Nouvelles approches du corpus en linguistique anglaise (9-10 juin 2016)

New approaches to corpus in English linguistics (9-10 June 2016)

Projet soumis à la SFR Agorantic 2015-2016

1 – Éléments de présentation

Le projet proposé porte sur un colloque en linguistique anglaise visant à associer, d'une part, les approches linguistiques dont la démarche relève d'une analyse plutôt qualitative des données, que ce soit en sémantique, en pragmatique, en analyse du discours, ou dans le domaine de l'énonciation, au sens large, et, d'autre part, les approches linguistiques s'appuyant sur des analyses d'ordre quantitatif – en termes d'une analyse statistique fréquentielle, et des collocations, colligations, prosodies sémantiques, etc. que celle-ci peut faire apparaître.

- Laboratoire porteur du projet : Identités Culturelles, Textes et Théâtralité (ICTT) – EA 4277.
- Laboratoire associé : Laboratoire Informatique d'Avignon (LIA) – EA 4128.

Le comité d'organisation :

- Aurélia Barrière, Université d'Avignon (Maison de la Recherche)
- Annemarie Dinvaut, Université d'Avignon (ICTT)
- Graham Ranger, Université d'Avignon (ICTT)

2 – Descriptif du projet :

Le colloque vise à rassembler des chercheurs en linguistique de divers horizons, dans la recherche de synergies potentielles entre, d'un côté, une linguistique dont l'objectif est la description des marqueurs, des constructions, des genres discursifs, au travers d'une modélisation théorique de l'activité du langage, et, de l'autre, une linguistique qui travaille sur des données numérisées dans l'optique d'une analyse statistique et fréquentielle.

On cherche ainsi à encourager un dialogue entre ces deux visions du langage, en considérant, d'une part, que les modèles théoriques issus d'une linguistique plutôt "qualitative" pourront aider à orienter les recherches sur les données et, d'autre part, que des analyses statistiques fines menées sur corpus pourront servir à vérifier ou à infirmer des hypothèses issues de micro-analyses qualitatives.

C'est dans cet esprit de rassemblement et de dialogue qu'a été redigé l'appel à communications, dont voici la version française (la version anglaise est donnée en annexe).

L'appel à communications :

Historiquement, la linguistique pragmatique – que ce soient les théories des actes de langage, de la pertinence, de l'argumentation, de l'énonciation, etc. – ne s'est pas toujours servie des corpus, ou alors s'en est servie, mais dans un esprit d'"éclectisme illustratif" (Kohnen 2015:56 [1]), en citant des occurrences authentiques pour illustrer telle ou telle perspective théorique. Plus récemment, diverses approches sémasiologiques ont considéré que l'apparente polysémie en contexte d'un marqueur donné peut s'expliquer de manière satisfaisante en termes de configurations complexes, construites par l'interaction des potentiels sémantiques – ou invariants – d'éléments en cooccurrence. Une telle approche ouvre de nouvelles perspectives pour l'exploitation quantitative des données de corpus dans le cadre des théories qui ont souvent privilégié l'analyse qualitative d'un

nombre limité de cas. Les technologies des requêtes sur corpus permettent désormais des interrogations sophistiquées en termes d'affinités de collocation, ciblant des données qui peuvent – en fonction des conventions d'étiquetage du corpus – impliquer aussi bien des traits linguistiques que paralinguistiques (pauses, chevauchements, identités des locuteurs, genres textuels, etc.).

Le présent colloque vise par conséquent à explorer l'utilité et la pertinence de l'analyse quantitative à partir de corpus d'anglais ou de variétés d'anglais dans des domaines de la recherche linguistique où une approche plus qualitative a jusqu'à présent prévalu. On pourra par exemple s'interroger sur:

- la description sémantico-pragmatique de marqueurs spécifiques ou de configurations de marqueurs à partir de données quantitatives issues de corpus;
- les exigences en termes de prétraitement des corpus en vue de telles recherches, ex. étiquetage spécifique ou ad hoc, conventions des identifiants employés;
- les types de requêtes sur corpus, les outils, la syntaxe et/ou les algorithmes pertinents pour de telles recherches.

Nous encourageons également les travaux qui ciblent l'utilisation des corpus dans l'étude de l'identité linguistique, thématique clé de notre laboratoire local Identité Culturelle, Textes et Théâtralité.

Langues du colloque: anglais et français.

[1] Kohnen, Thomas. 2014. "Speech Acts: a diachronic perspective", in Corpus Pragmatics. A Handbook. Aijmer, K. and Rühlemann, C. (eds.), Cambridge University Press.

État de l'organisation matérielle :

Nous prévoyons un public d'une soixantaine de personnes, composé de chercheurs français, internationaux et d'étudiants de linguistique (au niveau doctorant et master), d'Avignon et des universités voisines.

L'appel à communications est ouvert jusqu'au 29 février, 2016. Les propositions seront soumises à une double relecture, par les membres du comité scientifique (la composition est donnée en annexe). Les résultats seront communiqués aux auteurs avant le 31 mars, 2016.

Le site web du colloque est désormais ouvert : http://nacla1.sciencesconf.org/

Conférenciers invités :

Deux conférenciers invités ont accepté d'intervenir :

Lucie Gournay, Professeur de Linguistique et Vice-Présidente de la Commission de la Recherche du Conseil Académique de l'Université de Paris-Est Créteil, pratique une linguistique énonciative, orientée notamment vers les problématiques contrastives. Pr Gournay s'interrogera sur la constitution de corpus, dans l'étude de construction syntaxiques complexes.

Sebastian Hoffmann, Professeur de Linguistique à l'Université de Trèves, Allemagne, est un linguiste de corpus internationalement reconnu, co-auteur notamment de BNCweb, interface Web libre source d'un ensemble de scripts permettant l'interrogation multi-critériée du British National Corpus en CQL et en CQL simplifié. L'intervention du Pr Hoffmann reste à préciser mais pourra porter sur une démonstration de l'outil BNCweb.

3 – Objectifs et résultats attendus

- Recherche : L'objectif est en premier celui de faire avancer l'état de la recherche par le biais de nouvelles coopérations entre approches qui pourront, nous l'espérons, se montrer mutuellement enrichissantes. Ceci pourra ensuite donner lieu à des collaborations pérennes entre les informaticiens du LIA, et les linguistes, littéraires et civilisationnistes d'ICTT.
- Valorisation : Une publication des communications sous forme d'un volume thématique est prévue. Un premier éditeur international (Cambridge Scholars Publishing) a spontanément pris contact avec les organisateurs du colloque en vue d'une publication prochaine.

4 – Caractère innovant de ce projet

Le caractère innovant du projet provient des nouveaux éclairages qu'il se propose d'apporter sur des domaines de recherches qui, malgré un objet d'étude très proche, sinon commun, ont trop longtemps évolué séparément.

En effet, si les tenants des approches "qualitatives" sont issus très souvent d'une réflexion littéraire et / ou philosophique visant à modéliser l'activité du langage *in abstracto*, ceux des approches "quantitatives", issus des sciences de l'information ou de la communication, visent à décrire la production langagière, souvent à des fins prédictives, éventuellement commercialisables.

Malgré cette opposition dans les finalités des approches, nous restons convaincus qu'elles peuvent et doivent apprendre à dialoguer autour de l'objet commun d'étude qu'est le langage. Un tel rapprochement nous paraît représenter un tournant innovateur dans le contexte de l'état actuel de la recherche linguistique en France.

5 – Sa dimension interdisciplinaire (laboratoires de disciplines différentes)

Les thématiques abordées dans le colloque devront intéresser à la fois les linguistes, littéraires et civilisationnistes du laboratoire ICTT, les informaticiens du laboratoire LIA – dans l'optique de la recherche algorithmique permettant d'extraire des traits qualitatifs de données linguistiques numérisées, les chercheurs de l'équipe Culture et Communication – dans l'optique notamment du traitement des données issues de sites web ayant trait au fait culturel.

Le comité scientifique comprend, de fait, à l'intérieur de l'Université d'Avignon, deux membres du laboratoire ICTT, Anika Falkert, et Graham Ranger, et Richard Dufour, du laboratoire LIA.

6 – Positionnement dans Agorantic

- Axe 1 : Culture et numérique.

Le projet se situe au coeur de l'axe 1 de l'Agorantic, pour plusieurs raisons.

D'abord, il s'agit du repérage quantitatif – par l'exploitation du "numérique" – d'éléments relevant traditionnellement d'une appréciation qualitative de l'interprétant – c'est-à-dire, de sa "culture".

Ensuite, les outils informatiques permettant ce type de repérage sont de plus en plus souvent accessibles via Internet. La venue du Professor Sebastian Hoffmann de l'Université de Trèves, l'une des forces motrices derrière l'interface BNCweb et le projet dérivé CQPweb intéressera à ce titre aussi bien les informaticiens de LIA – volet "numérique" – que les chercheurs associés à ECC ou à ICTT – volet "culture".

[Note : L'interface BNCweb est une plateforme web en libre accès intégrant de nombreuses modalités d'interrogation du British National Corpus. L'interface CQPweb utilise cette architecture, qu'elle étend à d'autres corpus en accès libre.]

- Axe 2 : Réseaux sociaux, structures, contenus et usages.

De nombreux linguistes s'interrogent désormais sur les spécificités de certains types de communication médiée par le numérique. L'analyse quantitative de ce type de discours vise souvent à en extraire des éléments de contenu. L'association des deux perspectives pourront ouvrir la voie à de nouvelles modalités de recherche, en ciblant par exemple, des schèmes argumentatifs, *via* des recherches complexes qui associerait informations lexicales, grammaticales (aspect, modalité, voix) ou syntaxiques (ordre, proximité, répétition).

7 – Partenariats extérieurs (en cours et à venir)

Pas de partenariat extérieur au moment de la rédaction de ce document.

8 – Budget prévisionnel et tous les financements envisagés

DEPENSES		**RECETTES**	
Nature	**Montant**	**Origine**	**Montant**
Frais généraux : (à détailler)	€ 120 (édition du programme)	Acquises	
Déplacements :	€ 700	Sollicitées:	
Hébergement :	€ 480	ICTT SFR Agorantic Université	€ 1500 € 750 € 750
Repas, pause-café, … :	€ 1500		
Activités extra colloque: Autres :	€ 200	Droits d'inscription attendus :	
TOTAL 1	**€ 3000**	**TOTAL 2**	**€ 3000**

9 – Références bibliographiques (en appui du texte)

Aijmer, Karin & Christoph Rühlemann (eds.). 2014. Corpus pragmatics: a handbook. New York: Cambridge University Press.

Baker, Paul & Tony McEnery. 2015. Corpora and Discourse Studies: Integrating Discourse and Corpora (Palgrave Advances in Language and Linguistics). London: Palgrave Macmillan.

Biber, Douglas. 2009. Register, genre, and style. (Cambridge Textbooks in Linguistics). Cambridge, UK ; New York: Cambridge University Press.

Cappeau Paul, Hélène Chuquet & Freiderikos Valetopoulos (eds.). 2010. L'exemple et le corpus, quel statut ? Rennes: Presses universitaires de Rennes.

Evert, Stefan. 2006. "How Random is a Corpus?: The Library Metaphor." Zeitschrift für Anglistik und Amerikanistik 54(2). 177–190. doi:10.1515/zaa-2006-0208.

Glynn, Dylan. 2014. "Techniques and tools. Corpus methods and statistics for semantics." doi:10.13140/RG.2.1.1047.1842. http://dx.doi.org/10.13140/RG.2.1.1047.1842 (8 November, 2015).

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee, Ylva Berglund Prytz. 2008. Corpus linguistics with BNCweb: a practical guide. (English Corpus Linguistics v. 6). Frankfurt am Main: Peter Lang.

Hunston, Susan. 2001. "Colligation, lexis, pattern, and text." In Mike Scott & Geoff Thompson (eds.), *Patterns of Text*, 13–33. Amsterdam: John Benjamins Publishing Company. https://benjamins.com/catalog/z.107.03hun (27 May, 2015).

Legallois, Dominique. 2012. "La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ?" Corpus (11). http://corpus.revues.org/2202.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. Corpus-based language studies: an advanced resource book. New York: Routledge.

Stubbs, Michael. 1995. "Collocations and semantic profiles: On the cause of the trouble with quantitative studies." Functions of Language 2(1). 23–55. doi:10.1075/fol.2.1.03stu.

10 – Annexes

Annexe 1 :

Appel à communications (version anglaise)

Call for papers :

It is historically the case that linguists interested in pragmatic phenomena -- whether in terms of speech act theory, relevance theory, argumentation, enunciation, etc. -- have not always made use of corpora or, when they have done so, have used them in a spirit of "illustrative eclecticism" (Kohnen 2015:56 [1]), drawing on genuine examples to illustrate a theoretical perspective. In recent years various semasiological approaches have considered how the apparent polysemies of a given marker can be fruitfully explained in terms of complex configurations, as the postulated semantic potentials of cooccurring items interact to generate contextually situated values. Such an approach opens new perspectives for quantitative exploitation of corpus material in theories which have traditionally stressed qualitative analysis of a limited number of cases. The technology of corpus enquiry now enables us to conduct sophisticated searches in terms of collocational affinities, bearing on data which can -- according to how the corpus is tagged and marked up -- involve linguistic and paralinguistic features both (pauses, overlaps, speakers' identities, textual genres etc.).

The aim of the present conference is therefore to investigate the usefulness and relevance of the quantitative analysis of corpus data of English and Englishes, in those areas of linguistic research where a more qualitative, fine-grained approach has traditionally prevailed. Issues addressed might include:

- semantico-pragmatic profiling of specific markers or configurations of markers via quantitative corpus-based data;
- preprocessing requirements on corpora with a view to such research, i.e. specific or ad hoc tagging or mark up conventions;
- types of corpus enquiry, tools, syntax and/or algorithms relevant to such research.

In keeping with the programme of our local research group Identité Culturelle, Textes et Théâtralité, studies that address issues of corpus approaches to linguistic identities are also encouraged.

Conference languages: English and French.

[1] Kohnen, Thomas. 2014. "Speech Acts: a diachronic perspective", in Corpus Pragmatics. A Handbook. Aijmer, K. and Rühlemann, C. (eds.), Cambridge University Press.

Annexe 2 :

Comité scientifique :

Jean Albrespit, Université Bordeaux Montaigne, TELEM EA 4195

Agnès Celle, Université Paris Diderot, CLILLAC-ARP EA 3967

Guillaume Desagulier, Université Paris 8, MoDyCo UMR 7114

Lionel Dufaye, Université Paris-Est Marne-La-Vallée, LISAA EA 4120

Richard Dufour, Université d'Avignon, EA LIA 4128

Anika Falkert, Université d'Avignon ICTT EA 4277

Lucie Gournay, Université Paris-Est Créteil Val de Marne IMAGER EA 3958

Sebastian Hoffmann, Universität Trier, Allemagne

Rudy Loock, Université Charles de Gaulle, Lille 3 STL UMR 8163

Jean-Marie Merle, Université de Nice Sophia Antipolis, BCL UMR 7320

Blandine Pennec, Université Toulouse Le Mirail, CAS EA 801

Graham Ranger, Université d'Avignon, ICTT EA 4277

Antoinette Renouf, Birmingham City University

Martine Sekali, Université Paris 10, CREA EA 370