Colloque Jeunes chercheur(e)s en SDL
Normandie Université / Université de Rouen
Dynamique du Langage In Situ (DyLis)
23-24 novembre 2017

Diversité de la constitution des données : sur quoi travaille-t-on en Sciences du langage ?

La diversité des Sciences du langage (SDL) est sans doute à l'origine de la diversité des regards portés sur la constitution des *données*, dites aussi *observables*, ou encore *faits (linguistiques)*, eux-mêmes souvent regroupés en *corpus*. Ces données diffèrent d'un domaine à un autre en SDL : analyse de discours, sociolinguistique, didactique des langues, syntaxe, sémantique (formelle, textuelle), phonétique-phonologie…

Si le recueil des données et la constitution de corpus qui lui est généralement associée peuvent paraitre simples à décrire pour un(e) chercheur(e), ils sont néanmoins source de nombreux différends méthodologiques qui mériteraient d'être réinterrogés en SDL.

Une question méthodologique renvoie en effet très rapidement le(la) chercheur(e) à la question de la méthode qui la sous-tend, elle-même construite en lien avec une théorisation (socio)linguistique. Cela ouvre des questionnements épistémologiques et réinterroge très concrètement les choix que l'on fait, parcours obligé du(de la) jeune chercheur(e) en particulier.

On observe ainsi que ce questionnement autour de la constitution des données, parfois ramené (ou réduit ?) à une opposition caricaturale entre une linguistique qui serait *hors/sans* corpus et une linguistique *de/avec* corpus, est loin d'être clos.

Qu'en est-il de ce débat aujourd'hui ? Ne doit-on pas reconsidérer les notions de *données construites* et de *données attestées* ? Quelle place et quel rôle le(la) chercheur(e) attribue-t-il(elle) au « terrain » dans la constitution des données ? Quelle place et quel rôle s'attribue-t-il(elle) dans la démarche adoptée ? Comment ces choix participent-ils à la construction théorique de l'objet ? Comment objectiver ces questions de façon éclairante pour le processus de recherche ?

Ce questionnement invite à une redéfinition de la pratique des *corpus*, *données*, ou *observables* ; il soulève finalement la question de la ***scientificité*** en SDL, et peut-être en sciences humaines et sociales en général. Ce sont autant d'interrogations méthodologiques, théoriques, épistémologiques - et éthiques - que les jeunes chercheurs sont invités à partager lors de ce colloque.

La constitution des données ayant un réel impact sur les pratiques de recherche, sur les résultats qui en découlent, ainsi que sur leur diffusion et leurs usages, l'objectif de ce colloque ne consistera pas uniquement à débattre des notions de données, de corpus, d'observables, il entend également interroger les pratiques en proposant une réflexion autour de **l'objet d'étude** en SDL à travers les choix (théoriques et/ou méthodologiques) qu'impliquent la constitution de ces données et la formation conséquente aux outils actuellement disponibles pour les traiter.

* * *

La réflexion sera articulée autour des trois entrées suivantes :

Entrée 1 *Les données construites dans la recherche en SDL*

L'essor des *« linguistiques de corpus »* vient parfois, plus ou moins explicitement, mettre en cause la pertinence et la validité scientifiques d'une linguistique (dis)qualifiée de *« linguistique introspective »* (Jacques, 2005). Pourtant, avec Cori & David, on peut considérer que, indépendamment de la voie empruntée, la tâche du linguiste est toujours la même : *« choisir les données à considérer, les mettre en forme, les ramener à la dimension qui doit être étudiée, émettre un jugement »* (2008 : 127).

D'ailleurs, l'analyse linguistique à partir de données construites est bien ancrée dans les pratiques de recherche en SDL (notamment en France). Nous invitons le(la) jeune chercheur(e) à questionner cette « tradition » à partir des questions suivantes :

- Sur quelles bases et selon quelles modalités rassemble-t-on des données construites ? En fonction de quels objectifs de recherche ?
- Le recours aux données construites permet-il de développer une analyse plus pertinente que l'exploration de corpus forcément *limités* ? Cette analyse conduit-elle à des résultats plus généraux voire plus *scientifiques* ?
- L'analyse linguistique à partir de données construites est-elle remise en cause par le développement de nouveaux moyens d'investigation sur corpus ? ou, au contraire, doit-elle chercher en eux des pistes pour se renouveler ?

Entrée 2 *Les données attestées* : *de l'usage du/des corpus en SDL*

Le regain d'intérêt que connaît actuellement la linguistique de corpus, notamment dans les recherches francophones en SDL, vient raviver le questionnement autour des notions et des pratiques de corpus.

Il semblerait que tous(toutes) les chercheur(e)s en SDL, qui ancrent leurs travaux dans la constitution/délimitation d'un corpus, s'accordent sur le gage de scientificité que celui-ci

apporte à leurs analyses. Pourtant, il n'est pas sûr que la notion de corpus fasse consensus. Ainsi, aux rigoureuses définitions de Rastier (2005), certains opposent des pratiques beaucoup plus libres (Kilgariff & Grefenstette, 2003).

Dans ce contexte, nous proposons une réflexion autour des questions suivantes :

- Comment le(la) chercheur(e) définit-il(elle) la notion de corpus ? Quels sont ses critères de constitution ? Peut-il et doit-il être représentatif ? Selon quels critères ? Quelles pratiques pour quels objectifs de recherche ?

- Par exemple, doit-il y avoir homogénéisation et objectivation des corpus étudiés ? Peut-on alors parler de corpus sémantiquement auto-suffisants avec des univers interprétatifs clos ? (Mayaffre, 2002)

- Lorsque le terrain « fait corpus », comment est exploré et explicité le lien entre le matériau linguistique et l'usage que les locuteurs en font (Gadet & Wachs, 2015) ? S'oriente-t-on alors vers une linguistique de l'hétérogène ?

- Quelle place accorder aux corpus dans le domaine de l'enseignement-apprentissage des langues ? Favorisent-ils un apprentissage en autonomie (Ciekanski, 2014) ?

Entrée 3 *Les données (socio)linguistiques à l'ère du numérique*

Il est possible que l'accès toujours plus aisé à des bases de données textuelles toujours plus importantes ainsi que le développement d'outils de traitement linguistique (ex : *Alceste, Cordial, TXM, Hyberbase*...) bousculent les pratiques de recherche dans la constitution comme dans l'exploitation des données linguistiques. Ces outils peuvent servir, en effet, l'analyse d'un nombre important de données pour une même recherche, la constitution de bases de données exploitables selon différentes méthodologies ; ils peuvent servir aussi des analyses qualitatives de données empiriques. Tous nécessitent une maitrise technique de l'outil informatique mais aussi et surtout une réflexion quant à leur usage et les pistes d'analyses qu'ils permettent – ou pas.

Certains interrogent les liens entre *linguistique* et *informatique* comme les tenants du *Traitement automatique du langage naturel (TALn)* ou ceux de l'*Analyse des données textuelles (ADT)* notamment. Si les premiers visent l'automatisation, les seconds appréhendent plutôt l'informatique comme un outil de description linguistique (Valette, 2016). Pour d'autres encore, le Web est un corpus linguistique plus ou moins prêt à l'emploi, lieu de constitution de données attestées, qu'il s'agit pourtant, là encore, de (re)construire en tant qu'objet de recherche.

Ces pratiques invitent à se poser les questions suivantes :

- Quels outils numériques existent aujourd'hui en SDL, pour quels usages ? Quels types d'analyse du matériau linguistique favorisent-ils ? Quel protocole méthodologique nécessitent-ils ?

- Quels sont les critères pertinents pour la construction de corpus textuels dans un monde où foisonnent forums de discussion, réseaux sociaux et moteurs de recherche ? Et, notamment, quels enjeux dans le domaine de la didactique des langues à l'heure du *Data Driven Learning* (Sockett, 2014) ?

- Comment, concrètement, peut-on construire une recherche qui puisse répondre avec rigueur à ces questions ?

Les questions listées dans les trois entrées précédentes sont autant de pistes ouvertes à tous(toutes) les jeunes chercheur(e)s en SDL.

Références bibliographiques

Ciekanski Maud, 2014, « Les corpus : de nouvelles perspectives pour l'apprentissage des langues en autonomie ? », *Recherche en didactique des langues et des cultures* 11-1, pp. 111-135.

Cori Marcel & Sophie David, 2008, « Les corpus fondent-ils une nouvelle linguistique ? », *Langages* 171, pp. 111-129.

Gadet Françoise & Sandrine Wachs, 2015, « Comparer des données de corpus : évidence, illusion ou construction ? », *Langage et Société* 154, pp. 33-49

Jacques Marie-Paule, 2005, « Pourquoi une linguistique de corpus ? » *in* Geoffrey Williams (dir.), *La linguistique de corpus*, Rennes : Presses Universitaires de Rennes, pp. 22-30.

Kilgariff Adam & Gregory Grefenstette, 2003, « Introduction to the Special Issue on the Web-as Corpus », *Computational Linguistics* 29-3, pp. 333-347

Mayaffre Damon, « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus* [En ligne], décembre 2003, consulté le 15 septembre 2016. URL: https://corpus.revues.org/11

Rastier François, 2005, « Enjeux épistémologiques de la linguistique de corpus » *in* Geoffrey Williams (dir.), *La linguistique de corpus*, Rennes : Presses Universitaires de Rennes, pp. 31-45.

Sockett Geoffrey, 2014, « Corpus et perspectives pour l'enseignant », *Recherche en didactique des langues et des cultures* 11-1, pp. 79-91.

Valette Mathieu, « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée », *JADT 2016 : 13èmes Journées internationales d'Analyse statistique des Données Textuelles* [En ligne], juin 2016, consulté le 12 octobre 2016. URL : https://jadt2016.sciencesconf.org/84134/document

* * *

Modalités de soumission

Les propositions de communication d'environ 3000 signes, bibliographie non comprise, devront être envoyées avant le 20 mars 2017 à l'adresse mail suivante : cjcrouen2017@sciencesconf.org

Elles seront examinées anonymement par deux membres du comité scientifique.
Les résultats seront communiqués début juin.

Les communications donneront lieu à une publication dans les actes du colloque (modalités de soumission et éditeur à venir).

Tarif : Doctorants(es) communicants(es) : 25 €.

Les frais d'inscription comprennent l'entrée à toutes les conférences, les pauses-café, les deux repas du midi ainsi qu'un dîner de bienvenue dans le centre-ville de Rouen.

Comité d'organisation :

Hayat ALILOUCHE (Doctorante, Université de Rouen)
Selda ARAZ (Doctorante, Université de Rouen)
Thomas BERTIN (Doctorant, Université de Rouen)
Amandine DENIMAL (MCF, Université de Rouen)
Maryvonne HOLZEM (MCF, Université de Rouen)
Cleudir MOTA (Doctorant, Université de Rouen)
Fadila TALEB (Doctorante, Université de Rouen)

Comité Scientifique :

Salih AKIN (MCF-HDR, Université de Rouen)
Angelina ALEKSANDROVA (MCF, Université Paris Descartes)
Marc DEBONO (MCF, Université François Rabelais Tours)
Amandine DENIMAL (MCF, Université de Rouen)
Carine DUTEIL-MOUGEL (MCF, Université de Limoges)
François GAUDIN (PU, Université de Rouen)
Laurent GOSSELIN (PU, Université de Rouen)
Thierry GUILBERT (MCF-HDR, Université de Picardie Jules Verne)
Maryvonne HOLZEM (MCF, Université de Rouen)
Jean-Marc LEBLANC (MCF, Université Paris-Est Créteil Val de Marne)
Ksenija LEONARD (MCF, Université Paul-Valéry Montpellier 3)
Damon MAYAFFRE (PU, Université Nice Sophia-Antipolis)
Véronique MIGUEL-ADDISU (MCF, Université de Rouen)
Grégory MIRAS (MCF, Université de Rouen)
Thierry PONCHON (MCH-HDR, Université de Reims Champagne-Ardenne)
Vassil MOSTROV (MCF, Université de Valenciennes et du Hainaut-Cambrésis)
Christophe REY (PU, Université de Picardie Jules Verne)
Audrey ROIG (MCF, Université Paris Descartes)
Richard SABRIA (PU, Université de Rouen)
Mathieu VALETTE (PU, Inalco)

Contact 🖱-✉ :

Lien : https://cjcrouen2017.sciencesconf.org /// Courriel : cjcrouen2017@sciencesconf.org