés et structure », Pusch, Claus & Raible, Wolfgang ed., *Romanistische Korpuslinguistik - Korpora und Gesprochene Sprache / Romance Corpus Linguistics - Corpora and Spoken Language*. Tübingen, Gunter Narr Verlag, p. 93-106.

DURAND, Jacques, LAKS, Bernard, CALDERONE, Basilio & TCHOBANOV, Atanas. 2011. « Que savons-nous de la liaison aujourd'hui ? », *Langue française*, 169, p. 103-135.

ENCREVE, Pierre. 1976. « Labov, linguistique, sociolinguistique », Labov, William éd., *Sociolinguistique*, Paris, Les éditions de Minuit, p. 9-35.

ENCREVE, Pierre. 1988. *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*, Paris, Edition du Seuil.

ESHKOL-TARAVELLA Iris, BAUDE Olivier, Maurel Denis, HRIBA Linda, DUGUA Céline, TELLIER Isabelle. 2012. « Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012 », *Ressources linguistiques libres*, *TAL*, 52/3 (/2011), p. 17-46.

HABERT, Benoît. 2000. « Des corpus représentatifs : de quoi, pour quoi, comment ? », Bilger, Mireille éd., *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, p. 11-58.

LAKS, Bernard. 2005. « La liaison et l'illusion », *Langages*, 158, p. 101-125.

LAKS, Bernard. 2007, « Les hommes politiques français et la liaison (1908-1999) », Baronian, Luc & Martineau, France éd., *Modéliser le changement : Les voies du français*, Montréal, Presses de l'Université de Montréal, p. 237-269.

LAKS, Bernard. À paraitre. « Diachronie de la liaison 1999-2011 : le cas de la parole publique », Durand, Jacques *et al.* éd., *La phonologie du français : normes, périphéries, modélisation*, Presses Universitaires de Paris Ouest.

MULLINEAUX Alix & BLANC Michel. 1982. « The problems of classifying the population sample in the socio-linguistic survey of Orléans (1969) in terms of socio-economic, social and educational categories », R*eview of Applied Linguistics*, 55, p. 3-37.

VASLIN-CHESNEAU, Annie. 2008. *Analyse diachronique de la variation sociolinguistique à partir de deux corpus orléanais*, thèse de doctorat, Université d'Orléans.

WAUQUIER-GRAVELINES, Sophie, ENCREVE, Pierre & SCHEER, Tobias. 2005. « Liaison in French, towards an unified explanation of variation », Colloque Phonologie du Français Contemporain, *Phonological Variation, the case of French*, Tromsø, Norway, 25–27 August 2005.

LES CORPUS ET LA PARTITION DES STRUCTURALISMES

Gabriel BERGOUNIOUX
Laboratoire Ligérien de Linguistique (UMR 7270) / Université d'Orléans

Les anthologies de textes intitulées « corpus » sont de tradition ancienne. Au XIX^e^ s., malgré une extension du concept à l'épigraphie – les *Corpus Inscriptionum* du latin, du grec, des langues sémitiques… –, les corpus linguistiques qui servent aujourd'hui de référence à la discipline ne sont pas issus de ces compilations. Quelle explication donner à cet état de fait ? Le désintérêt pour l'étude des formes non littéraires, a fortiori non écrites, dans les traditions académiques occidentales, n'a pas résisté à la confrontation directe, en Amérique et en Asie centrale, entre immigrants et populations locutrices de langues non indo-européennes, les premières présentes sur des territoires que l'expansion impérialiste annexait. La distance qui existait entre la métropole et ses conquêtes coloniales se trouvait abolie aux États Unis et en Russie où les « indigènes » se trouvaient inclus dans l'espace englobé par les fronts pionniers. Ainsi, tandis que l'Europe se suffisait, au-dehors, d'un relevé topographique des langues exotiques et de quelques grammaires tout en représentant sous forme d'atlas les variations dialectales des langues nationales, les cultures submergées par le flux migratoire russe ou anglo-saxon, déterritorialisées, ne devraient de subsister dans la mémoire de l'humanité qu'à proportion des transcriptions de leurs récits. La confection de corpus de textes pour des langues généralement à tradition orale devenait le format linguistique de leur représentation, dissociant les données et les analyses. À côté des travaux de l'école russe dans le Caucase et en Sibérie, l'œuvre de F. Boas a été fondatrice dans l'organisation de l'enquête et la formation des chercheurs comme dans la présentation matérielle des corpus. Outre ses élèves, en particulier E. Sapir, d'autres chercheurs américains, comme L. Bloomfield (1928), ont continué dans cette voie. Ce recours au corpus a distingué, dans leurs pratiques et leurs conceptions, les structuralismes russe et américain d'un côté et français de l'autre. La transcription pour des langues sans écriture a favorisé la réflexion phonologique et la définition d'une division du travail avec les enquêtes ethnographiques. En Europe occidentale, à l'inverse, la grammaire des langues indo-européennes se focalisait sur le fonctionnement de langues écrites, souvent mortes, qui faisait abstraction de toute considération anthropologique pour se concentrer sur l'organisation interne des systèmes. Dans cette perspective, l'emploi de corpus de langues à tradition orale apparaît à la fois comme le point de départ et l'indicateur d'une formulation du structuralisme qui diffère dans ses sources et sa méthode du structuralisme saussurien.

Mots-clés
CORPUS, LANGUE A TRADITION ORALE, PHONOLOGIE, ETHNOGRAPHE

Bibliographie

BENVENISTE, Émile. 1966. *Problèmes de linguistique générale*, Paris, Gallimard.
BLOOMFIELD, Leonard. 1928. *Menomini texts*, New York, AMS Press.
ESPAGNE, Michel & KALINOWSKI, Isabelle. 2013. *F. Boas, le travail du regard*, Paris, A. Colin.
SERIOT, Patrick. 1999. *Structure et totalité*, Paris, PUF.

L'APPORT DES CORPUS AUX GRAMMAIRES ET/OU AUX DICTIONNAIRES : L'EXEMPLE DE CONTRE, MEME ET ENTRE

Mireille BILGER
Université de Perpignan-Via-Domitia
Paul Cappeau
Université de Poitiers

S'il est acquis que les corpus occupent une place majeure dans la linguistique contemporaine, il n'en demeure pas moins vrai que la question des données ainsi rassemblées est majeure et que l'on interroge trop peu, parfois, la question de leur représentativité. De fait, de nombreux outils disponibles (comme le *Trésor de la Langue Française* et bon nombre de grammaires) continuent à ne représenter et décrire que l'usage écrit littéraire. Or, en ce qui concerne la description syntaxique, l'appui sur des corpus de nature et de genres différents (oral, écrit, littérature, presse, explications techniques, récits de vie, etc.) se révèle primordial si l'on souhaite proposer une représentation plus fournie et plus sensible à certains faits de variation. On retrouve là des éléments du débat devenu classique entre *corpus-driven* et *corpus-based* (Tognini-Bonelli 2001).

Nous prendrons l'exemple de plusieurs formes grammaticales (telles que *contre, même* et *entre*) afin d'illustrer le décalage qui se manifeste entre les descriptions existantes (dans les grammaires ou les dictionnaires) et les pistes que fournissent les emplois attestés.

Si l'on prend l'exemple de *contre*, on peut observer dans des dictionnaires - comme le *TLF* ou une édition récente du *Grand Robert* - qu'une même hiérarchie apparaît :

En premier, la valeur locative (*se serrer contre qq'un*), suivie de la valeur d'opposition (*agir contre*). Cet ordre aurait été inversé si les dictionnaires avaient utilisé comme base de données un corpus de Presse plutôt qu'un corpus Littéraire. Par ailleurs, ces mêmes ouvrages ne signalent qu'accessoirement un emploi pourtant là encore bien représenté dans la Presse : *le prix retenu est de 200 euros par action contre 180 euros au départ*, dans lequel on note que *contre* et *pour* sont quasi interchangeables. Enfin, les dictionnaires n'accordent qu'une part mineure - et régulièrement accompagnée de commentaires sur le registre de langue - à la locution *par contre* qui renvoie cependant à plus de la moitié des emplois rencontrés dans un corpus oral de parole privée ou publique.
Ce rapide constat illustre comment la portée - ou encore la validité - de la description basée sur les données de corpus risque d'être fort différente, *voire discutable*, selon la façon dont ce dernier a été conçu et les raisons pour lesquelles il l'a été.
Se pose notamment le problème de l'identification et du traitement des collocations que l'appui sur corpus permettrait de préciser. Ainsi, dans *Le Petit Robert* l'une des rubriques de l'article *contre* fournit l'énumération suivante :
Se battre, être en colère contre qqu'un

Cette succession peut laisser croire que les deux formulations sont comparables. Or, un sondage dans *Frantext* catégorisé laisse penser que *se battre contre* (296 occurrences) est bien une collocation, statut plus difficile à accorder à *être en colère contre* (15 occurrences).

En ce qui concerne le domaine (morpho)syntaxique, il en est de même. Les phénomènes grammaticaux peuvent varier de manière significative selon le type de production ou la situation de parole d'où l'intérêt de s'appuyer sur des corpus variés et échantillonnés afin d'en parfaire la description, d'en renouveler l'analyse ou d'en proposer une nouvelle présentation.

En ce sens, la description linguistique fondée sur corpus met en jeu la conception même de la grammaire. Elle oblige à revenir sur des oppositions souvent présentées comme fondamentales, telles que celles que l'on a pu poser entre Oral et Écrit, entre Lexique et Grammaire, ou encore entre « système » et « usages » du système (Halliday 1987).

Mots clés

CORPUS, ORAL, ECRIT, SYNTAXE, GENRES

Éléments bibliographiques

BIBER, Douglas. 1988. *Variation accross speech and writing*, Cambridge, Cambridge University Press.

BILGER, Mireille et CAPPEAU, Paul. 2003. « Les emplois de "contre" dans les corpus de français parlé et de presse écrite », *Recherches linguistiques*, 26, p. 91-111.

BULTON, Alex & TYNE, Henry. 2014. *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues*, Paris, Didier.

HALLIDAY, Michael A.K. 1987. « Spoken and Written Modes of Meaning », Horowitz, Rosalind & Samuels, S. Jay ed., *Comprehending oral and written language*, New York. Academic Press.

HALLIDAY, Michael A.K., TEUBERT, Wolfgang, YALLOP, Colin & CERMÁKOVÁ, Anna. 2004. *Lexicology and Corpus Linguistics. An introduction*, London, Continuum.

LAKS, Bernard. 2010. « La linguistique des usages : de l'*exemplum* au *datum* », *Travaux linguistiques du CerLiCO*, 23, p. 11-26.

SINCLAIR, John. 1991. *Corpus Concordance Collocation*, Oxford, Oxford University Press.

TOGNINI-BONELLI, Elena. 2001. *Corpus linguistics at work,* Amsterdam, John Benjamins Publishing Company.

REPENSER L'HISTOIRE DE LA LINGUISTIQUE ROMANE PAR LA CONSTITUTION DE NOUVEAUX CORPUS : L'EXPERIENCE DU PROJET *DICTIONNAIRE HISTORIQUE DES CONCEPTS DESCRIPTIFS DE L'ENTITE ROMANE* (D.HI.CO.D.E.R.)

Anne-Marie CHABROLLE-CERRETINI
Université de Lorraine-ATILF
Cyril DE PINS
Université ParisVII-UMR HTL
Narcís IGLESIAS FRANCH
Université de Gérone
Christophe REY
Université de Picardie, LESCLAP

Notre proposition de communication s'inscrit dans le cadre du projet de recherche D.HI.CO.D.E.R qui vise à recenser les concepts ayant permis de décrire l'entité romane depuis le XIX^{e} siècle jusqu'à aujourd'hui (Chabrolle-Cerretini, à paraître et 2009).
Alors même qu'il y a généralement consensus sur les débuts de la linguistique romane - dès les années 1836-44 avec la parution de la *Grammaire des langues romanes* de F. Diez - et que l'on évoque ensuite généralement les années 1880, puis la période 1930-1950 comme moments-clés de cette linguistique, le projet D.HI.CO.D.E.R. reconsidère cette périodisation et se positionne par rapport à des notions comme celles de « paradigme » et de « texte fondamental » pour questionner ce qui est présenté comme des écoles ou des approches théoriques particulières constitutives de l'étude des langues romanes.
Cette recherche, qui induit implicitement que la linguistique romane dont nous cherchons à étudier les étapes fondatrices est bien celle qui depuis ses débuts s'est donné pour objet les langues romanes, toutes les langues romanes prises ensemble, repose sur la nécessité de faire émerger de nouveaux ensembles textuels (ouvrages théoriques, grammaires, correspondances) pour chacune de ces variétés linguistiques.
À travers cette communication permettant de présenter les grands principes méthodologiques de D.HI.CO.D.E.R., d'énoncer ses principaux écueils et verrous scientifiques, mais aussi d'exposer ses premiers résultats, nous souhaitons mettre en évidence le besoin indispensable – pour ce type de travail – de constituer de nouveaux regroupements de données.
La reconstruction des ensembles conceptuels envisagée doit en effet se faire par une recherche et une étude nouvelle de textes théoriques et descriptifs de toutes les aires linguistiques de la Romania s'appuyant notamment sur la prise en compte des différentes traditions linguistiques nationales qui, on le constate, s'accentuent de plus en plus, pour des raisons institutionnelles (cf. M. Glessgen 2000).
Les regroupements théoriques envisagés nous amènent logiquement à un questionnement de la notion de « corpus » (matérialité- exhaustivité et clôture- dépouillement- objet de l'analyse) et à une présentation de notre corpus de nature bibliographique, la finalité même de ce projet étant de faire émerger une ressource lexicographique tout à fait nouvelle : le *Dictionnaire HIstorique des COncepts Descriptifs de l'Entité Romane*.
À l'image des approches défendues dans le champ de la métalexicographie - discipline qui « fait des dictionnaires, de leur histoire, de leur mode de traitement sémantique du lexique, et des problèmes résultant du travail lexicographique, son objet de réflexion et de recherche » (Neveu, F., 2011 : 229) -, le dictionnaire numérique élaboré dans le cadre de ce projet pourra être considéré comme un corpus à part entière pour (re)découvrir les frontières conceptuelles de l'entité linguistique romane.

Mots-clés

LINGUISTIQUE ROMANE, ROMANIA, CONCEPTS, DICTIONNAIRE

Bibliographie générale

BAHNER, Werner. 1986. « Quelques problèmes méthodologiques dans l'historiographie de la linguistique romane », *Actes du XVIIIe Congrès International de linguistique et de Philologie romanes*, Université de Trèves, Tübingen, tome VII, p. 4-10.

CHABROLLE-CERRETINI, Anne-Marie. 2009. « La linguistique romane : un champ épistémologique pour penser la diversité linguistique aujourd'hui ? », Carmen Alén GARABATO, Teddy ARNAVIELLE, Christian CAMPS éd., *La romanistique dans tous ses états*, *o*rganisé par l'EA 739 Dipralang, la collaboration du Cerc et de Redoc, Université de Montpellier III, Paris, L'Harmattan (Langue et parole), p. 125-137.

CHABROLLE-CERRETINI, Anne-Marie. À paraître, « De la constitution des paradigmes en histoire de la linguistique romane : un enjeu du D.HI.CO.D.E.R. », Anne-Marie Chabrolle-Cerretini éd., *Paradigmes et concepts pour une histoire de la linguistique romane*, Actes du premier colloque de l'équipe D.HI.CO.D.E.R., Limoges, Éditions Lambert-Lucas.

GLESSGEN, Martin-Dietrich. 2000. « Les manuels de linguistique romane, source pour l'histoire d'un canon disciplinaire », *Romanistisches Kolloquium* XIV, Tübingen, Gunter Narr, p. 189-299.

NEVEU, Franck. 2011. *Dictionnaire des sciences du langage*, Paris, Armand Colin.

OESTERREICHER, Wulf. 2000. « L'étude des langues romanes », Sylvain Auroux éd., *Histoire des idées linguistiques*, tome 3, Sprimont, Mardaga, p. 183-192.

PATRIMONIALISATION ET ARCHIVAGE : L'EXEMPLE DES CORPUS ORAUX

Pascal CORDEREIX
Bibliothèque nationale de France, département de l'Audiovisuel ; Laboratoire Ligérien de Linguistique

La patrimonialisation des corpus oraux fait désormais partie de leur cycle de vie. Mais ce déplacement du « scientifique » au « patrimoine » n'est pas neutre ; il implique au contraire un certain nombre d'opérations qu'on pourrait qualifier, suivant Claude Levi-Strauss, de « miniaturisation ». Le geste de « mettre à part » (Michel de Certeau) qui caractérise toute entrée en archives amène en effet à un ensemble d'actions descriptives (inventaire, catalogage…), techniques (numérisation…), etc., normées, qui vont permettre la consultation, la diffusion, la conservation pérenne, etc. du corpus, mais qui procèdent dans le même temps à une forme de « réduction » de l'objet scientifique, en l'inscrivant dans un ensemble plus vaste qui fait sens : celui de la collection, celui de l'archive.
Cette réduction est-elle sans perte ? C'est tout l'enjeu du processus qui opère la translation entre l'objet premier : le corpus constitué dans le cadre scientifique, et son artefact le corpus devenu archive.
À partir de cas concrets, nous observerons ces dispositifs de patrimonialisation à l'œuvre, en tentant d'apporter des éclairages sur cette question : que se passe-t-il lorsqu'un corpus devient une « archive » ?

Mots-clés
ARCHIVES SONORES, CORPUS ORAUX, ARCHIVE NUMERIQUE, CONSERVATION PERENNE, BIBLIOTHEQUE NATIONALE DE FRANCE

TEXTES ET ***DOCUMENTS*** DANS L'ANALYSE DES CORPUS : NOUVEAUX OBJETS POUR LA LINGUISTIQUE ?

Rossana DE ANGELIS
Post-doctorante
Université de Calabre (Italie)
LATTICE (CNRS, ENS, Sorbonne Nouvelle)

La *linguistique des textes* a cédé sa place aujourd'hui à une *linguistique des corpus*. En passant de la dimension du *texte* – ex. l'approche de l'École sémiotique de Tartu et la *Textlinguistik* allemande – à la dimension du *corpus* – ex. les conceptions du texte de J. R. Firth (1957) aux fondements de la *Corpus Linguistics*, cf. les travaux de J. Sinclair (1991) – on peut observer une remise en question de certaines notions fondamentales: *texte*, *document*, *corpus*.
Par exemple, au sein de la *sémantique interprétative* (Rastier 1987, 1989) la description des objets linguistiques suppose l'individuation de *parcours interprétatifs* à la fois dans la dimension du *texte* (les *parcours textuels*) et dans la dimension du *corpus* (les *parcours intertextuels*). Les parcours interprétatifs dépendent essentiellement de trois facteurs différents: 1) une *pratique* descriptive; 2) un contrat interprétatif propre au *genre* et/ou au *discours*; 3) les structures particulières qui le réalisent. Néanmoins, ces critères guident à la fois l'analyse des *textes* (Rastier 2001) et l'analyse des *documents* composant les *corpus* (Rastier 2011).
Les *corpus* se présentent comme un ensemble de données construit par les linguistes en vue de l'analyse. Les *documents* dont se composent sont donc des «objets linguistiques» produits par des *pratiques* particulières. Pour comprendre la relation entre *textes*, *documents* et *corpus*, il faut donc considérer l'apport des *pratiques* (Bourdieu 1972, 1982), mais aussi la matérialité spécifique des *textes* et des *documents* dont le *corpus* est composé.
L'*herméneutique matérielle* (Szondi 1975, Molinié 2005, Mayhew 2007), par exemple, porte précisément sur la relation entre les pratiques et les objets linguistiques produits. Inspiré de la tradition herméneutique, par exemple, le philosophe M. Ferraris développe une *théorie de la documentalité.* Selon cette perspective, en analysant le lien entre l'écriture, la trace, la réalité et la mémoire, mais aussi sur l'impact du numérique dans les pratiques linguistiques, Ferraris ébauche les traits d'une théorie du *document* (Ferraris 2009, 2012).
Toute en considérant ces différentes approches, quel type d'objet linguistique et social est donc le *document* ? Quelle différence on suppose entre des *textes* et des *documents* au sein d'une linguistique des *corpus ?* Quelle est la nature des objets linguistiques faisant partie d'un *corpus ?* C'est donc à ces questions qu'on voudrait essayer de répondre.

Mots-clés
TEXTES, DOCUMENTS, PRATIQUES, LINGUISTIQUE DES TEXTES, LINGUISTIQUE DES CORPUS

Bibliographie

BAKER M., FRANCIS G., TOGNINI-BONELLI E. ed. 1993. *Text and Technology: In Honour of John Sinclair*, Amsterdam, John Benjamins.
BOURDIEU P. 1972. *Esquisse d'une théorie de la pratique précédé de Trois études d'ethnologie kabyle*, Seuil, Paris.

BOURDIEU P. 1982. *Ce que parler veut dire: l'économie des échanges linguistiques*, A. Fayard, Paris.

DUTEIL-MOUGEL C. 2007. « Groupements de textes et corpus : point de vue de linguiste », *Corpus en Lettres et Sciences sociales – Des documents numériques à l'interprétation*, éd. par C. Duteil-Mougel et B. Foulquié, Presses de l'Université de Toulouse Le Mirail, 2007, p. 225-235. Disponible sur : http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Duteil.pdf (Consultée le 04 juin 2014).

FERRARIS M. 2009. *Documentalità. Perché è necessario lasciar tracce*, Roma-Bari, Laterza.

FERRARIS M. 2012. *Lasciar tracce. Documentalità e architettura*, Roma, Mimesis.

FIRTH J. R. 1957. *Papers in Linguistics 1934–1951*, London, Oxford University Press.

MAYHEW R. J., 2007. « Materialistic hermeneutics, textuality and the history of geography: print spaces in British geography, c. 1500-1900 », *Journal of Historical Geography*, n. 33, p. 466-488.

MOLINIÉ G. 2005. *Hermès mutilé. Vers une herméneutique matérielle. Essai de philosophie du langage*, Paris, Honoré Champion.

RASTIER F. [1987] 2009. *Sémantique interprétative*, Paris, Presses Universitaires de France.

RASTIER F. 1989. *Sens et textualité*, Paris, Hachette.

RASTIER F. 2001. *Arts et sciences du texte*, Paris, Presses Universitaires de France.

RASTIER F., 2004. « Enjeux épistémologiques de la linguistique de corpus », *Texto !* [en ligne], juin 2004.
Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>. (Consultée le 04 juin 2014).

RASTIER F. 2007. « Le corpus en questions », *Corpus en Lettres et Sciences sociales – Des documents numériques à l'interprétation*, éd. par C. Duteil-Mougel et B. Foulquié, Presses de l'Université de Toulouse Le Mirail, 2007, p. VIII-XIII.

RASTIER F. 2011. *La mesure et le grain : sémantique de corpus*, Paris, Honoré Champion.

SINCLAIR J. McH. 1991. *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

SZONDI P. 1975. *Einführung in die literarische Hermeneutik*, a cura di Jean Bollack e Helen Stierlin, Frankfurt-am-Main: Suhrkamp [trad. it. par B. Cetti Marinoni, *Introduzione all'ermeneutica letteraria*, Parma, Pratiche, 1979].

LES CORPUS METALINGUISTIQUES ET L'HISTOIRE CONCEPTUELLE DES THEORIES LINGUISTIQUES – UNE CONTRADICTION ?

Gerda HAßLER
Université de Potsdam

Les difficultés du travail conceptuel à base de corpus métalinguistiques sont évidentes : les corpus permettent la recherche de signifiants, pas de signifiés. Si l'on veut étudier l'histoire des concepts linguistiques, il faut prendre en compte le changement de leurs dénominations qui, parfois, reflètent des points de vue théoriques ou idéologiques. D'autre part, les mêmes signifiants ne garantissent pas l'identité du contenu conceptuel. Comment peut-on sortir de ce dilemme ? Nous nous sommes posé cette question lors du travail sur le lexique des théories linguistiques du XVII^e^ et du XVIII^e^ siècle (Haßler/Neis 2009) et dans le travail ultérieur sur le XIX^e^. À partir de cela, nous voulons partager l'expérience faite et proposer à la discussion les hypothèses suivantes.

Les concepts linguistiques se développent d'une manière discursive et leurs dénominations ne s'imposent que dans un processus de négociation et de dominance de textes de référence. Un texte de référence peut être préparé par une série de textes qui prépare le concept en question et plusieurs autres séries qui le reprennent, le divulguent et le modifient. L'étude de l'interaction entre des textes de référence et des textes sériels est une approche qui permet de relever le contenu des concepts linguistiques et leurs dénominations. De plus, il y a des textes qui sont produits avec une volonté de conceptualisation qui ne correspond pas toujours à leur réception comme ouvrages de référence.

Nous avons établi un corpus de paragraphes dans lesquels apparaissent des concepts linguistiques dans des grammaires et des traités théoriques sur les langues, des dictionnaires, des articles ou des manuels, en nous limitant au XVII^e^ et au XVIII^e^ siècles. En dehors de ce que propose le CTLF, des corpus spécialisés doivent être jusqu'à présent établis pour chaque projet. L'usage de grands corpus diachroniques tels que CORDE (Corpus Diacrónico del Español) ou Frantext peut être utile à l'étude de l'histoire des théories linguistiques dans la mesure où ils permettent d'étudier la divulgation et la généralisation de certains termes dans des textes non spécialisés.

Nous avons pu constater, entre autres, une ampleur de variation assez large des dénominations de concepts linguistiques de l'époque étudiée. Ainsi le concept de 'clarté' est exprimé par lat. *perspicuitas*, all. *Klarheit, Deutlichkeit*, *Reinigkeit*, fr. *clarté, netteté*, esp. *claridad, perspicuidad*. La traduction du terme *génie de la langue* pose problème dans les autres langues européennes qui recourent à des solutions parfois insuffisantes et qui sont discutées dans les textes : all. *Genie der Sprache*, *Genius der Sprache*, *Sprache ein Spiegel des Verstandes* (Leibniz); *besondere Art einer Sprache* (Lambert); angl. *genius of a language, particular words in every language* (Locke); *Tenor of the Language* (Hartley), *the structure or genius of the languages* (Priestley); ital. *genio della lingua*, *la varia indole delle lingue* (Algarotti). Un terme tel qu'*analogie* qui remonte à la logique aristotélicienne a connu une différenciation de sa signification : en dehors des règles de conjugaison, il désigne les principes de métaphores et un état de langue souhaitable. Au début du XIX^e^ siècle, le terme *analogie* évolue et il exprime un processus de changement d'une langue et, à partir de là, s'utilise moins pour dénommer des structures cohérentes.

Il faut, pour annoter les textes, élaborer un outil qui permette de manier cette diversité des signifiants et la diversification des signifiés. Le travail onomasiologique doit être combiné avec une étude sémasiologique des termes utilisés qui tienne compte de leurs contextes et de la spécificité des emplois.

Mots clés

CORPUS METALINGUISTIQUE, CONCEPTS LINGUISTIQUES, TEXTES DE REFERENCE, DES TEXTES SERIELS, ONOMASIOLOGIQUE, SEMASIOLOGIQUE

Bibliographie

ARCHAIMBAULT, Sylvie. 2006. « L'histoire de la linguistique, un élément d'une culture linguistique nationale », *Histoire Épistémologie Langage* 28/1, p. 77-88.

AUROUX, Sylvain & COLOMBAT, Bernard & LALLOT, Jean. 1998. « Dictionnaire de la terminologie linguistique », *Histoire Épistémologie Langage* 20/1, p. 147-165.

BUSSE, Dietrich. 2005. *Brisante Semantik: neuere Konzepte und Forschungsergebnisse einer kulturwissenschaftlichen Linguistik*, Tübingen, Niemeyer.

COLOMBAT, Bernard & SAVELLI, Marie éd. 2001. *Métalangage et terminologie linguistique : actes du colloque international de Grenoble, Université Stendhal, Grenoble III, 14-16 mai 1998*, Leuven & Paris, Peeters.

HAßLER, Gerda & NEIS, Cordula. 2009. *Lexikon sprachtheoretischer Grundbegriffe des 17. und 18. Jahrhunderts*, 2 vol., Berlin, New York, Walter de Gruyter.

HAßLER, Gerda. 2005. « Dictionnaire onomasiologique et métalangage des XVIIème et XVIIIème siècles », *Lexicographica* 21/2005, 58-70.

KÖLLER, Wilhelm. 2006. *Narrative Formen der Sprachreflexion: Interpretationen zu Geschichten über Sprache von der Antike bis zur Gegenwart*, Berlin & New York, de Gruyter.

UNE LINGUISTIQUE OUTILLEE, POUR QUELS OBJETS ?

Marie-Paule JACQUES
Université Grenoble Alpes, UJF / LIDILEM

La facilité améliorée de la constitution de corpus à la fois étendus et spécialisés, de l'accès à divers ensembles textuels de taille variable – via notamment des portails spécialisés tels le *Virtual Language Observatory*[1] ou le portail *Ortolang*[2] – font de l'appui sur corpus une forme quasi-obligée de la linguistique moderne.

Nous abordons ici la question des implications de cette linguistique de corpus à l'égard des objets traités. Historiquement, la linguistique de corpus, tout du moins l'école anglaise, s'est focalisée sur les formes, plus particulièrement le lexique, avec pour objectif de caractériser le sens, notamment par la co-occurrence et la collocation (Léon 2007 ; Léon 2008). A côté d'un appareillage conceptuel tout à la fois de l'essence de la linguistique de corpus et de la sémantique (Teubert 2005), s'est développé un appareillage logiciel propre à permettre aux chercheurs de recueillir les données idoines pour la caractérisation des mots et de leurs sens, notamment à travers les concordances (Pincemin 2007). L'approche de cette linguistique est clairement onomasiologique et son programme concerne tout à la fois la façon dont une même forme est susceptible de recevoir diverses significations selon son contexte et les combinaisons récurrentes au sein des textes. La focalisation sur les formes et l'hypothèse que les co-occurrences fournissent des clefs pour le fonctionnement des textes et du discours ont rencontré en France le courant de l'analyse de discours (Charolles et Combettes 1999), s'ouvrant alors sur la statistique textuelle (Sueur 1982) puis récemment sur la textométrie[3]. Ce sur quoi nous voulons insister avec ce parcours schématique et parcellaire est qu'il ne s'agit pas uniquement ici d'une question de méthode mais d'une position forte sur l'objet de la linguistique : se préoccuper des formes et aller vers le sens.

De ce fait, le chemin inverse, du sens vers les formes, parait hétéronyme à une linguistique de corpus. Or, répondre à la question « comment s'exprime... » est pour la caractérisation de la langue tout aussi profitable que répondre à la question « que signifie... » mais pratiquement bien moins aisé. En effet, ce second versant de questionnement implique que l'on ne connaisse pas à l'avance les formes sur lesquelles la recherche porte. Par exemple, lorsque l'on se demande comment s'opère une antonomase (Leroy, 2004), comment se formulent une définition naturelle (Rebeyrolle 2000), une relation d'hyperonymie (Hearst 1992 ; Borillo 1996) ou « partie-tout » (Jackiewicz 1996), une cause ou une conséquence (Nazarenko 2000), une référence à autrui dans un article scientifique (Teufel *et al.* 2006), etc. Pour répondre à de telles questions, impossible de s'appuyer au départ sur la recherche de formes car l'expression d'un même contenu sémantique peut impliquer une variété d'unités lexicales de diverses natures. Le programme de la recherche change : moins s'interroger sur la façon dont le sens vient aux mots ou dont il se construit et se propage au sein d'un texte qu'explorer systématiquement et modéliser la variété de l'expression d'un contenu. Cette orientation suppose une description de constructions sub-phrastiques en termes d'inventaire des éléments grammaticaux et lexicaux potentiels et de définition des places syntaxiques qu'ils peuvent occuper. Les outils de TAL permettent alors d'exploiter un étiquetage morpho-syntaxique, voire des relations syntaxiques, et d'évaluer la polyfonctionnalité des constructions décrites (Jacques et Aussenac 2006).

1 http://www.clarin.eu/content/virtual-language-observatory
2 http://www.ortolang.fr
3 http://textometrie.ens-lyon.fr/spip.php?rubrique80

Mots-clés

APPROCHE SEMASIOLOGIQUE, OBJETS LINGUISTIQUES, OUTILS, METHODOLOGIE, TAL

Références

BORILLO, Andrée. 1996. « Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie », *LINX,* 34-35, p. 113-124.

CHAROLLES, Michel & COMBETTES, Bernard. 1999. « Contribution pour une histoire récente de l'analyse du discours », *Langue Française,* 121, p. 76-116.

HEARST, Marti 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora", *COLING-92,* p. 539-545.

JACKIEWICZ, Agatha 1996. « L'expression lexicale de la relation d'ingrédience partie-tout », *Faits de langues,* 7, p. 53-62.

JACQUES, Marie-Paule & AUSSENAC-GILLES, Nathalie. 2006. « Variabilité des performances des outils de TAL et genre textuel ». *T.A.L.*, *47/1*. Consulté à partir de : http://www.atala.org/Variabilite-des-performances-des

LEROY, Sarah. 2004. « Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique », *Revue française de linguistique appliquée,* 91, p. 25-43.

LÉON, Jacqueline. 2007. "Meaning by collocation. The Firthian filiation of Corpus Linguistics", *ICHoLS X, 10th International Conference on the History of Language Sciences,* p. 404-415.

LEON, Jacqueline. 2008. « Aux sources de la "Corpus Linguistics" : Firth et la London School », *Langages,* 171, p. 12-33.

NAZARENKO, Adeline. 2000. *La cause et son expression en français*, Ophrys, Paris.

PINCEMIN, Bénédicte. 2007. « Concordances et concordanciers : de l'art du bon KWAC », *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation : actes du XXVII*^{e} *colloque d'Albi Langages et signification,* p. 33-42. Disponible à l'adresse http://halshs.archives-ouvertes.fr/halshs-00356008.

REBEYROLLE, Josette. 2000. *Forme et fonction de la définition en discours*, thèse de doctorat. Université Toulouse II Le Mirail, Equipe de Recherche en Syntaxe et Sémantique.

SUEUR, Jean-Pierre. 1982. « Pour une grammaire du discours », *Mots,* 5, p. 143-185.

TEUBERT, Wolfgang. 2005. "My version of corpus linguistics", *International Journal of Corpus Linguistics,* 10/1, p. 1-13.

TEUFEL, Simone, SIDDHARTHAN, Advaith & TIDHAR, Dan. 2006. "Automatic classification of citation function", *EMNLP*, p. 103-110.

L'INTERNET EST UNE ARCHIVE OUVERTE, NON UN CORPUS, MAIS UNE ATTAQUE CONTRASTIVE PEUT QUAND MEME EN EXTRAIRE DE LA VALEUR

René-Joseph LAVIE
MoDyCo (Université Paris Ouest & CNRS)

Dans un travail fait avec Pierre Cadiot (Cadiot & Lavie 2014) sur les mots 'instant' et 'moment', nous proposons l'hypothèse suivante et la validons sur un grand nombre de cas.

- En cas d'individuation, le denotatum d'*instant* est repéré par rapport à l'origo temporel et celui de *moment* ne l'est pas.
- En l'absence d'individuation, *instant* et *moment* peuvent s'opposer selon l'étendue : le denotatum de *moment* est étendu, celui d'*instant* est hors étendue.

Ainsi, nous subordonnons à l'individuation le contraste, classique, de l'étendue le complétant, ceci est nouveau, par un contraste d'ancrage à l'origo temporel NUNC. Après de multiples essais c'est là l'hypothèse qui a le meilleur rendement explicatif, même si sa forme composite est critiquable.

La validation s'appuie sur des tests et sur les intuitions des auteurs et de leur classe sociolectale ; mais elle recourt aussi à l'Internet. Prendre l'Internet comme 'corpus' ne va pas de soi : il n'a pas de 'ratio' (cf. la polysémie de ce mot en latin), il transgresse tous les standards de pratique progressivement élaborés pour la confection de corpora.

Pour contenir l'effet de la variation - de genre textuel, d'âge, de sociolecte - nous avons recherché des paires d'occurrences (a) contenant 'instant' ou 'moment', (b) dans la même construction, (c) en contexte proche et (d) de la même main, pour tenter voir tout de même ce qu'un scripteur rend sensible quand il opte pour 'instant' ou pour 'moment'.

Cette quadruple condition est si stricte que dans un corpus même grand elle ne produirait rien ou presque ; mais dans le gigantesque Internet on ramasse bien quelques dizaines de paires. La pêche est intéressante. On trouve que l'occurrence avec 'instant' domine dans les titres – de dépêches d'agences, d'articles de journaux, d'annonces de produits, de comptes rendus sportifs – ou dans du discours rapporté direct, et celle avec 'moment' dans les corps ou dans du discours indirect. C'est de la main de scripteurs d'âges différents, de sociolectes différents et dans des genres textuels variés que le même contraste se trouve. Nous analysons que le contraste ci-dessus d'attachement – détachement vis-à-vis de l'origo temporel NUNC se transfère en un attachement – détachement vis-à-vis de l'origo personnel EGO.

Ce fait, surprise intrigante imperceptible avec des moyens plus conventionnels, devient apparent dans une archive immense, mais sans 'ratio', si c'est des rapports que l'on y recherche ; des 'ratios' justement.

On explorera les raisons du succès de cette méthode, et ses limites.

Mots clés

CONTRASTE ; CORPUS ; DECROCHEMENT ; DISCOURS RAPPORTE ; DISCOURS DIRECT ; DISCOURS INDIRECT ; EGO ; ETENDUE : HETEROGENEITE ; NUNC ; ORIGO ; TITRE ; VARIATION ;

Bibliographie

Cadiot, Pierre & Lavie, René-Joseph. 2014. "Les noms français *instant* et *moment*, une hypothèse originale", *Cuadernos de Filología Francesa*, 24, Cáceres, Universidad de Extremadura, p. 289-322.

LES GLOSSAIRES BRESILIENS DANS LA LITTERATURE: LES SAVOIRS SUR LA LANGUE

Vanise MEDEIROS
Université Fédérale Fluminense – UFF ; Laboratoire Archives du Sujet ; CNPq ; FAPERJ

La production de glossaires est vraiment ancienne. Généralement, ils sont spécifiques, soit à une région, soit à un genre de texte, entre autres motivations. Si l'on considère le discours littéraire dans l'élaboration de ces matériaux linguistiques en (ainsi appelée) « langue portugaise », on peut les diviser en deux grands axes : ceux qui se servent du discours littéraire, c'est-à-dire qui utilisent le texte de l'écrivain comme un exemple pour parler de la langue ; et ceux qui sont produits à partir du (et pour le) discours littéraire, c'est-à-dire qui surgissent du texte de l'écrivain et se tournent vers ce même texte. Il s'agit, donc, d'encadrer le double jeu d'exposer et de composer un corpus de matériaux distincts : le premier est élaboré à partir de la position discursive du philologue (dans ce cas, on ne différencie pas le philologue du lexicographe); le deuxième, par contre, est élaboré à partir de trois positions discursives : le philologue, l'écrivain, l'éditeur. Les deux sont sporadiques, mais les glossaires du deuxième groupe se situent généralement dans les livres de littérature; ceux qui constituent un travail indépendant sont rares. La sporadicité et le fait de faire partie du livre entraînent des problèmes pour l'élaboration d'un corpus et, en même temps, permettent qu'il soit envisagé comme un corpus d'exemplification de questions théoriques et analytiques de l'objet d'étude. Le fait d'être présent dans certains ouvrages, dans certains moments et dans certains auteurs conduit aussi à des questions et hypothèses qui dirigent la recherche ; par exemple : quand sont-ils produits ? Pourquoi ? Dans quelles conditions ? Dans quelle mesure peut-on affirmer que la présence des glossaires provient de politiques linguistiques, comme l'Accord Orthographique qui essaie d'établir une unité linguistique entre différentes nations ? Ou dans quelle mesure peut-on affirmer que les glossaires font partie de certains mouvements littéraires ? Quant aux glossaires que nous analysons, il faut dire qu'ils se déploient comme métatextes : ils s'articulent à partir du texte et vers le texte. Quand ils mettent en relief des mots et des expressions d'un texte, ils produisent un discours sur les frontières de la langue et partitionnent cette langue. Plutôt qu'aider à la compréhension d'un texte, ces glossaires s'articulent à partir d'une langue imaginaire envisagée comme langue nationale. Autrement dit, ils se constituent dans leur rapport avec les grammaires et les dictionnaires d'une langue. Cette recherche essaie aussi de comprendre le fonctionnement de ce rapport et les tensions qui en surgissent. Ainsi, le but de ce travail est de réfléchir sur la question de l'élaboration et de l'articulation d'un corpus à partir de notre objet d'étude, c'est-à-dire les glossaires produits pour les livres de littérature, afin de comprendre le fonctionnement de ces glossaires par rapport aux savoirs linguistiques qu'ils présentent.

Mots-clés
GLOSSAIRES, LITTERATURE, SAVOIRS LINGUISTIQUES

Bibliographie

AUROUX, Sylvain. 1992. *Histoire des idées linguistiques,* tome 1 et 2, Liège, Mardaga.
AUROUX, Sylvain. 1998. *La raison, le langage et les normes*, 1re éd., Paris, Presses Universitaires de France.

AUTHIER-REVUZ, Jacqueline. 1982. “Hétérogénéité montrée et hétérogénéité constitutive: éléments pour une approche de l’autre dans le discours”*, DRLAV*, 26, p. 91-151.

AUTHIER-REVUZ, Jacqueline. 1995. *Ces mots qui ne sont pas de soi: boucles réflexives et non-coïncidences du dire*, Paris, Larousse.

COLLINOT, André & MAZIERE, Francine. 1997. *Un prêt à parler : le dictionnaire,* Paris, Presses Universitaires de France.

COLOMBAT, Bernard, FOURNIER, Jean-Marie; PUECH, Christian. 2010. *Histoire des idées sur le langage et les langues*, Paris, Klincksieck (50 questions).

MARIANI, Bethania. 2004. *Colonização lingüística,* Campinas, São Paulo, Pontes.

ORLANDI, Eni ed. 2001. *História das ideias políticas: construção do saber metalingüístico e constituição de língua nacional*, Campinas, São Paulo, Pontes ; Cáceres, Mato Grosso, UNEMAT.

NUNES, José Horta. 2006. *Dicionários no Brasil*, Campinas, Pontes ; São Paulo: FAPESP ; São José do Rio Preto, FAPERP.

DE L'EMPIRIQUE AU THEORIQUE OU DE LA DIFFICULTE A OBJECTIVER LES « PHENOMENES » PEU VISIBLES : LE CAS DE L'IMPLICATION COGNITIVE DE L'ENONCIATEUR DANS L'ACQUISITION DE PRATIQUES LITTERATIEES

Christiane MORINET
Fédération Clesthia, Université Paris 3

Ce travail voudrait interroger le statut des données extraites par le chercheur pour la construction d'une représentation linguistique lors de leur traitement dans un article ou une communication en colloque. L'objectivation, par la constitution de données, d'un phénomène plus ou moins labile, comme l'implication d'un énonciateur dans les activités langagières à visée scientifique, semble faire perdre ce que l'observation empirique a manifesté. Que ce soit par la connaissance du terrain et de ce qu'il implique comme données sous-jacentes d'ordre extralinguistique ou de l'effet de transformation en produit de ce qui est repéré comme processus, les données ne changent-elles pas de « *nature* » lors de la construction ou de la diffusion de savoirs linguistiques ? De l'isolement du contexte à son traitement dans le cadre d'une démonstration, les données, justifiées par l'observation empirique, semblent perdre de leur validité, voire de leur pertinence, et ne plus soutenir aussi explicitement la progression théorique au moment de l'exposé devant la communauté des chercheurs.

Dans le cadre d'une recherche sur l'acquisition en milieu scolaire (Lycée) de « *pratiques littératiées* », permettant le « *travail de la pensée* » conséquence d'un « *grapholecte*[4]» (Ong, 2014, 34), l'observation linguistique se fait à partir de corpus de copies écrites, recueillies en situation d'enseignement et confrontées aux intentions des auteurs. L'objectif est de saisir empiriquement le « *décalage* » entre ces intentions recueillies sous formes d'échanges informels et les données écrites. Les pratiques langagières parlées apparaissent « *différenciatrices* » car elles sont plus ou moins adaptées à l'acquisition des pratiques littératiées. L'intuition de l'observateur le mène à être sensible, de façon fulgurante, au croisement de certaines données qui, une fois extraites du contexte, peinent à garder leur force révélatrice sur le phénomène traité.

Ainsi, une étude de l'ancrage énonciatif (l'usage de la première personne – *je* - face à - *on/il*-) dans des productions écrites atteste de la difficulté à objectiver cet enjeu linguistique hors du dynamisme de la relation pédagogique. L'approche énonciative de l'acquisition/mise en œuvre du « *rapport scriptural* » au langage est déterminante dans l'observation des pratiques littératiées. Mais, la réalité psychocognitive échappe par définition à toute objectivation sauf dans le résultat de ses effets, elle semble invisible à qui n'a pas empiriquement assisté à sa manifestation (cf. Bautier, 1995, p. 20-26). Pour valider l'objectivation, le chercheur a besoin de mesurer le degré de recouvrement des processus dans la construction du scripteur et dans le travail d'objectivation qu'il produit. Peut-être que l'outil de mesure prend pour étalon une petite dose de confiance aveugle lovée dans l'expérience littératiée de chaque pair et de lui-même.

Le cas du conditionnel et de son traitement dans un exercice de synthèse peut servir de second exemple. En effet, connaitre le conditionnel en tant que réalité morphologique de conjugaison et prendre acte de son effet interprétatif sont empiriquement séparables. La différence entre les deux perceptions dépend de l'implication du sujet de l'énonciation dans ses activités cognitivo-langagières. Le traitement des données se compose de la description du fait linguistique et de l'observation de son acquisition dynamique visible par sa manipulation dans

[4] « *L'écriture ayant pris possession de la psyché* » Ong, 2014, 33

un produit. Mais la coordination des deux réussit-elle à convaincre quand l'implicite suppose une activité de l'élève dans la relation pédagogique même ?
Le problème des savoirs linguistiques et de leur transmission se résout par l'expérience énonciative dont seules les conditions de mise en œuvre sont objectivables. C'est pourquoi la perspective d'une linguistique de l'acquisition des pratiques littératiées, encore en chantier, centre le traitement de ses données sur leur lien à un sujet de l'énonciation dans la réalité d'une intersubjectivité discursive.

Mots clés

DONNEES ECRITES EN LIEN AVEC DES DONNEES PARLEES, EMPIRISME, SUJET DE L'ENONCIATION, SUJET DE LA COGNITION, PHENOMENES OBSERVABLES, INTERPRETATION DES DONNEES, CONTEXTE D'OBSERVATION

Éléments bibliographiques

ADAMI. Hervé. 2000. « La maitrise de la langue, des enjeux idéologiques aux enjeux scientifiques et pédagogiques », Nancy, ATILF.

BALSIGER, Claudine, BEATRIX KÖHLER, Dominique, PIETRO, Jean-François de & PERREGAUX, Christiane. 2012. *Eveil aux langues et approches plurielles, de la formation des enseignants aux pratiques de classe,* Paris, L'Harmattan.

BARRE DE MINIAC, Christine. 2002. « La notion de littératie et les principaux courants de recherche », *La lettre de la DFLM*, 30-1, p. 27-36.

BAUTIER, Elisabeth. 1995. *Pratiques langagières, pratiques sociales*, Paris, L'Harmattan.

BEILLEROT, Jacky. 1987. *Savoir et rapport au savoir. Disposition intime et grammaire sociale*, Paris, Université Paris V - René Descartes.

BENTOLILA, Alain. 1996. *De l'illetrisme en général et de l'école en particulier*, Paris, Plon.

BENVENISTE, Émile. 1966. *Problèmes de linguistique générale,* Paris, Gallimard.

BERNIE, Jean-Paul. 2004. « L'écriture et le contexte : quelles situations d'apprentissage ? Vers une recomposition de la discipline "français" », *Linx*, 51, p. 25-39.

BERNSTEIN, Basil. 1975. *Langage et classes sociales, codes socio-linguistiques et contrôle social*, Paris, Minuit.

BILLIEZ, Jacqueline & BUSON, Laurence. 2013. « Perspectives diglossique et variationnelle-complémentarité ou incompatibilité ? », *French Language studies*, 23, p. 135-149.

BLANCHE-BENVENISTE, Claire. 2010. *Le francais. Usages de la langue parlée*, Leuven, Paris, Peeters.

BOURDIEU, Pierre. 1982. *Ce que parler veut dire. L'économie des échanges linguistiques*, Paris, Fayard.

BRONCKART, Jean-Paul. 1985. *Le fonctionnement des discours. Un modèle psychologique et une méthode d'analyse,* Paris, Delachaux & Niestlé.

CANUT, Emmanuelle & VERTALIER, Martine. 2009. *L'apprentissage du langage, une approche interactionnelle, Réflexions théoriques et pratiques de terrain*, Paris, L'Harmattan.

CHARLOT, Bernard. 1999. *Le rapport au savoir en milieu populaire*, Paris, Anthropos.

CONEIN, Bernard & GADET, Françoise. 1998. « Le français populaire des jeunes de la banlieue parisienne, entre permanence et innovation », Jannis K. Androutsopoulos & Arno Scholz éd., *Jugendsprache – langue des jeunes – youth langage. Linguistische und soziolinguistische Perspektiven, Actes du colloque de Heidelberg*, Frankfurt, Peter Lang, p. 105-123.

DANON-BOILEAU, Laurent. 1987. *Le sujet de l'énonciation*, Paris, Ophrys.

DORAY, Philippe. *Etymo-notions*, http://www.thesaurus-etynotions.fr/ETYNOTIONS.

ESPINOSA, Natacha & MORINET, Christiane. 2009. « Evaluation de la compréhension de textes narratifs », *Caractère*, 34, p. 15-20.
GADET, Françoise. 2007. *La variation sociale en français.* Paris, Ophrys.
GOODY, Jack. 2006. « La littératie, un chantier toujours ouvert », *Pratiques*, 131-132, p. 69-82.
JAUBERT, Martine & REBIERE, Maryse. 2003. « Langage, savoir, développement : quelle articulation pour quelles didactiques ? », Bordeaux, Cdrom.
KERBRAT-ORRECHIONI, Catherine. 1980. *L'énonciation de la subjectivité dans le langage,* Paris, PUF.
LAMBERT, Patricia & TRIMAILLE, Cyril. 2012. « La variation stylistique : un contenu à intégrer dans la formation des enseignants », Balsiger, Köhler, de Pietro & Perregaux, *Eveil aux langues et approches plurielles*, Paris, L'Harmattan, p. 255-267.
MORINET, Christiane. 1998. « La ponctuation entre logique de l'oral et logique de l'écrit », Defays, Jean-Marc, Rosier, Laurence, & Tilkin, Françoise, *A qui appartient la ponctuation ?* Paris, Bruxelles, Duculot, p. 275-288.
MORINET, Christiane. 2012a. *Du parlé à l'écrit dans les études*, Paris, L'Harmattan.
MORINET, Christiane. 2012b. « Quand la pratique langagière est plus qu'une affaire de langage ! Le cas du "bilinguisme social" rendu explicite par les variantes énonciatives », Françoise Demougin & Jérémi Sauvage éd., *La construction identitaire à l'école*, Paris, L'Harmattan, p. 331-340.
ONG, Walter. J. 2014. *Oralité et écriture, la technologie de la parole*, Paris, Les Belles Lettres.
SCRIBNER, Sylvia & COLE, Michael. 2010. « La littératie sans l'école, à la recherche des effets intellectuels de l'écriture », *Langage et société*, 133, p. 25-42.
TIRVASSEN, Rada. 2012. *Langages de jeunes, plurilinguisme et urbanisation*, Paris, L'Harmattan.
TRIMAILLE, Cyril & BILLIEZ, Jacqueline. 2004. *Pratiques langagières de jeunes urbains : peut-on parler de "parler" ?*, Enrica Galazzi & Chiara Molinari éd., *Les Français en émergence*, Berne, Peter Lang, p. 95-109.
VYGOTSKI, Lev. 1997. *Pensée et langage*, Paris, La Dispute.

DONNEES AUTHENTIQUES : UN GRAND CORPUS DE SMS EN FRANÇAIS

Rachel PANCKHURST*, Mathieu ROCHE**, Cédric LOPEZ***
* Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3
rachel.panckhurst@univ-montp3.fr
** Tetis, MTD (Maison de la Télédétection) UMR TETIS
mathieu.roche@cirad.fr
*** Viseo Technologies
cedric.lopez@viseo.com

À Augusta Mela,
en mémoire de son œuvre interdisciplinaire entre linguistique et informatique

Qu'est-ce que la donnée écrite en sciences du langage ? Trois types se distinguent : 1) la *donnée lexicale*, qui se présente essentiellement sous forme d'une entrée lexicale, regroupant un ensemble de propriétés ; 2) » le nom spécifique de la donnée observable en linguistique est l'*exemple* » et renvoie à « un énoncé qui pourrait être effectivement prononcé, même s'il ne l'est pas dans les faits » (Milner 1989, p. 51-52) ; 3) la donnée en tant que texte brut, i.e. *le corpus*. En linguistique(s) de corpus, il s'agit d'analyser les productions *authentiques* contenues dans le corpus. Dans certaines écoles linguistiques, au contraire, l'étude du corpus tout-venant n'a pas lieu d'être. Ainsi, perdure le débat concernant l'opposition (ou, tout au moins, la différenciation) entre exemples linguistiques (éventuellement « fabriqués ») et productions authentiques relevées dans des corpus (*cf.* entre autres, pour le français, Bilger *et al*. 2000, Cori *et al*. 2008, Habert *et al*.1997, Péry-Woodley 1995).

En vingt ans, notre propre approche a évolué : d'une analyse linguistique-informatique basée sur l'*exemple* (Panckhurst 1994, p. 39), nous sommes passée à une analyse de la donnée *authentique* figurant dans des corpus (Panckhurst 2013, p. 97, Panckhurst *et al.* 2014). Pour nous, cette mutation s'explique, d'une part, par l'évolution de l'accès aux données, et, d'autre part, par le *discours électronique médié* (Panckhurst 1997, 2006), circulant entre individus se servant d'outils électroniques (ordinateurs, tablettes, téléphones portables, etc.), qui induit des pratiques et des usages émergents. En deux décennies, la constitution de corpus numérisés ou nativement numériques est devenue monnaie courante, et cette accessibilité massive constitue en soi une nouveauté. Les données authentiques existant sous la forme de courriels, forums, chats, blogs, réseaux sociaux, et, plus récemment de SMS, facilement exploitables par les chercheurs, permettent l'observation, la fouille et l'analyse des pratiques et des usages (novateurs ou non) des scripteurs.

Dans le cadre de cette communication, nous expliquerons ce cheminement, en nous focalisant sur des recherches récentes, portant sur le recueil, le traitement et l'analyse d'un grand corpus de SMS en français, intitulé « 88milSMS » (consultable sur la grille de services d'Huma-Num). En 2004, des universitaires belges ont lancé un projet international, *sms4science* (www.sms4science.org, Fairon *et al.* 2006, Cougnon 2015), afin de constituer une grande base de données mondiale de SMS authentiques. D'autres collectes ont suivi : en 2011, plus de 93 000 SMS ont été recueillis auprès du grand public (qui pouvait également répondre à un questionnaire sociolinguistique) par un groupe de chercheurs dans la région Languedoc-Roussillon (projet *sud4science LR*, www.sud4science.org, Panckhurst *et al*. 2013, Panckhurst & Moïse 2014). À l'aide d'exemples extraits de « 88milSMS », nous montrerons que les données peuvent être appréhendées selon deux approches, « fondée sur corpus » ('corpus-based') et « guidée par corpus » ('corpus-driven'), et que le va-et-vient constant entre les hypothèses et l'observation des données constitue le point essentiel de notre démarche. L'élaboration de ce corpus a participé au développement d'un logiciel

d'anonymisation semi-automatique, *Seek&Hide*, par des étudiants (Accorsi *et al*. 2014, Patel *et al*. 2013), et d'un prototype, permettant la construction automatique de dictionnaires électroniques de SMS selon une méthode d'alignement statistique (Lopez *et al.* 2014).

Mots clefs

CORPUS, SMS, DONNEES AUTHENTIQUES, TAL, DISCOURS ELECTRONIQUE MEDIE, LOGICIEL D'ANONYMISATION, DICTIONNAIRES ELECTRONIQUES, ALIGNEMENT

Éléments bibliographiques

ACCORSI P., PATEL N., LOPEZ C., PANCKHURST R., ROCHE M. 2014. « Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques », L.-A. Cougnon, C. Fairon ed., *SMS Communication. A Linguistic Approach*, Amsterdam, Philadelphia, John Benjamins, p. 11-28.

BILGER M. 2000. *Corpus : Méthodologie et applications linguistiques*, Paris, Champion.

CORI M., DAVID S., LEON J. éd. 2008. « Construction des faits en linguistique : la place des corpus », *Langages*, 171, Paris, Larousse, septembre 2008.

COUGNON L.-A. à paraître, 2015. *Langage et sms. Une étude internationale des pratiques actuelles.* Presses universitaires de Louvain.

FAIRON C., KLEIN J.-R., PAUMIER S. 2006. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*, Presses universitaires de Louvain, Louvain-la-Neuve, Manuel + CD-Rom, http://www.smspourlascience.be/

HABERT, B., NAZARENKO, A., & SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin.

LOPEZ C., BESTANDJI R., ROCHE M., PANCKHURST R. 2014. "Towards Electronic SMS Dictionary Construction: An Alignment-based Approach", Actes du colloque LREC, Reykjavik, Islande, May 26-31, 2833-2838.
www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf

MILNER, J.-C. 1989. *Introduction à une science du langage*, Paris, Seuil.

PANCKHURST R. 1994. "A Database for linguists: intelligent querying and increase of data", *Computers and the Humanities*, 28, p. 39-52.

PANCKHURST R. 1997. « La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? », *Terminologies nouvelles*, 17, p. 56-58.

PANCKHURST R. 2006. « Le discours électronique médié : bilan et perspectives », A. Piolat éd. *Lire, écrire, communiquer et apprendre avec Internet*, Marseille, Éditions Solal, p. 345-366.

PANCKHURST R. 2013. "A large SMS corpus in French : from design and collation to anonymisation, transcoding and analysis", Colloque CILC2013, Alicante, 14-16 mars : http://web.ua.es/en/cilc2013/, Actes du colloque, Procedia — Social and Behavioural Sciences, Elsevier.
http://www.sciencedirect.com/science/article/pii/S1877042813041475

PANCKHURST R. et MOÏSE C. 2014, "French text messages. From SMS data collection to preliminary analysis", L.-A. Cougnon, C. Fairon, ed., *SMS Communication. A Linguistic Approach*, Amsterdam/Philadelphia, John Benjamins, p. 141-168.

PANCKHURST R., DETRIE C., LOPEZ C., MOÏSE C., ROCHE M., VERINE B. 2013. « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS », *Épistémè — revue internationale de sciences sociales appliquées*, 9 : « Des usages numériques aux pratiques scripturales électroniques », p. 107-138.

PANCKHURST R., DETRIE C., LOPEZ C., MOÏSE C., ROCHE M., VERINE B. 2014. "88milSMS. A corpus of authentic text messages in French", produit par l'Université Paul-Valéry Montpellier 3 et le CNRS, en collaboration avec l'Université catholique de Louvain,

financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8.

PATEL N., ACCORSI P., INKPEN D., LOPEZ C., ROCHE M. 2013. "Approaches of anonymisation of an SMS corpus", Alexander Gelbukh ed., *Computational Linguistics and Intelligent Text Processing*, 14th International Conference, CICLing 2013, Samos, Greece, Berlin, Heidelberg, Springer Verlag, p. 77-88.

PERY-WOODLEY, M.-P. 1995. « Quels corpus pour quels traitements automatiques ? », *T.A.L.*, 36/1-2, p. 213-232.

CORPUS ET LANGUE ARABE : UN CHANGEMENT DE PARADIGME

Catherine PINON
IREMAM

Cette communication vise à interroger le rapport entre la grammaire arabe et les corpus, pour expliquer la place de plus en plus importante que tiennent actuellement les corpus en linguistique arabe.

La tradition grammaticale arabe a fourni de véritables sommes sur la langue. Mais de quelle langue s'agit-il au juste ? Les grammairiens, qui se sont toujours basés sur des corpus, ont-ils réellement décrit ce qu'ils prétendaient décrire ? Quelle est la réalité de cet arabe classique forgé par les grammairiens ? Quels rapports entretient-il avec le Coran et la poésie archaïque souvent cités comme exemples ? Ce corpus, sur lequel la grammaire arabe s'est élaborée, peut être considéré comme restrictif. Les grammaires anciennes ont ensuite elles-mêmes servi de corpus-exemplier pour les grammairiens postérieurs, jusqu'à l'époque moderne, selon un phénomène d'accrétion qui a contribué à construire ce que l'on peut nommer une *sui-langue* arabe classique.

Les questions liées à la description technique de l'arabe sont nombreuses et brûlantes, tant la religion et les mythes en découlant les ont orientées depuis l'origine. Après les avoir résumées, des exemples précis illustreront la question du changement de paradigme qui s'opère dès lors que l'on a accepté de dépasser les considérations socio-mystico-religieuses qui pèsent sur la langue arabe, et dès lors que l'on a pour objectif de proposer une grammaire descriptive réaliste et dynamique de l'arabe contemporain. Des questions d'ordre épistémologique liées aux fondements des corpus surgiront alors : comment le choix du corpus oriente-t-il les descriptions ? Comment reflète-t-il le point de vue du chercheur sur la langue arabe ? Comment la construction d'un corpus tendant à un maximum de représentativité amène-t-elle à réfléchir à la langue elle-même, dans son infinie richesse et dans toute sa complexité ? Comment intégrer la pluriglossie ? Comment statuer sur les différents états de la langue ? On constatera que la méthodologie du corpus amène *de facto* à s'interroger sur la finalité d'un travail de description technique de la langue et permet de faire évoluer positivement les connaissances sur la langue arabe.

Mots clés
GRAMMAIRE ARABE – GRAMMAIRE REALISTE – ARABE CONTEMPORAIN – EPISTEMOLOGIE – LINGUISTIQUE DE CORPUS – LINGUISTIQUE HISTORIQUE - INTERNET

Bibliographie indicative

AUROUX, Sylvain. 1998. *La raison, le langage et les normes*, Paris, PUF.

BOHAS, Georges. 1981. « Quelques aspects de l'argumentation et de l'explication chez les grammairiens arabes », *Arabica*, XXVIII/2-3, Leiden, Brill, p. 204-221.

CAPPEAU, Paul, CHUQUET, Hélène & VALETOUPOLOS, Freiderikos éd. 2010. *L'exemple et le corpus, quel statut ?*, Rennes, PUR.

CALABRESE, Laura dir. 2010. *L'internet, corpus sauvage ; nouvelles ressources, nouveaux problèmes ?*, Bruxelles, Fernelmont, EME.

DELAHAYE, Jean-Paul & GAUVRIT, Nicolas. 2013. *Culturomics : le numérique et la culture*, Paris, Odile Jacob.

DELESALLE, Simone & MAZIERE, Francine. 2002. « La liste dans le développement des grammaires », *Histoire Épistémologie Langage*, 24/1, p. 65-92.
FLEISCH, Henri. 1956. *L'arabe classique. Esquisse d'une structure linguistique*. Beyrouth, Imprimerie catholique.
FLEISCH, Henri. 1979. *Traité de philologie arabe*, vol. II, Beyrouth, Imprimerie catholique.
GALA, Nuria & ZOCK, Michael. 2013. *Ressources lexicales : contenu, construction, utilisation, évaluation*, Amsterdam-Philadelphie, Benjamins.
LARCHER, Pierre. 2003. « Diglossie arabisante et *fushâ* vs *'âmmiyya* arabes : essai d'histoire parallèle », Auroux, Sylvain ed., *History of Linguistics* 1999. Selected Papers from the Eight International Conference on the History of the Language Sciences (ICHoLS VIII), Fontenay-St. Cloud, France, 14-19 September 1999, Amsterdam, Philadelphia, Benjamins (SIHoLS 99), p. 47-61.
LARCHER, Pierre. 2005. « Arabe préislamique, arabe coranique, arabe classique : un continuum ? », Ohlig, Karl-Heinz & Puin, Gerd-Rüdiger ed., *Die dunklen Anfänge. Neue Forschungen zur Enstehung und frühen Geschichte des Islam*, Berlin, Verlag Hans Schiler, p. 248-265.
LARCHER, Pierre. 2008. « *Al-lugha al-fushâ* : archéologie d'un concept 'idéolinguistique' », Catherine Miller & Niloofar Haeri dir., *Langues, religion et modernité dans l'espace musulman*, *REMMM* 124, p. 263-278.
LARCHER, Pierre. 2010. « In search of a standard: dialect variation and New Arabic features in the oldest Arabic written documents », Macdonald, Michael dir., *The development of Arabic as a written language* (Supplement to the Proceedings of the Seminar for Arabian Studies 40), Oxford, Archaeopress, p. 103-112.
PINON, Catherine. 2011. « La grammaire arabe : entre théories linguistiques et applications didactiques », *Essais de linguistique arabe*, Synergie-Monde arabe n°7, p. 75-86. En ligne (http://ressources-cla.univ-fcomte.fr/gerflint/Mondearabe7/mondearabe7.html).
PINON, Catherine. 2012. « Les enjeux épistémologiques et didactiques d'une grammaire arabe fondée sur corpus », Arnavielle, Teddy éd., *Voyages grammairiens*, Paris, L'Harmattan, p. 83-101.
PINON, Catherine. 2013. « Quel corpus peut aider à fonder la grammaire d'une langue pluriglossique ? Exemple de l'arabe contemporain », *Cahiers de Praxématique*, 54-55, Corpus, données, modèles, PUM, p. 39-58.