

Résumés

Regards nouveaux sur les liens entre théories, méthodes et données en linguistique française

8 au 10 septembre 2011
Nancy, Palais des Congrès

Colloque annuel de
l'Association for French Language Studies
organisé à l'occasion de ses 30 ans



contact : afls2011@atilf.fr



Nancy-Université



Conférences plénières

Lou BURNARD - Oxford University Computing Services (1975-2010) (Royaume-Uni) & TGE ADONIS (France)

Qu'est-ce que l'annotation du corpus et pourquoi en parle-t-on de manière si inquiétante ?

In this introductory talk, I will discuss the purpose, practice, and consequences of annotation in the construction of language corpora. Some have argued that a language corpus should be entirely free of additional annotation, that the corpus should, in John Sinclair's phrase, "speak for itself". Others have argued that a corpus may actually consist of nothing but annotation. I will attempt to review some of the wide range of current practices falling between these two extreme positions. I will suggest that the elastic term "annotation" can be applied to a wide range of interpretative acts applied to linguistic materials, often responding to very different needs, and that there has consequently been a tendency to implement these using very different styles or mechanisms. I will argue for a generic annotation scheme that can encompass the full range of such requirements.

Jenny CHESHIRE - Queen Mary, University of London (Royaume-Uni)

Language contact and language change in European cities

Recent large-scale patterns of immigration to many of the urban centres of Europe have resulted in new 'multiethnolects' that owe their form either directly or indirectly to language contact. Relevant research has been carried out in multilingual cities in Scandinavia, the Netherlands, Germany, London and is currently ongoing in Paris. In many cases this research shows that linguistic innovations begin in the speech of ethnic minorities and may then diffuse to monolingual speakers of the local vernaculars.

In this talk I will mainly focus on innovations in grammar and discourse in 'Multicultural London English': a variable repertoire of distinctive features used by monolingual and bilingual speakers alike. I will discuss the sociohistorical context in which these innovative features have emerged, comparing this context with those in which other contact varieties have developed (such as pidgin and creole languages, and indigenized post-colonial varieties) and arguing that the urban multiethnolects of Europe constitute a typologically distinct language contact variety. I will present some analyses of the social and linguistic factors that account for the emergence and diffusion of some of the distinctive features of Multicultural London English, suggesting that these analyses can contribute to our more general understanding of the processes leading to the actuation of linguistic change.

Tom COBB - Université du Québec à Montréal (Canada)

A Wish-List for Data-Driven Learning in French

Input alone may be inadequate to expose the patterns of a second language, but input (data) explored with a computer program seems a useful addition. Data-driven language learning

relies on the development of both computer programs, which in principle will work on any language, and a learning-friendly analysis of particular languages, which has been accomplished to varying degrees. English has been particularly privileged owing to the work of both its linguists and its pedagogical adaptors. My plenary talk will show examples of how the data of English has been made learner ready and set some goals for a parallel treatment of French.

Bernard COMBETTES - ATILF UMR 7118 CNRS, Nancy-Université (France)

Linguistique diachronique et corpus : problèmes théoriques et méthodologiques

En s'attachant essentiellement aux aspects théoriques et méthodologiques, on essaiera d'évaluer l'importance des principaux changements que le développement des grands corpus informatisés a entraînés dans le domaine de la linguistique historique. Si certaines questions sont communes à l'étude des états de langue anciens et à celle de la synchronie de la langue moderne, en particulier en ce qui concerne la représentativité des corpus, d'autres sont propres à l'étude historique. Parmi les points positifs, qui marquent une rupture nette avec les pratiques plus anciennes, on mettra en avant la possibilité de travailler sur des diachronies larges et de mettre plus facilement en relation divers paramètres, ce qui permet une nouvelle approche de la variation, considérée alors comme moteur du changement et non comme une de ses conséquences. La conception même de la périodisation se trouve également modifiée (on prendra comme exemple le moyen français et le français préclassique). Une plus grande souplesse est obtenue dans la modification, en cours d'étude, des corpus de travail, ce qui était plus difficilement réalisable. On signalera enfin, comme conséquence positive, les modifications du fonctionnement de la communauté scientifique : l'utilisation de l'outil informatique a permis à des non spécialistes de la diachronie, qui ne se seraient pas aventurés dans la lecture des textes, de se pencher, davantage qu'ils ne l'auraient fait auparavant, sur des faits de la langue ancienne dans le but d'éclairer les usages actuels. L'élaboration des grands corpus a ainsi incontestablement favorisé le travail d'équipe.

On insistera également sur quelques points plus négatifs, qu'il conviendrait de prendre en compte dans un emploi "raisonné" des ressources informatisées : les études se trouvent trop souvent limitées aux phénomènes qui se prêtent facilement à la recherche automatisée, d'où l'abondance de travaux sur des expressions, des formes identifiables après lemmatisation, et la mise à l'écart de nombreux sous-domaines de la syntaxe. L'observation des résultats fournis par l'outil informatique éloigne par ailleurs de la lecture linéaire, continue ; l'appréhension syntagmatique est remplacée par une appréhension paradigmatique, qui, si elle fait apparaître des régularités, empêche la prise en compte d'un contexte large (non seulement du cotexte, mais, plus généralement, du type de production textuel). Il faut enfin signaler les problèmes posés par l'indexation et l'étiquetage des corpus. On pourrait considérer que cette question se pose également pour les corpus de langue moderne ; la diachronie ajoute toutefois une difficulté supplémentaire, dans la mesure où les catégories, les fonctions syntaxiques, les unités (comme la phrase ou la proposition, par exemple) sont elles-mêmes en évolution, ce qui multiplie les dangers d'"anachronisme" dans le codage. On ajoutera que l'évolution concerne aussi les types de textes et les genres et que, dans ce domaine aussi, il conviendrait de relativiser la plupart des notions utilisées.

Communications

Hervé ADAMI - ATILF UMR 7118 CNRS, Nancy-Université (France)

Limites théoriques et méthodologiques de l'analyse des corpus d'interactions verbales : retour vers Bakhtine

L'analyse des corpus peut poser des problèmes pratiques, méthodologiques et théoriques parce que les interactions transcrites sont décontextualisées. L'absence de données concernant les interactants, leurs positions sociales, leur histoire conversationnelle, les finalités de l'interaction ou la situation précise dans laquelle ils se trouvent, laisse apparaître de nombreuses formes d'implicites dont le sens échappe à l'analyse. Au risque d'en rester à une analyse de la mécanique interactionnelle, il semble opportun de repenser un cadre théorique en s'appuyant sur Bakhtine et ses concepts de polyphonie ou son approche de l'énonciation et de l'énoncé entendus comme des instances interactionnelles connectées à la réalité sociale dont elles procèdent.

Parastoo AGHAJANZADEH-DARZI - Université Sorbonne Nouvelle - Paris 3 (France) [PRIX DOC]

Perception et intégration de l'image dans le cadre d'un enseignement / apprentissage du F.L.E.

Notre communication porte sur l'intégration des images dans les méthodes de FLE, en soulignant ses avantages mais aussi ses limites. Cette recherche, en se basant sur le concept d'"image" en tant que "support visuel didactique" dans l'enseignement-apprentissage des langues-cultures, cherche à cerner dans quelle mesure l'application d'une image favorise l'apprentissage d'une langue chez les apprenants.

Notre étude empirique a été effectuée au sein du DULF (Diplôme Universitaire de Langue Française) à l'université de la Sorbonne à Paris. De ce fait, nous avons recueilli des données en observant les interactions et les productions orales et écrites d'un public étudiantin « multiculturel » face aux images.

Nous esquissons des pistes qui nous permettront d'explorer le besoin des apprenants, la configuration méthodologique de l'étude et les outils de recueil de données.

Virginie ANDRÉ - ATILF UMR 7118 CNRS, Nancy-Université (France)

« Ton corpus, j'y comprends rien ! » : de la difficulté de partager des données orales

L'accès à des corpus oraux est de plus en plus facilité. Néanmoins, apparaissent des problèmes méthodologiques, d'une part, et théoriques, d'autre part, liés à l'exploitation de ces données partagées. L'intérêt de partager des données est notamment d'aboutir à un grand corpus de français parlé diversifié, et de permettre des analyses partageables et vérifiables. Cependant, ces données n'ont initialement pas été recueillies et traitées dans le but d'intégrer de potentielles analyses. Nous montrerons quelles sont les difficultés qu'un chercheur peut rencontrer lors de l'analyse d'un corpus constitué par un autre chercheur

dans un cadre théorique, méthodologique et analytique particulier. Les données situationnelles accompagnant les données primaires ne sont pas, peu ou rarement explicitées (les métadonnées sont souvent insuffisantes). La simple compréhension des données est alors difficile puisque la situation de communication dans laquelle se déroulent les interactions n'est que peu connue.

Virginie ANDRÉ¹, Christophe BENZITOUN¹, Emmanuelle CANUT¹, Jeanne-Marie DEBAISIEUX², Bertrand GAIFFE¹ - ¹ATILF UMR 7118 CNRS, Nancy-Université (France), ²Université Sorbonne Nouvelle - Paris 3 (France)

Le projet TCOF (Traitement de Corpus Oraux en Français) : enjeux pour l'étude des pratiques langagières

Le projet TCOF (« Traitement de Corpus Oraux en Français », ATILF) rassemble des données orales authentiques : conversations entre adultes (récits divers, réunions de travail en entreprise, entretiens...), entre adulte et enfant de moins de sept ans en cours d'apprentissage du langage (conversations libres, narrations à partir de livres illustrés). Nous présenterons tout d'abord les différentes étapes du projet, avant de nous intéresser aux questions actuelles liées *in fine* à l'analyse de ces dialogues de l'oral « spontané » chez l'adulte et chez le jeune enfant.

Marta ANDRONACHE - ATILF UMR 7118 CNRS, Nancy-Université (France)

Une édition critique de texte : liens entre sources informatisées, bases de données numériques, sources spécialisées en ligne et supports informatiques

Nous voudrions retracer le chemin d'une entreprise scientifique qui se propose de faire le lien entre l'utilisation de ressources linguistiques et informatiques, dans le cadre du programme de recherche ERC : « Histoire du lexique politique français ».

Le but de notre communication est de montrer le travail mené pour la réalisation de la double édition de texte – électronique et papier – de la première traduction en français de la Cité de Dieu de saint Augustin, traduction réalisée par Raoul de Presles (1371-1375). Notre édition de texte repose sur l'examen d'une riche panoplie de sources linguistiques informatisées et sur l'utilisation de moyens informatiques adaptés. Dans ce cadre, nous aborderons des questions pratiques (encodage, lemmatisation), pour finir avec quelques aspects linguistiques : nous nous proposons d'illustrer nos propos avec une série de néologismes introduits en français par cette traduction, lexèmes que nous avons pu repérer aussi grâce à l'utilisation de supports informatiques.

Sílvia ARAÚJO, Conceição VARELA, Natália FUMEGA - Universidade do Minho, Braga (Portugal)

Linguistique de corpus et traduction des constructions françaises en se voir + INF/PP en portugais et en espagnol

Dans le cadre de cette communication, nous tenons, tout d'abord, à présenter le projet *Perfide* qui vise la compilation, avec des textes appartenant à différents domaines (littéraire, journalistique, technico-scientifique, juridique et religieux), de corpora parallèles entre le portugais et six autres langues (l'espagnol, le russe, le français, l'italien, l'allemand et l'anglais), dans le but de les rendre librement accessibles aux linguistes, enseignants, étudiants et traducteurs (Araújo *et al.* 2009).

Nous entendons proposer ici une réflexion sur l'utilité d'un corpus multilingue dans les domaines touchant à l'étude du langage et de la traduction (Granger *et al.* 2003), à la fois d'un point de vue théorique et sur la base d'un exemple concret, celui des constructions en *se voir + infinitif/PP* (Shyldkrot 1981) et ses constructions *homologues* en portugais et en espagnol. [...]

Nathalie AUGER, Christelle DODANE, Guillaume ROUX, Jérémie SAUVAGE - DIPRALANG - EA 739, Université Paul Valéry - Montpellier 3 (France)

Fonctionnements mimétiques dans l'usage de marqueurs discursifs de la mère à l'enfant

Nous nous proposons de dépasser le clivage entre approche qualitative et quantitative dans l'étude d'un vaste corpus longitudinal rassemblant les productions spontanées de 8 enfants filmés mensuellement entre l'âge de 12 et 48 mois, en interaction avec leurs mères (corpus NIMH de Lyon et corpus de Paris). Nous nous concentrerons sur l'usage pragmatiqué des marqueurs discursifs « *enfin* » et « *en fait* » en nous demandant si ce type d'usage pourrait se transmettre à l'enfant via le langage de la mère. L'étude des interactions mère-enfant nous permet en effet de suivre l'évolution de ces phénomènes de mimétisme de l'enfant en fonction de l'âge et des stades de l'acquisition du langage. Nous présenterons les différentes étapes de notre analyse, tout en nous questionnant sur les problèmes méthodologiques liés au fait de concilier un traitement de type automatique et quantitatif, à une approche qualitative des usages étudiés.

Fabienne BAIDER¹ & Évelyne JACQUEY² - ¹Université de Chypre (Chypre), ²ATILF UMR 7118 CNRS, Nancy-Université (France)

Comparaison entre les résultats obtenus entre une approche distributionnelle et une approche lexicométrique pour l'analyse des possibles a priori dans un corpus journalistique

Des analyses distributionnelles ont conclu à des prises de position des journalistes en faveur des candidats les plus chevronnés au premier tour de la campagne présidentielle de 2007, hommes ou femmes, sauf dans le cas où les enjeux étaient de taille (Baider & Jacquy 2008 et 2010). Parallèlement, d'autres recherches ont démontré comment la prise en compte de

choix stylistiques, thèmes récurrents, références historiques étaient *spécifiques* à certaines idéologies (Valette & Rastier 2006 ; Reutenauer et al. 2010 ; Valette 2010).

L'objet de la présente proposition est d'examiner si effectivement le critère de *spécificité* permet d'obtenir les mêmes résultats que ceux obtenus lors d'une prise en compte d'emplois linguistiques. Pour accéder à l'information linguistique relevant du niveau du texte, nous appliquons une approche textométrique telle qu'elle est définie dans Brunet (1993) et mise en œuvre à l'aide du logiciel *Termostat* (Drouin 2003 ; Lesage et al. 1993) le corpus ayant déjà été étudié préalablement de manière distributionnelle.

Luciana Beatriz BASTOS AVILA¹, Priscila OSÓRIO CÔRTEZ², Raíssa VITÓRIA CAETANO³, Heliana RIBEIRO DE MELLO⁴ - ¹PosLin -UFMG/UFV (Brésil), ²FALE-UFMG/ Santander (Brésil), ³FALE-UFMG/CNPq (Brésil), ⁴FALE-UFMG/ CNPq/Fapemig (Brésil)

Searching for modality – new methodologies for the study of modal indexes

Modality is traditionally defined, from a semantic point of view, as the speaker's "opinion or attitude towards the proposition that the sentence expresses or the situation that the proposition describes" (Lyons 1977: 452). It is usually codified by different indexes that express the speaker's degree of certainty or commitment towards a given propositional content or locution. Those can be realized as grammatical or morphological categories such as future markers, conditionals, modal verbs, adverbs, belief verbs, periphrastic forms, evaluative adjectives in nominal predicates etc. However, as new technologies foster the compilation of oral corpora, it has become evident that the pursuit of modal indexes at the sentential level is but one small portion of the work that needs to be done. [...]

Olivier BAUDE - Laboratoire Ligérien de Linguistique - EA 3850, Université d'Orléans (France) & DGLFLF (Délégation générale à la Langue française et aux Langues de France)

Le programme Corpus de la parole

Le programme *Corpus de la parole* initié en 2005 par la Délégation générale à la langue française et aux langues de France du Ministère de la Culture en partenariat avec le CNRS a pour vocation la constitution, la conservation, l'exploitation et la diffusion de corpus oraux, en français et langues de France.

Ce programme s'est concrétisé d'une part par un travail sur les *bonnes pratiques* et d'autre part, par la réalisation d'une base de données conçue comme un réservoir accueillant, après cinq années d'existence, plus de mille heures d'enregistrements en quarante deux langues (ainsi que des annotations et des transcriptions) provenant de projets scientifiques aux pratiques et objectifs très diversifiés. [...]

Olivier BAUDE¹⁻², Céline DUGUA¹ - ¹Laboratoire Ligérien de Linguistique - EA 3850, Université d'Orléans (France), ²DGLFLF (Délégation générale à la Langue française et aux Langues de France)

Les ESLOs : la variation en réserve

Notre communication vise à porter un regard réflexif sur la prise en compte de la *variation* dans le programme de constitution et d'exploitation des Enquêtes sociolinguistiques à Orléans. La compilation de ces deux enquêtes constituera à terme le plus grand corpus de français parlé disponible avec 10.000.000 de mots pour plusieurs centaines d'heures d'enregistrement. La mise à disposition de la plus grande partie de ce corpus, à partir du milieu de l'année 2011 et après presque dix ans de travail consacrés aux opérations de numérisation, collecte, enregistrement, transcription, annotation, constitution d'une base de données, élaboration d'outils de requête – sans oublier les aspects juridiques et éthiques – offre l'occasion d'interroger les choix techniques et théoriques mis en œuvre. [...]

Kate BEECHING - University of the West of England, Bristol (Royaume-Uni)

Indexicality and heteroglossia in French/English translation

In a series of recent books, chapters and articles, Aijmer and others (2006a and b and 2007) have proposed the use of translation as a useful heuristic in the exploration of the semantic status of pragmatic markers. By pragmatic markers, I refer to such terms in French as *bon, enfin, hein, quand même* and *quoi*. The meaning of pragmatic markers is notoriously hard to pin down: they are semantically bleached and often multifunctional. Aijmer et al. (2006b: 105-111) usefully highlight their functions in three domains: reflexivity (their ability to allow the speaker to reformulate or reflect on their wording), indexicality (the way that they point to different aspects of the immediate context – or the wider context, such as their indication of social categories) and heteroglossia (the stance or attitude of the speaker to other people and to other opinions). [...]

Bernard BEL, Médéric GASQUET-CYRUS - Laboratoire Parole et Langage UMR 6057 CNRS, Université de Provence (France)

Interdisciplinarity and the sharing of oral data open new perspectives to field linguistics

Our laboratory is engaged in resource collection, analysis and theorizing (socio)linguistics with particular focus on links between experimental and field approaches. Two projects dealing with endangered languages on the border area of Provençal and Francoprovençal (Valjouffrey and Valbonnais, Isère) are addressing this issue. Their fieldwork builds on multichannel recordings and video coverage allowing accurate studies of interactions.

Dealing with large amounts of data beyond the scope of projects is justified by their sharing in archives hosted by the Resource Centre for the Description of Oral (CRDO-Aix, <http://crdo.fr>). Resource pooling contributes to mobilizing amateurs and professionals handing over unpublished data for its preservation and non-commercial distribution. This phenomenon empowered our informants who became members of the team in full right,

setting new priorities on the research agenda: the design of a script for their revitalized language and an inventory of place names that delineate their living space.

L'interdisciplinarité et la mutualisation des données orales ouvrent de nouvelles perspectives à la linguistique de terrain

Notre laboratoire est dans une phase de collection de ressources, d'analyse et de théorisation (socio)linguistique visant à renforcer les liens entre approches expérimentales et de terrain. Deux projets s'intéressent à des langues menacées de disparition à la lisière des aires occitane et francoprovençale (Valjouffrey et Valbonnais, Isère). Le travail sur le terrain bénéficie d'une haute technicité (audio multipistes et vidéo) permettant l'analyse précise des interactions.

Collecter des données au delà des besoins projetés justifie leur partage dans une archive hébergée par le Centre de ressources pour la description de l'oral (CRDO-Aix, <http://crdo.fr>). La mutualisation des ressources conduit des amateurs et des professionnels à soumettre des données inédites pour assurer leur pérennité et leur libre distribution. Ce phénomène contribue à donner à nos informateurs le statut de chercheurs dans une équipe dont ils réorientent les priorités : la définition d'une graphie pour leur langue revitalisée et un inventaire des toponymes qui délimitent leur espace de vie.

Jean-David BELLONIE¹, Françoise GADET², Emmanuelle GUÉRIN³ - ¹Université des Antilles et de la Guyane (France), ²MoDyCo UMR 7114 CNRS, Université Paris Ouest Nanterre La Défense (France), ³Laboratoire Ligérien de Linguistique - EA 3850, Université d'Orléans (France)

Réflexions autour d'un corpus : des faits émergents dans le contact

Cette communication a pour cadre le projet ANR/ESRC MLE-MPF, qui a pour objectif la comparaison de deux capitales occidentales (Londres et Paris) du double point de vue sociolinguistique et écolinguistique d'influence des langues de l'immigration et des changements multiculturels pouvant affecter la langue nationale. Au-delà des aspects linguistiques, la visée ultime concerne des enjeux sociaux en rapport avec la migration, l'intégration et leurs conséquences éducatives.

L'arrière-plan concerne des interrogations sur la mesure où les processus du changement linguistique dépendent des langues et des contextes où elles sont mises en œuvre. Quels sont les effets sur une langue d'être parlée dans des métropoles à fort brassage multiculturel (processus de véhicularisation, nivellement, koïnésation..., mais aussi faits plus directement liés au contact) ? [...]

Karima BEN ABBES - University of Essex (Royaume-Uni) [PRIX DOC]

The acquisition of Gender/Number Concord and Definiteness in Early L3 French interlanguage

This paper investigates, within a generative paradigm, the acquisition of *gender/number* concord and *Definiteness* in early L3 French syntax by native Turkish and Spanish who have English as an L2.

Four hypotheses in the literature will serve as the framework of this research (i) L1 is the source of CLI in L3A, (ii) L2 is the source of influence, (iii) both L1 and L2 can be influential and finally (iv) the language which is “typologically” the closest to the L3 is the most influential.

This study partially agrees with all these hypotheses. However, our main prediction is that CLI is highly guided by the learner’s perception and linguistic judgments towards the syntactic property itself. Hence, the present work aims to test the tenability of these hypotheses, and compare them with the prediction of the present study which has been named the *perceived property-based hypothesis* (PPBH).

Christophe BENZITOUN, Lolita BÉRARD - ATILF UMR 7118 CNRS, Nancy-Université (France)

Premiers tests en vue de l’étiquetage d’un corpus oral par apprentissage automatique exclusivement endogène.

Nous souhaitons étiqueter en parties du discours et lemmatiser une banque de données orales de plus de deux millions de mots. Or, les corpus oraux présentent des spécificités qui ne sont pas prises en compte par les étiqueteurs morphosyntaxiques classiques. Pour ce faire, nous voulons développer dans Treetagger un fichier de paramétrage spécifique aux corpus de français parlé, obtenu par apprentissage automatique. Le principe général consiste à faire un premier passage avec Treetagger puis à corriger manuellement les étiquettes. Nous avons mis en œuvre une série d’heuristiques afin d’accélérer le travail de correction. Quand le volume de transcriptions vérifiées atteint un seuil (10.000 mots), nous lançons l’apprentissage à l’aide du module TrainTreetagger et évaluons les résultats obtenus. Nous comptons reproduire ce processus jusqu’à obtenir des résultats convenables puis diffuser les fichiers paramètres et les corpus étiquetés. Ce travail est mené, pour l’instant, sur les transcriptions d’adultes du projet TCOF.

Lolita BÉRARD - ATILF UMR 7118 CNRS, Nancy-Université (France) [PRIX DOC]

Corpus et relation lexique-grammaire dans les contrôles à distance

Pour révéler les liens entre lexique et grammaire, il est nécessaire d’avoir des corpus de très grande taille. Dans les constructions de dépendance à distance, du type “c’est génial quoi faire un truc que tu aimes quoi et que [tu *sais* que] tu maîtrises” les corpus semblent montrer que le lexique du verbe pont est restreint. Néanmoins, certains verbes tels que *aimer* ou *craindre*, sont absents, ce qui paraît contre intuitif. Sur le web, ces verbes, et beaucoup d’autres, sont attestés en verbe pont. Nous nous demanderons pourquoi il y a une

telle différence entre les corpus et le Web et dans quelle mesure nous pouvons parler de relation lexicque-grammaire.

Tiphanie BERTIN - ATILF UMR 7118 CNRS, Nancy-Université (France) [PRIX DOC]

Complémentarité des approches quantitatives et qualitatives dans l'étude du développement langagier.

Dans une étude sur le processus de développement des articles et des clitiques sujets chez des enfants français, nous soulignons l'importance de la complémentarité entre approche quantitative et qualitative des données pour rendre compte du rôle de l'interaction adulte-enfant dans ce processus d'appropriation. L'analyse quantitative de l'évolution du langage de l'enfant et de l'interaction fait ressortir une trame générale du développement des éléments et la progression vers des productions autonomes chez l'enfant. Une observation attentive des corpus permet de constater une évolution non linéaire, où l'enfant met en place des systèmes transitoires. Une approche qualitative des données permet de montrer comment ces systèmes se modifient progressivement dans l'interaction avec l'adulte en illustrant la façon dont se déroulent les échanges entre l'enfant et l'adulte.

Hélène BLONDEAU - University of Florida (Etats-Unis)

Regards croisés sur deux variables de l'irréalis : l'apport des données longitudinales

Cette communication présente une étude longitudinales de la variation dans l'emploi des formes verbales de deux variables du domaine de l'irréalis : la référence temporelle au futur et la protase des hypothétiques en *si*. Les recherches antérieures sont remises en perspective à partir de données longitudinales recueillies à Montréal auprès de mêmes individus au cours d'une période de 24 ans. L'analyse de la variation montre que plutôt que de prendre part au changement, les individus de la cohorte retiennent leur usage des variantes conservatrices et montrent une constance dans les contraintes linguistiques affectant la variation au fil du temps. Ce comportement tranche avec leur adoption pourtant rapide du changement sociophonétique ayant affecté la prononciation du (r) à Montréal. Cette étude longitudinale permet de documenter la contribution des individus aux changements linguistiques affectant la communauté tout entière.

Jacques BRES¹, Emmanuelle LABEAU² - ¹Praxiling UMR 5267 CNRS, Université Paul Valéry - Montpellier 3 (France), ²Aston University, Birmingham (Royaume-Uni)

Venir de et l'effet de sens de récence

[...] La présente recherche, dans le cadre d'un travail plus général sur la grammaticalisation des formes itive (*aller*) et ventive (*venir*) en français et dans les langues romanes et en appui sur un corpus nombreux d'occurrences authentiques, se propose :

- d'établir que *venir de* peut être actualisé non seulement au présent et à l'imparfait et contrairement à *aller* futur, également au futur (contra Golian 1979 et Schogt 1968) et au conditionnel (contra Vet 1978, 1980), voire au participe présent et au subjonctif présent (Vetters 1989, Havu 2005) [...]

Sylvie BRUXELLES, Carole ÉTIENNE, Lorenza MONDADA, Florence OLOFF, Sandra TESTON, Véronique TRAVERSO - ICAR UMR 5191 CNRS, Université de Lyon (France)

Doubles contraintes et injonctions paradoxales: (mais pourquoi faites-vous des) corpus et bases de données de LPI !

[...] Notre objectif est de montrer comment il peut être possible de développer une base de données

- dédiée à un type de recherche spécifique,
- intégrant les possibilités d'autres recherches en linguistique et en SHS,
- conjuguant les possibilités nouvelles offertes par la base (en termes de quantité de données traitées, en termes d'outils et de systématisme) avec le travail du chercheur qui continue à faire du terrain et à confectionner ses propres corpus. Ces deux pratiques sont en effet essentielles pour l'avancée de la recherche.

Laure BUDZINSKI - ATILF UMR 7118 CNRS, Nancy-Université (France) [PRIX DOC]

Exploitation des données numérisées pour un travail de recherche en étymologie

Cette communication montrera l'avantage d'une utilisation des plates-formes d'ouvrages numérisés (comme la recherche de livres sur Google) pour l'étymologiste. Nous nous appuierons sur plusieurs exemples concrets, dont *dialectologie* et *phonologie*, afin de constater quelles sont les possibilités de recherche offertes par ces outils internet en accès libre : large choix de littérature permettant l'antédation des lexèmes jusqu'à présent décrits, possibilité d'une comparaison inter-linguistique dans le cadre d'un emprunt, possibilité d'élargissement à d'autres domaines sectoriels, faculté d'orienter ses recherches afin d'asseoir la description sémantique d'un vocable... Nous parlerons aussi de quelques-unes des lacunes et limites des plates-formes d'ouvrages numérisés que l'étymologiste doit connaître.

Emmanuelle CANUT¹, Martine VERTALIER² - ¹ATILF UMR 7118 CNRS, Nancy-Université (France), ²Fédération CLESTHIA, Université Sorbonne Nouvelle - Paris 3 (France)

Un grand corpus français d'interaction entre adulte et enfant : pour quoi faire en acquisition ?

A partir d'un corpus d'interaction adulte-enfant de 8h30 (76.000 mots), nous souhaitons vérifier certains des résultats issus d'analyses qualitatives en linguistique de l'acquisition et les confronter aux résultats issus d'analyses quantitatives en psycholinguistique

développementale pour voir les points de convergence et divergence. Nous nous intéresserons particulièrement à l'appropriation de constructions syntaxiques complexes (des relatives et des complétives).

Janice CARRUTHERS - Queen's University Belfast (Royaume-Uni)

La construction d'un corpus de contes oraux : objectifs, méthodologie et exploitation

Cette communication concerne la construction et l'exploitation d'un corpus de narrations orales. Après une présentation des objectifs du corpus, il y aura une discussion sur la méthodologie employée pour la sélection des données (sélection des participants, des performances et des conte-types) ainsi que pour l'encodage linguistique. Je présenterai ensuite une analyse de deux phénomènes linguistiques, c'est-à-dire l'encadrement temporel et l'inversion sujet-verbe, en essayant de situer l'analyse de ces phénomènes linguistiques dans le contexte discursif d'une performance orale.

Marie CHAGNOUX¹, Marion LAIGNELET² - ¹CREM - EA 3476, Université Paul Verlaine - Metz (France), ²CLLE-ERSS UMR 5263 CNRS, Université Toulouse - Le Mirail (France)

Appréhender le discours: relations hiérarchiques et linéaires en corpus

La nécessité d'accéder rapidement au contenu sémantique de données numériques de plus en plus volumineuses a marqué un tournant dans le domaine du traitement automatique des textes dans les années 90 en favorisant le recours aux méthodes d'apprentissage basées sur des modèles probabilistes où le texte est considéré comme un ensemble d'unités. En parallèle, le champ de la linguistique discursive se développait à partir de l'idée fondamentale que le sens résidait davantage dans les relations entre ces unités.

Dans notre présentation, nous interrogeons la possibilité d'une représentation formalisée des structures textuelles qui puisse être opératoire pour le traitement de larges corpus : comment construire une représentation qui articule différents concepts théoriques issus de la linguistique discursive et une mise en œuvre par des outils informatiques ?

Michel CHAROLLES - LaTTiCe UMR 8094 CNRS ENS, Université Sorbonne Nouvelle - Paris 3 (France)

Des données attestées aux données provoquées : complémentarité des études sur corpus et des études expérimentales

La communication proposée porte sur le fonctionnement des adverbiaux en discours et, plus spécifiquement, sur le rôle qu'ils peuvent assumer dans la structuration des informations textuelles quand ils sont détachés en tête de phrase, en position cadrative. La communication vise globalement à étayer l'idée que, pour tester des hypothèses de ce type, les données que l'on peut tirer de corpus d'emplois attestés doivent être complétées par des données expérimentales provoquées (comportementales et/ou de neuro-imagerie). Ce point

sera illustré à partir de données sur le fonctionnement de groupes prépositionnels temporels et spatiaux.

Kenza CHERKAOUI MESSIN - DiLTeC - EA 2288, Université Sorbonne Nouvelle - Paris 3 (France)

Construction et exploitation d'un corpus hétérogène : communauté discursive et circulation des discours

L'activité discursive des hommes politiques est nécessairement hétérogène : débattre en sessions parlementaires, écrire des lois, rendre des rapports ou des avis écrits, les présenter à l'oral, interpellier le Gouvernement, s'adresser aux citoyens au travers des médias sont autant d'occasions de « prendre la parole », à l'écrit ou à l'oral.

Analyser le « dire » de cette « communauté discursive » (Beacco 1999 : 14) revient à suivre les personnels politiques au travers de leurs différents lieux d'expressions et à comprendre les conditions de production de ces discours (Charaudeau 2005 : 11-35). Les personnes participant à une même communauté discursive produisent des discours que l'on peut lire de manière topologique (Beacco & Moirand 1995), c'est à dire en identifiant les différents lieux d'expression pour cette communauté et en envisageant ses discours comme circulant à l'intérieur ou vers l'extérieur de la communauté. [...]

Marie-Hélène CONDETTE, Rafael MARÍN, Aurélie MERLO - Savoirs Textes Langage UMR 8163 CNRS, Université de Lille 3 (France)

La structure argumentale des nominalisations : de la théorie au corpus

Malgré le vif intérêt soulevé par la structure argumentale (SA) des nominalisations, aucun outil pour l'exploitation en corpus de ce type d'information n'avait été développé jusqu'à présent pour le français. Nous prétendons combler ce vide avec la présentation d'un lexique électronique des noms déverbaux élaboré à partir d'un corpus sémantiquement annoté.

Notre lexique inclut 656 noms déverbaux différents (et 737 acceptions) liés aux 4 046 occurrences de ces noms extraites du *French Treebank* (Abeillé 2003), ce qui fait une moyenne de 6,2 occurrences par lemme. Nous avons d'abord codifié la SA de chacune des nominalisations dans le lexique et celle des verbes correspondants, pour décrire ensuite systématiquement la réalisation syntaxique de la SA des 4 046 occurrences répertoriées dans le corpus. [...]

Caroline CORDIER - Newcastle University (Royaume-Uni) [PRIX DOC]

The presence, nature and role of formulaic sequences in the oral productions of English advanced learners of French

This paper deals with the longitudinal investigation of formulaic sequences (FS) in advanced English university students of French, before and after a seven-month stay in France. FS are

studied from a psycholinguistic perspective and defined as prefabricated word-strings retrieved whole from memory, rather than generated by the language grammar (Wray, 2002).

The learners' oral productions on four tasks are examined to shed light on the presence and nature of FS, whether there are different learner profiles and the impact of FS on fluency.

István CSÚRY - Université de Debrecen (Hongrie)

Les structurants de discours à la lumière des études de corpus : approches théoriques, données empiriques et problèmes de méthode

La communication aborde la problématique du traitement d'une catégorie linguistique dont certains représentants ont été étudiés en linguistique française sous l'étiquette de connecteurs. Les fondements empiriques de ces travaux ne sont pas toujours solides et les expressions étudiées n'ont pas fait l'objet d'une systématisation suffisamment efficace. Quant à l'étude du discours, elle se caractérise par une approche qualitative et subjective d'une faible quantité de données. Cependant, la présence de nouvelles approches sensibles à l'attestation devient de plus en plus marquée.

En partant des aspects de la catégorisation des « mots du discours » et de leur étude sur corpus, nous analyserons l'interrelation entre syntaxe, articulation du discours et sémantique.

Annabelle DAVID - Newcastle University (Royaume-Uni)

Receptive lexical knowledge in beginner learners of French: a corpus-based study of 5, 7 and 11 year olds

[...] It is generally assumed that most of the vocabulary learnt is learnt through oral input (e.g. Ellis, 1997). Numerous studies have shown that various factors have very different effects on the learning of vocabulary (e.g. intra-lexicon factors (Laufer 1997); frequency (David 2008); memory (Thornbury 2002); first language (Swan 1997); teaching strategies (Nation 2001), etc. However, we know little about the effects of these factors in relation to age.

This paper investigates different factors influencing the acquisition of lexis in beginner learners of French with different starting ages. This paper will address the following questions: what major factors play a role in the vocabulary acquisition of young beginner learners? And what are the differences between the three age groups? [...]

Alain DAWSON, Jean-Michel ELOY, Christophe REY - LESCLaP-CERCLL - EA 4283, Université de Picardie Jules Verne (France)

Vue perspective sur le français à partir d'une base de données textuelles en domaine d'oïl

Cette communication présente sous une double approche la constitution et l'exploitation d'une base de données textuelles littéraires écrites en picard : PICARTEXT.

1) Une approche va des données vers la théorie. On peut montrer des phénomènes linguistiques communs au picard et au français mais décrits comme grammaticalisés seulement en picard. Ces faits mettent en question l'idée de discontinuité entre les deux variétés.

2) L'autre approche, partant de la Théorie des Correspondances, organise à priori le corpus en le lemmatisant. Elle suggère que les relations entre variantes dialectales reposent sur des potentialités du système exploitées à un autre niveau par les locuteurs dans l'identification interdialectale et même dans l'intercompréhension.

Le corpus d'une variété « collatérale » nous permet donc – *modulo* un enrichissement sociolinguistique des données – d'examiner comment se construisent des distinctions-frontières entre le français et cette variété collatérale.

Michèle DEBRENNE - Université d'Etat de Novossibirsk (Russie) & MoDyCo UMR 7114 CNRS, Université Paris Ouest Nanterre La Défense (France)

L'apport des dictionnaires d'associations lexicales en sémantique comparative des concepts : Svoboda vs liberté

Le concept de « LIBERTE » est un mot-clé de la langue culture tant russe que française. Les méthodes d'études des concepts se sont enrichies récemment des données récoltées au cours d'expériences d'associations lexicales. Malgré le fait que les lexèmes « liberté » et « svoboda » soient considérés comme équivalents par les dictionnaires, leurs champs associatifs dans les deux langues sont très différents : pour les Français les premières associations sont, sans surprise, « égalité » et « fraternité », alors que les Russes privilégient des expressions figées comme « liberté d'action » ou « liberté de parole ».

Henri-José DEULOFEU, Nathalie RIGAUD - TALEP LIF UMR 6166 CNRS, Université de Provence (France)

Etude sur corpus de la construction dite « ellipse modale ». Il fallait la faire et on a pu

Cette communication présente une étude sur corpus (Corpus CERF établi par Jean Véronis 5M de mots de français écrit et oral) des constructions des verbes comme *pouvoir* et *devoir* connues sous la dénomination d' « ellipse modale » (Busquets & Denis 2001 ; Dagnac 2008), du type : *il voulait venir, mais il pouvait pas (venir)*. Notre objectif était double : mesurer leur importance quantitative par rapport aux constructions à l'infinitif non elliptiques de ces verbes et trouver des arguments empiriques pour évaluer deux analyses possibles de ces constructions : l'analyse classique par ellipse de l'infinitif et une analyse par non spécification lexicale du complément d'objet avec valeur anaphorique (*j'aime, j'accepte...*). Nous

présentons les résultats suivants : quantitativement le corpus révèle un nombre assez restreint d'occurrences au regard de celles des constructions « complètes » de ces verbes (*Je peux regarder*). L'« ellipse modale » n'est donc pas d'une structure majeure du français. [...]

Sonia DI VITO - Università degli Studi di Cassino (Italie)

L'utilisation des corpus dans l'analyse linguistique, dans l'enseignement /apprentissage du FLE et dans les pratiques d'intercompréhension entre langues romanes.

L'emploi de ressources numérisées ainsi que l'étude des collocations (selon la définition donnée par Sinclair 1991) peuvent constituer des outils particulièrement utiles pour l'analyse linguistique et la didactique des langues.

Nous avons mis au point une méthode pour désambigüiser un certain nombre de parasyonymes, suivant les critères de collocation, colligation, prosodie sémantique et préférence sémantique (proposés par le courant de la linguistique de corpus d'inspiration britannique). Notre conclusion porte sur le fait que les analyses linguistiques selon ces critères peuvent être mises à profit pour la constitution de ressources lexicographiques monolingues (dictionnaires de langue et des synonymes) et bilingues.

Dans une perspective didactique se basant sur l'intercompréhension (L. Dabène 1996, Blanche-Benveniste et *alii* 1997, Jamet 2003), l'emploi de corpus bilingues et l'extraction des équivalents (Kraif 2003) nous permettent de mettre en évidence les points de contact et les différences entre deux ou plusieurs langues parentes. Les aspects étudiés sont d'ordre syntaxique et sémantique (pour l'analyse des intensifieurs, des connecteurs logiques et du lexique spécialisé).

Maud DUBOIS, Alain KAMBER, Carine SKUPIEN DEKENS - Université de Neuchâtel (Suisse)

Analyse quantitative et qualitative de problèmes orthographiques d'apprenants du français

Sur la base d'un corpus de textes écrits (75.000 mots) d'apprenants du français de niveau B1 et de six langues maternelles différentes (anglais, chinois, espagnol, kurde, portugais et russe), la présente communication a pour objectif de dégager une distribution chiffrée des erreurs d'orthographe grammaticale dans le domaine de l'accord de l'adjectif (épithète, attribut ou détaché), puis de passer à une analyse plus qualitative. Les résultats obtenus permettront d'esquisser des tendances communes à tous les apprenants ou particulières à chacune des langues représentées. Il s'agira de donner des pistes d'explication des erreurs commises et de proposer des moyens d'y remédier.

Marion DUFOUR - CEPED UMR 196 CNRS, INED, IRD, Université Paris Descartes - Paris 5 (France) [PRIX DOC]

Corpus de référence du français parlé et didactique du FLE : comment réconcilier données en linguistique et méthodes d'enseignement ?

Malgré la publication en 2001 d'un *Cadre Européen Commun de Référence pour les Langues* (CECRL) à forte résonance sociolinguistique et impliquant une approche situationnelle du discours, force est de constater que les méthodes de didactique du FLE ont bien du mal à se conformer à cette perspective et à se renouveler. Il semble donc que la dichotomie entre la réalité des pratiques observées sur le terrain et les représentations de la langue véhiculées par les méthodes, reste forte. Dans notre communication, nous mettrons en évidence tout d'abord l'obstacle que représente la conception de la langue des manuels d'enseignement pour la mise à jour des ressources pédagogiques, l'intérêt de sélectionner les contenus linguistiques des programmes d'enseignement compte tenu des genres discursifs et enfin l'opportunité que représentent les corpus de référence du français parlé pour l'élaboration d'une telle méthodologie.

Jacques DURAND¹, Bernard LAKS², Chantal LYCHE³ - ¹CLLE-ERSS UMR 5263 CNRS, Université Toulouse - Le Mirail (France), ²MoDyCo UMR 7114 CNRS, Université Paris Ouest Nanterre La Défense (France), ³Université d'Oslo (Norvège)

Que peut-on dire et ne pas dire sur la liaison en français contemporain ?

Dans cette communication, nous présentons une méthodologie pour étudier la liaison fondée sur le programme PFC (*Phonologie du Français Contemporain : usages, variétés et structures*, Durand, Laks & Lyche 2009). Nous partirons de 35 enquêtes transcrites orthographiquement, codées et vérifiées, qui mettent en scène 372 locuteurs pour lesquels nous disposons ainsi de 49.728 codages de liaisons. Nous défendons l'idée qu'une base empiriquement plus adéquate montre la liaison comme un phénomène multifactoriel et inter-niveaux par excellence présentant un enchevêtrement complexe de zones stables et de zones profondément variables. Notre argumentation s'accompagnera de résultats quantitatifs précis sur divers domaines de variation (en particulier registre, âge, sexe, niveau d'études et facteurs diatopiques), qui complètent Pagliano & Laks (2005), Durand et Lyche (2008), Mallet (2008). Nos conclusions seront comparées à celles obtenues sur les mêmes paramètres par, entre autres, Malécot (1975), Ashby (1981), van Ameringen & Cedergren (1981), Booij & de Jong (1987), De Jong (1994) et Ranson (2008).

Soodeh EGHESAD - Université Sorbonne Nouvelle - Paris 3 (France) [PRIX DOC]

Implications culturelles des pratiques d'évaluation en classe de langue étrangère en Iran, en France et aux Etats-Unis

Cette communication propose une étude des pratiques d'évaluation utilisées dans les cours de langue étrangère en Iran, en France et aux États-Unis. L'objectif est d'analyser les différences et les points communs de ces pratiques dans ces pays afin de comprendre si les processus évaluatifs de ces contextes sont universelles ou culturellement spécifiques et s'ils

sont spécifiques, quelle est la dynamique qui met en jeu leurs spécificités, c'est-à-dire les réalités sociales, politiques ou économiques qui les forment. Nous nous appuierons sur l'analyse des facteurs comme la conception de l'évaluation, le rôle de l'examen et les objectifs d'apprentissage des langues pour étudier les liens entre ceux-ci et les « ethos » de chaque culture et contexte. À partir de 48 entretiens compréhensifs avec des enseignants de langues (français, anglais et persan), nous avons montré que malgré l'universalité des instruments d'évaluation, l'influence des traditions pédagogiques et des cultures éducatives propres à chaque pays, ainsi que les différents besoins des apprenants, les contraintes institutionnelles et les exigences politico-socio-culturelles de chaque contexte suscitent diverses démarches évaluatives dans chaque contexte.

Natacha ESPINOSA - Fédération CLESTHIA, Université Sorbonne Nouvelle - Paris 3 (France)

Quand la recherche-action-formation aide à repenser la pratique didactique. Un exemple sur la maîtrise du langage oral en maternelle.

Nous présenterons un exemple de recherche-action menée conjointement par des linguistes et des enseignantes d'écoles maternelle, sur l'étude des situations d'interaction élève-enseignant les plus efficaces pour l'apprentissage du langage.

Nous réinvestirons, à partir des exemples de corpus (séances filmées), la question de la place de la dimension linguistique dans la maîtrise du langage à l'école maternelle (syntaxe, lexique), puis du rôle des adultes dans l'accès à la maîtrise des fonctionnements langagiers. Nous présenterons brièvement quelques exemples de situations et de supports qui peuvent favoriser des interactions langagières efficaces pour l'apprentissage du langage oral et écrit.

Danièle FLAMENT-BOISTRANCOURT¹, Juliette DELAHAIE² - ¹MoDyCo UMR 7114 CNRS, Université Paris Ouest Nanterre La Défense (France), ²Laboratoire de Linguistique Formelle UMR 7110 CNRS, Université Paris Diderot - Paris 7 (France)

Constitution et exploitation de corpus d'interactions verbales pour le FLE : problèmes et programme

Depuis les débuts de l'approche actionnelle, les interactions verbales sont au cœur de l'enseignement/apprentissage du FLE, cependant l'exploitation de données numérisées y reste confidentielle. Nous voudrions donner une explication à cette situation et définir les modalités de constitution/exploitation d'un vrai corpus d'interactions verbales pour la didactique du FLE. Notre exposé abordera trois points problématiques : la pertinence et la représentativité des données numériques actuellement disponibles pour le FLE ; la question de l'analyse linguistique de ces données ; enfin, le problème de la pédagogisation des corpus.

Michel FRANCARD - VALIBEL, Université Catholique de Louvain (Belgique)**Le corpus d'exclusion en lexicographie différentielle du français. Une pratique, des questions.**

Le corpus d'exclusion est celui constitué par des données contenues dans des ouvrages dits « de référence », qui décrivent, pour l'essentiel, une langue normée (dictionnaires, grammaires, etc.). Les données étudiées dans les travaux de néologie ou de lexicographie différentielle sont celles qui ne trouvent pas leur place dans ces ouvrages, desquels elles sont exclues en raison de leur statut de variantes non standard (diatopismes, néologismes, etc.).

L'approche fondée sur un corpus d'exclusion, intuitivement simple à appréhender, pose une série de problèmes théoriques et méthodologiques qui seront abordés dans cette communication du seul point de vue des travaux récents en lexicographie différentielle du français. On analysera pour ce faire quatre dictionnaires (Thibault 1997, Poirier 1998, Rézeau 2001, Francard *et al.* 2010), en montrant les logiques à l'œuvre dans ces ouvrages (et dans des travaux connexes des mêmes auteurs) du point de vue de la pratique du corpus d'exclusion. Des logiques largement convergentes, mais avec quelques différences significatives, en particulier dans la délimitation du corpus d'exclusion.

Itsuko FUJIMURA - Université de Nagoya (Japon)**Etude sur corpus des groupes Nom + Adjectif de couleur dans la presse française**

Ayant remarqué l'emploi courant dans la presse française de groupes formés d'un nom et d'un adjectif de couleur <N+Adjcou>, nous nous proposons d'en préciser le sens et la fonction. Dans cet objectif, nous avons relevé 78.500 occurrences de <N+Adjcou>, constituant 2.800 combinaisons distinctes, dans deux corpus journalistique et littéraire :

1) Le quotidien *Le Monde* de 1988 à 2006 : 144 millions de mots

2) *Corpatext* : 37 millions de mots.

Une double analyse, quantitative (fréquence, information mutuelle) et sémantique, permettra, ensuite, de mettre au jour des caractéristiques de ces groupes propres à l'écriture journalistique : formes courtes, noms simples et adjectifs basiques ; fonction non de sous-catégorisation mais de dénomination anaphorique, imagée, plus ou moins stéréotypée, d'un référent préidentifié (*blouses blanches* = personnel médical) ; matrice de formations néologiques pittoresques : *lunettes roses/lunettes noires* = vision optimiste/pessimiste (cf. *voir la vie en rose/en noir*).

Nathalie GETTLIFFE¹, Susan DIXON² - ¹Université de Strasbourg (France), ²Washington and Lee University (Etats-Unis)**Les corpus d'interactions pédagogiques en ligne et la création de nouvelles activités pour l'apprentissage des langues par visioconférence**

Dans le cadre d'un cours sur le tutorat en ligne suivi par des apprentis-tuteurs de Master 2 en Conception Multimédia en Ligne (CML), nous nous sommes appuyés sur un corpus

d'interactions pédagogiques (le français en première ligne) pour créer de nouveaux exercices autour de la visioconférence de poste à poste destinés à des étudiants de Virginie. Enregistrements, entretiens et travaux réflexifs nous ont permis de mieux comprendre comment les tuteurs s'étaient appuyés sur la base de données d'interactions pédagogiques existantes pour initier des conversations pédagogiques ciblant au mieux les attentes de leurs partenaires américains.

Wajih GUEHRIA - MoDyCo UMR 7114 CNRS, Université Paris Ouest Nanterre La Défense (France)

Le "code-switching" chez les jeunes Mantais : de l'expression d'un conflit à la constitution d'un idiome symbiotique

L'engouement pour l'arabe maghrébin incite les jeunes de la ville de Mantes-la-Jolie à encadrer l'arabe dialectal dans un discours en français. L'intérêt pour l'arabe ne provient pas du seul fait qu'elle soit la langue d'origine de leurs parents puisqu'un certain nombre de Mantais, sans aucune attache avec le Maghreb, pratiquent, tant bien que mal, un codeswitching français/arabe. Cette pratique linguistique permet aux personnes interrogées de mettre en avant une identité qui transcende le clivage Maghreb/France dans lequel ils ne se reconnaissent pas. Ils se revendiquent plutôt d'une catégorie où ils sont créateurs de tendances, linguistique et culturelle, qui leur sont propres : le codeswitching, le style vestimentaire, les arts urbains, etc.

Le code-switching est donc, dans le domaine de l'expression verbale, le moyen d'instituer une identité propre, originale, qu'on pourrait dire "symbiotique".

Emmanuelle GUÉRIN¹, Stéphanie MALAUSA² - ¹Laboratoire Ligérien de Linguistique - EA 3850, Université d'Orléans (France), ²CEPED UMR 196 CNRS, INED, IRD, Université Paris Descartes - Paris 5 (France)

« Langue des jeunes », « langue des cités », « langue de banlieue », etc. : quand la dénomination fait barrage à la caractérisation de formes de langue

Le travail présenté participe de la réflexion menée dans le cadre du projet MLE-MPF qui s'intéresse aux effets du multiculturalisme sur les pratiques langagières des « jeunes » à Paris et à Londres. Un des objectifs est de proposer un éclairage sur des formes de langue que la pensée commune, relayée par les médias (et inversement), exclut du champ de variation de la langue.

Du projet naît un corpus de données qui s'oppose à la caractérisation réductrice que supposent les dénominations qui circulent dans la société : bien qu'on admette que ces formes de langue s'observent particulièrement dans les pratiques de jeunes locuteurs de milieu dit « populaire », on émet l'hypothèse qu'elles concernent une communauté bien plus large. [...]

Marie-Noëlle GUILLOT - University of East Anglia, Norwich (Royaume-Uni)

Le corpus restreint comme plateforme de recherche : faces cachées du développement pragmatique en langue étrangère, nouveaux éclairages sur la langue source

Cette communication s'appuiera sur la synthèse de trois études relatives au même corpus restreint d'interactions verbales en français et anglais langue maternelle et seconde, rassemblé pour les besoins d'un projet consacré au développement pragmatique en français langue étrangère, pour aborder diverses questions relatives au corpus lui-même. Ces études mettent en relief des phénomènes significatifs dans le comportement verbal des apprenants de français et leur développement pragmatique, et apportent un éclairage du français sous un autre angle qui en révèle des aspects insoupçonnés par rapport à l'anglais. Mais comment les caractéristiques du corpus affectent-ils la valeur et la portée de ces résultats ?

Olga GUSHCHINA-MAGNO - Université Paris-Sorbonne - Paris 4 (France)

Utilisation de la base de données catégorielle Frantext et création de glossaire sous Access

L'objet de cette communication est la recherche et l'analyse des items construits selon le modèle (Article) + Substantif + Substantif, effectuées, d'une part, grâce à l'utilisation de la base catégorielle Frantext et, d'autre part, à la collecte personnelle (textes de magazines et journaux, œuvres littéraires, publicité orale et écrite, radio, télévision, discours oral et écrit des locuteurs natifs français). Nous parlerons des difficultés rencontrées dans le traitement des données et des solutions proposées dans le cadre de notre recherche. Le travail a donné lieu à la création d'une base de données sous Access que nous souhaiterions valoriser.

Cécile HAUT - ATILF UMR 7118 CNRS, Nancy-Université (France) [PRIX DOC]

Intérêt des données issues de corpus informatisés en linguistique historique

L'objet de notre étude, à savoir la formation des locutions conjonctives temporelles en français, s'appréhende à travers le cadre de la linguistique diachronique par le biais de diverses ressources linguistiques informatisées permettant de nous éclairer sur leurs faits de grammaticalisation. L'exploitation des données issues de bases textuelles facilite, d'une part, la mise en évidence d'une succession de faits diachroniques et, d'autre part, la mise en exergue d'un état de langue donné permettant de vérifier l'existence d'une corrélation entre niveau de langue et degré de grammaticalisation. Il s'agit ainsi de démontrer l'intérêt des méthodes de la linguistique de corpus dans toute étude s'inscrivant dans le cadre processuel de la grammaticalisation.

Siham HOCINI - Université de Bejaia (Algérie) & Université de Tizi-Ouzou (Algérie)

La place des données naturelles et situées dans l'étude de la langue française

L'objectif de cette communication est de montrer la place des données naturelles pour l'étude de la langue française : quel est l'apport de la prise en compte de ces données pour

la linguistique ? Quels en sont les enjeux méthodologiques ? Quelles sont les spécificités de ces données par rapport aux autres types de données (les textes écrits, élaborés, issus de sources légitimes, comme la presse ou les textes syndicaux ; les productions langagières orales provoquées par le chercheur comme les entretiens) ? Nous visons donc la présentation de quelques-unes des méthodes utilisées pour leur exploitation. Pour cela, nous nous focaliserons sur l'analyse d'un corpus de données naturelles : un ensemble de conversations produites en français que nous avons enregistrées dans un milieu professionnel.

Magali HUSIANYCIA - ATILF UMR 7118 CNRS, Nancy-Université (France) [PRIX DOC]

Séquentialité et type linguistique : deux notions à revisiter pour l'analyse des genres de discours en situation professionnelle

Dans le cadre d'une recherche s'intéressant aux productions verbales entre professionnels en situation de travail, nous tentons de déterminer des genres de discours en situation professionnelle. Cette étude nécessite de s'interroger sur les questions méthodologiques de recueil et d'analyse des données. Nous avons eu recours à la notion de « séquence » (J.-M. Adam, 1998, 2001) articulée à la celle de « type linguistique » (J.-P. Bronckart, 1996) que nous avons adaptées à notre propre corpus. Nous exposons les raisons de ce choix et le découpage que nous avons opéré. Nous avons ainsi déterminé 7 types de séquences : séquences opérative, explicative, prescriptive, descriptive, encadrante, interstitielle et informelle relevant chacune d'un des trois types linguistiques : co-constructif, expositif ou péri-professionnel.

Nous posons l'hypothèse que le genre de discours d'une réunion est déterminé par la combinatoire de ces différentes séquences, mais que le type de séquence dominant (type majeur) détermine le genre de discours.

Richard INGHAM - Birmingham City University (Royaume-Uni)

Les corpus d'anglo-normand électroniques comme ressources pour l'histoire plurielle du français: les connecteurs *pourtant* et *à cause que*

Anglo-Norman and the 'plural history' of French: the connectives *pourtant* and *à cause que*.

Les textes de la pratique et autres ressources documentaires incorporés dans trois corpus d'anglo-normand électroniques permettent une perspective nouvelle sur la variation en français d'avant l'ère moderne. Ils sont capables de nous renseigner sur des aspects importants, diatopiques ainsi que diaphasiques, de la variation en diachronie. Grâce à ces nouvelles ressources, on arrive à rendre compte de façon plus complète de l'évolution des tendances sémantico-pragmatiques du lexique français qu'en suivant les approches traditionnelles qui privilégiaient le développement du francien et du français standard. Il est démontré dans cette étude que les connecteurs *pourtant* et *à cause que* pouvaient s'employer en anglo-normand avec leurs valeurs modernes bien avant les dates proposées par les dictionnaires historiques de référence. Les citations nombreuses relevées dans les

textes anglo-normands du XIV^{ème} siècle offrent ainsi la possibilité de découvrir les débuts des évolutions en sémantique lexicale du français post-médiéval.

Paul ISAMBERT - LaTTiCe UMR 8094 CNRS ENS, Université Sorbonne Nouvelle - Paris 3 (France)

Concilier les attentes théoriques et des corpus en apparence contradictoires

Accommodating theoretical expectations with apparently conflicting corpora

Les corpus forment des ensembles de données obtuses, accidentelles dans une certaine mesure, qui ne peuvent pas être acceptées d'emblée; les corpus assurément ne doivent pas être rejetés sans justification - en fait ne doivent pas être rejetés du tout - mais il ne faut pas les considérer aveuglément comme des images fidèles de la réalité, particulièrement quand ils entrent en conflit avec des attentes théoriques. La question qui se pose est donc : quand doit-on abandonner une hypothèse démentie par les données, et quand doit-on au contraire penser que les faits connus ne donnent qu'une image incomplète de la réalité ? Cette dernière possibilité nécessite bien sûr que l'on *montre* que les données sont partielles ; une simple affirmation pour sauver une théorie n'a aucune valeur scientifique. On illustrera ce problème avec l'adverbe *autrement*, pour lequel les données historiques contredisent la théorie de la grammaticalisation.

Betsy KERR - University of Minnesota - Twin Cities (Etats-Unis)

Semantic Anglicisms in Contemporary Metropolitan French

The present study provides a comprehensive and detailed description of recent usage of selected semantic Anglicisms (e.g. *réaliser* = 'se rendre compte') which have entered the language or significantly increased in frequency over the last half-century. Anglicisms for study were identified with the help of Rifelj 1996 and Picone 1996 as well as didactic texts and purist commentaries. Data have been gathered through the use of three primary corpora: *Le Monde* 1998, the Chambers-Rostand Journalistic Corpus and the Leeds Internet Corpus. The study examines relative frequencies of the given lexical item with both the traditional and the innovative senses, as well as relative frequencies and contexts of use of any pre-existing synonyms of the Anglicism. The data belie both the purists' claim of wholesale transfer of the English item to French and their prediction of displacement of the traditional meaning of the lexical form and the neologism's synonym.

Aurore KOEHL - ATILF UMR 7118 CNRS, Nancy-Université (France) [PRIX DOC]

Étude morphologique des noms désadjectivaux : expliquer la profusion des suffixations par les données

Un même adjectif peut être la base morphologique de plusieurs noms (enregistrés dans les dictionnaires ou néologismes), sans distinction sémantique apparente (e.g.

tendreur/tendreté). L'étude de la pluralité des suffixations permettant de construire un nom désadjectival en français (-ité, -erie, -itude, -eur, -esse, -ise, etc.) est basée sur l'analyse de 3.750 couples adjectifs-noms, issus du TLF, du journal Le Monde ou collectés sur la Toile. 1.358 des 3.750 noms sont concernés par le doublonnage. Quelles contraintes motivent le choix d'une suffixation plutôt qu'une autre ? L'étude de données extensives en morphologie permet de mettre en avant les facteurs favorisant l'apparition de doublons et d'appuyer les résultats obtenus par l'étude individuelle de chaque suffixation en jeu.

Hélène LEFRANÇOIS - Université d'Ottawa (Canada)

Espace littéraire et oralité dans les contes populaires normands de Gaston Demongé

Cette étude portera sur les enjeux méthodologiques et sociolinguistiques de la représentation de la langue orale dans les contes normands au début du XXe siècle. Nous examinerons comment ces contes peuvent servir à la reconstruction de la langue parlée ancienne et dans quelle mesure l'étalonnage avec les sources orales est possible. À cette fin, nous analyserons les contes de Gaston Demongé en les mettant en rapport avec des sources dialectologiques de l'époque et d'aujourd'hui, mais aussi en les comparant à des corpus oraux. Cela permettra d'analyser les relations entre la langue établie dans les contes de Demongé et la langue répertoriée dans les glossaires ou autres sources scientifiques. Du reste, à la lumière du travail effectué par les associations de préservation de la langue normande, nous pourrons observer comment Demongé est interprété aujourd'hui au niveau linguistique.

Iryna LEHKA-LEMARCHAND¹, Maria CANDEA², Cyril TRIMAILLE³ – ¹LiDiFra - EA 4305, Université de Rouen (France), ²Fédération CLESTHIA, Université Sorbonne nouvelle - Paris 3 (France), ³LIDILEM - EA 609, Université Stendhal - Grenoble 3 (France)

Etudier la perception de phénomènes sociophonétiques : méthodes « artisanales » et/ou interfaces en ligne ?

Cette communication propose une réflexion épistémologique sur l'étude de la perception de traits linguistiques, à partir d'une confrontation de différentes méthodes étudiant la perception de marqueurs de l'accent populaire dit « de banlieue ». En s'appuyant sur diverses enquêtes menées par les auteurs, notre contribution cherche à dégager les apports et les limites de deux types d'approche (enquêtes quantitatives en ligne vs tests perceptifs plus artisanaux basés sur des interactions en focus groupe par exemple), puis interroge les modes de production et d'exposition du savoir qui y sont liés. Ce faisant, elle problématise les rapports entre méthodes d'investigation, positionnements théoriques et résultats visés.

Erin MACMURRAY, Marguerite LEENHARDT - SYLED/CLA2T - EA2290, Université Sorbonne Nouvelle - Paris 3 (France) [PRIX DOC]

La textométrie appliquée aux corpus francophones : statut théorique et état méthodologique

Comment articuler corpus et méthodes d'analyses pour l'étude linguistique du français ? Cette question, qui n'est pas récente, a déjà été abordée par de nombreux chercheurs en linguistique de corpus durant les deux dernières décennies. Toujours d'actualité, elle se pose dans les mêmes termes quant à l'étude textométrique des phénomènes linguistiques.

Nous présentons une synthèse des différents outils et analyses statistiques effectués sur des corpus francophones discutés lors des conférences JADT depuis 2000.

Nous nous concentrons sur les méthodes de calcul implémentées dans la plupart de ces outils existants, l'impact de ces outils et leur utilisation sur une large variété de corpus, afin de faire émerger une esquisse des courants textométriques récents en Francophonie.

Anna MAKEROVA - Université Paris Ouest Nanterre la Défense (France) & Université d'Etat de Tcheliabinsk (Russie)

« Im/pertinence lexicale » dans la pratique langagière chez les jeunes Montreuillois

Cette proposition émane d'une enquête sociolinguistique menée auprès des jeunes de Montreuil (Ile-de-France), dans un contexte très multilingue. C'est à cette fin que, dans une tradition sociolinguistique, des entretiens semi-directifs et des enregistrements écologiques ont été effectués au studio de rap *La Pêche (Montreuil)* et au squat montreuillois depuis l'an 2008.

Je me suis immergée dans la vie du studio de rap et du squat où j'ai pu observer les interactions langagières des jeunes. Le but était de vérifier l'hypothèse selon laquelle *la diffusion des phénomènes de contact de langues va au-delà des communautés initialement concernées* : « sous quel angle cette diffusion est-elle abordée dans la pratique langagière des Montreuillois ? ». [...]

Mario MARCON - Université d'Udine (Italie) [PRIX DOC]

GRAPHraséologie. Modélisation de graphes pour la description et la reconnaissance automatique de données phraséologiques

Notre communication se focalise sur l'utilisation des « graphes » (automates à états finis) pour le traitement des données phraséologiques. Nous proposons la modélisation de graphes qui décrivent des formes proverbiales françaises tirés de la base de données *DicAuPro*. Les graphes sont conçus en vue de la reconnaissance de formes proverbiales dans des corpus (statiques et dynamiques). D'autre part, nous visons la reconnaissance d'autres données phraséologiques (patrons, formules, cadres lexico-grammaticaux etc.) qui caractérisent le(s) corpus utilisé(s). Sur la base des résultats obtenus, nous essayons d'estimer les relations entre les descriptions lexico-grammaticales des formes proverbiales par rapport au continuum des données phraséologiques non proverbiales. Nous souhaitons

mettre en évidence une fois de plus la souplesse du graphe et son adaptation aux besoins des recherches linguistiques, tout en questionnant le linguiste au fur et à mesure de sa modélisation.

Philippe MARTIN - CLILLAC-ARP - EA 3967, Université Paris Diderot - Paris 7 (France)

Un logiciel multifonctions pour l'analyse des corrélations texte-intonation et l'enseignement de la prononciation d'apprenants du FLE

Pour permettre plus efficacement l'analyse d'un grand nombre de données (de type ponctuant et conjonction), et ainsi en établir les corrélats prosodiques, on a implémenté dans un logiciel d'analyse acoustique de la parole une fonction permettant d'afficher en un seul clic de souris un mot et son contexte ainsi que l'analyse acoustique correspondante (spectrogramme, fréquence fondamentale, intensité, durées syllabiques). Chaque occurrence du mot ainsi qu'éventuellement son contexte gauche et droite spécifié par l'utilisateur est retrouvé automatiquement dans un ensemble de fichiers alignés texte-son au format XML regroupés dans un même répertoire, et l'analyse acoustique correspondante est immédiatement affichée aux fins d'analyse.

La fonction d'affichage sonore d'occurrences spécifiques permet à ce même logiciel de présenter à des apprenants de FLE des exemples sonores extraits de corpus de parole lue ou spontanée.

Branislav MESZAROS - Université Paris Descartes - Paris 5 (France) [PRIX DOC]

Vers un corpus dynamique en FLE via l'analyse des correspondances entre les supports textuels et les niveaux du CECR

Bien que l'enseignement du français langue étrangère (FLE) dispose de nombreuses ressources pédagogiques, on est encore loin de pouvoir constater l'existence d'un corpus de référence. Cela se ressent d'autant plus depuis l'apparition du *Cadre Européen Commun de Référence (CECR)* et la place de plus en plus prépondérante des certifications en langues. La mise en place d'un tel corpus soulève cependant plusieurs difficultés, parmi elles le choix des critères dits constitutifs. Il semblerait alors qu'une piste se basant sur la mise en correspondance des supports textuels avec les niveaux du *CECR* via les *Référentiels* puisse être envisagée comme une des solutions.

Rosamond MITCHELL¹, Annabelle DAVID², Florence MYLES³ - ¹University of Southampton (Royaume-Uni), ²Newcastle University (Royaume-Uni), ³University of Essex (Royaume-Uni)

Teacher gesture in the L2 classroom and its role in L2 vocabulary acquisition: a corpus based study

Gesture is an integral part of human communication, and performs a variety of functions to support meaning making. However, the role of gesture in instructed language learning has

hardly been studied. This paper presents preliminary findings from a study of the learning and teaching of French as a foreign language with English L1 children aged 5, 7 and 11. A corpus of lessons has been videorecorded and transcribed, and this paper examines a) the nature and extent of teacher gesture during classroom interaction; and b) the influence of teacher gesture on children's vocabulary learning, as measured in a receptive and frequency based vocabulary test.

Raymond MOUGEON¹, Françoise MOUGEON¹, Katherine REHNER² - ¹Université York, Glendon (Canada), ²Université de Toronto à Mississauga (Canada)

Usage des prépositions à/en/au(x) avec les toponymes sur un continuum L1-L2

Notre étude examine l'usage des prépositions à/en/au(x) avec les noms de localité ou de pays. Elle repose sur des corpus d'entrevues enregistrées auprès d'adolescents ou jeunes adultes que l'on peut situer sur un continuum FL1—FL2. Le choix entre ces prépositions étant déterminé par des facteurs contextuels dont la motivation sémantique est réduite et différent de l'anglais qui emploie la même préposition quel que soit le type de toponyme, nous avons trouvé que seuls les locuteurs FL1 de la ville de Québec utilisent ces prépositions selon les règles du FS et que parmi les autres locuteurs, on observe une tendance variable à substituer d'autres prépositions à à/en/au(x). Une analyse statistique révèle que cette tendance est surtout contrainte par le contexte linguistique et le niveau d'exposition au français. Nous situons nos résultats par rapport à ceux des recherches sur la variation du FL2 et du FL1 en situation de contact linguistique.

Dominique NOUVEAU - Radboud University Nijmegen (Pays-Bas)

E final et e caduc en FLE : Exploitation du corpus PFC par des apprenants néerlandophones

Le corpus oral *Phonologie du Français Contemporain* (Durand, Laks & Lyche, 2002) offre un accès direct (<http://www.projet-pfc.net/>) à une importante base de données sonores authentiques transcrites orthographiquement et codées, étroitement associées à un moteur de recherche. Dans cette communication, nous nous proposons d'illustrer comment le recours à ce corpus de 77 enquêtes (liste de mots et texte lus, entretiens guidé et libre), et à ses ressources complémentaires destinées à l'enseignement, favorise le traitement de points délicats de la phonologie du français dans l'enseignement de la linguistique à des étudiants de FLE. A titre d'exemple, nous présentons des activités réalisées avec/par des étudiants néerlandophones, dont l'objectif était d'éclairer grâce à la collecte d'exemples ciblés contextualisés le comportement des *e final* et *e caduc*, et la loi des trois consonnes (Grammont, 1926) tout en documentant la variation diatopique de l'Hexagone.

Roberto PATERNOSTRO - Università degli Studi di Brescia (Italie) & MoDyCo UMR 7114 CNRS, Université Paris Ouest Nanterre La Défense (France) [PRIX DOC]

Aspects phonétiques de l'accent parisien « multiculturel » : défis et enjeux d'une approche écolinguistique.

Notre recherche, élaborée dans le cadre du projet « *Multicultural London English – Multicultural Paris French* » (MLE-MPF), vise à étudier d'un point de vue sociolinguistique le processus de variation phonétique se déroulant dans une grande ville occidentale telle que Paris, tout spécialement en ce qui concerne le parler de jeunes locuteurs issus de l'immigration et demeurant en banlieue.

Notre but est d'investiguer la dynamique du contact des langues, de manière à comprendre si le parler de groupes ethniques importants, en tant que possible moteur d'innovation, exerce une influence sur la langue commune.

Mes analyses – à la fois acoustiques, perceptives et statistiques – portent sur la comparaison d'énoncés produits par trois groupes de locuteurs (FM = français issus de milieux multiculturels ; FF = franco-français ayant peu ou pas de contacts avec des milieux multiculturels ; FC = franco-français ayant des contacts avec des milieux multiculturels), notamment en ce qui concerne la présence d'un contours prosodique final montant/descendant, caractérisé par une chute abrupte de F0.

Chrysta PELISSIER - Praxiling UMR 5267 CNRS, Université Paul Valéry - Montpellier 3

Usage de ressources linguistiques pour la caractérisation de la notion d' « aide dans la formation » : le cas du dictionnaire des synonymes et du TLFi

Dans cette communication, nous souhaitons présenter comment le TLFi et le dictionnaire des synonymes proposé par le CRISCO nous ont permis d'avancer sur la caractérisation de la notion d'aide dans l'enseignement en ligne.

D'un point de vue méthodologique, nous avons sélectionné les 10 premiers synonymes du dictionnaire des synonymes en ligne. Chaque synonyme a ensuite fait l'objet d'une étude dans le TLFi : pour chaque synonyme/entrée, nous avons isolé les définitions. La mise en commun de ces définitions a ensuite permis d'identifier 5 « orientations » de caractérisation de la notion d'aide : acteur, situation, but, action, type.

Mélanie PETIT, François NEMO - Laboratoire Ligérien de Linguistique - EA 3850, Université d'Orléans (France)

Corpus oraux et théories sémantiques

L'objet de notre communication sera de montrer l'importance des corpus oraux pour les sémantiques linguistiques.

Nous montrerons comment l'étude des corpus oraux conduit à transformer la vision que le sémanticien a de la diversité des emplois écrits, par exemple en imposant de remplacer des distinctions sémantiques éprouvées comme les distinctions signification/sens (binaire) ou signification/profil/sens (ternaire) par des distinctions encore plus fines.

Nous montrerons également que les résultats obtenus amènent à s'interroger sur la notion de forme sonore, ou encore de signifiant, en rompant avec une représentation purement phonématique de la forme sonore au profit de représentations plus complexes intégrant forme phonématique et forme prosodique.

Nous montrerons enfin en quoi les corpus oraux sont aussi à même de trancher empiriquement des débats, restés théoriques, comme le débat sur l'argumentativité de la langue ou encore le débat qui oppose les « use-based semantics » aux sémantiques dites « minimales ».

Vera REGAN¹, Isabelle LEMÉE² - ¹University College Dublin (Irlande), ²St Patrick's College, Drumcondra (Irlande)

Quantitative analysis of a corpus of Polish speakers of French L2: sociolinguistic competence and variation patterns

This study on the Polish diaspora investigates variation patterns in the speech of Polish migrants living in France. Our corpus constitutes of eleven speakers from two principal waves of migration: post world war two and 'solidarity' migration. Through both quantitative and qualitative approaches, our study analyses whether the speakers 'integrated' into French society and whether their language practices were indicative of that integration. Our triangulated results reveal a complex story of people who actively seek to ensure the transmission of Polish language and culture as well as a highly successful integrationist enterprise into French society. This picture emerges both from the stories the speakers tell about their lives and experiences of migration to France as well as from the multifactorial (Goldvarb) analysis of their French L2 speech patterns. The analysis of the stable variable 'ne deletion' found considerable interspeaker variation which was constrained by social factors relating to migration patterns.

Christophe REY - LESCLaP-CERCLL - EA 4283, Université de Picardie Jules Verne (France)

La Métalexicographie, une discipline résolument tournée vers la modernité

Situé dans la continuité de l'article fondateur de Gilles Petrequin et Pierre Swiggers (2007), notre exposé se propose d'évoquer les relations entretenues entre l'outil informatique et la jeune discipline qu'il est convenu d'appeler depuis les travaux précurseurs de Bernard Quemada (1968) et de Josette Rey-Debove (1971), la « Métalexicographie ».

Nous nous attacherons à montrer ici que la Métalexicographie semble avoir connu grâce au « souffle informatique de la langue » (Pruvost, J., 2006) une mutation sans précédent qui a considérablement renouvelé les méthodes et les pratiques d'investigation de ses chercheurs.

Lucie RIOU - University of Exeter (Royaume-Uni)**Comment gérer les actes de parole potentiellement conflictuels ? Analyse pragmatique de l'expression du désaccord entre amis**

En s'appuyant sur des exemples de désaccords, critiques et autres actes de paroles potentiellement conflictuels, cette communication vise à explorer les différents procédés mis en oeuvre par les interlocuteurs afin de modérer l'effet dommageable des Actes Menaçants pour la Face, lors de conversations entre amis, et ainsi d'éviter le conflit. Nous tenterons d'appréhender en contexte ces procédés adoucisseurs, et d'en explorer certains aspects plus en détail, notamment leur variété d'emploi ainsi que la différence interpersonnelle dans le contexte spécifique de la conversation entre pairs, où les rapports de pouvoir sont souvent moins évidents que dans les contextes de discours hiérarchisés (relations médecin-patient, professeur-élève, etc.)

Inès SADDOUR - Aston University (Royaume-Uni)**The role of aspect in expressing simultaneity in native and learner productions**

This paper examines the use of aspectual perspective in expressing the simultaneity relation of two events. Strongly inspired by Schmiechtová's (2004) categorisation, we investigate the different aspectual styles used in native and adult learner productions. We compare two unrelated native languages, Tunisian Arabic (TAL1) and French (FrL1), and we look at the way Tunisian learners of French as a second language (FrL2) use aspect to express simultaneity at different acquisitional stages.

Gilles SIOUFFI¹, Agnès STEUCKARDT², Chantal WIONET³ - ¹Université Paul Valéry - Montpellier 3 (France), ²Université du Sud Toulon-Var (France), ³Université d'Avignon (France)

Comment enquêter sur la diffusion des normes à partir de corpus écrits ?

Dans le cadre d'une enquête sur ce qu'on pourrait appeler la « normativité des usages », il convient de s'interroger sur les matériaux et les types de corpus qui sont les plus appropriés : corpus déclaratifs (entretiens, interviews), enregistrements et corpus oraux, corpus écrits. En effet, les dynamiques implicites qui poussent les locuteurs à choisir une forme langagière sont beaucoup plus difficiles à cerner que celles qu'on relie traditionnellement à « la norme », car elles ne sont pas toujours appuyées sur une référence à de l'explicite (« règle », « exemple », « modèle », ...). Par ailleurs, on sent intuitivement qu'une investigation poussée de la dynamique de la diffusion paraît nécessiter l'interrogation de grands corpus.

Cette communication présentera deux types de questions qui peuvent naître au moment d'initier semblables enquêtes :

- Comment isole-t-on le fait sur lequel va porter l'investigation ? Quelles limites rencontre ici le repérage de fréquence en corpus ? Quelle méthodologie requiert le recours aux commentaires épilinguistiques ?

- Quels types de corpus interroger ? Quel corpus est susceptible de nous donner les meilleurs résultats ?

Il s'agira de développer ces réflexions à partir de l'expérience que procure une recherche sur les outils de modalisation (type *c'est vrai que, tu vois, juste*) dans une diachronie très récente. On mettra en regard les avantages et les inconvénients que présentent les bases Frantext et Europresse pour une semblable enquête. Nous poserons également la question de savoir comment utiliser internet, notamment google, et certains sites de discussion. La communication fera état pour finir de quelques besoins de la recherche à l'égard des corpus, des bases et des méthodes d'extraction d'informations actuellement disponibles.

Juliette THUILIER, Anne ABEILLÉ, Benoît CRABBÉ - Laboratoire de Linguistique Formelle UMR 7110 CNRS, Alpage, INRIA, Université Paris Diderot - Paris 7 (France)

Préférences concernant l'ordre relatif des compléments du verbe en français

Nous étudions les facteurs influençant l'ordre relatif des compléments postverbaux de 150 verbes ditransitifs du français (donner, dire...). Dans la lignée des travaux sur l'alternance dative en anglais (Bresnan et al. 2007), nous avons mené une étude sur corpus : deux corpus journalistiques (Le Monde, Est-Républicain) et un radiophonique (Ester). On observe une préférence générale pour l'ordre V NP PP, ainsi que le rôle crucial de la longueur respective des compléments, et dans une moindre mesure, celui du lemme verbal et du caractère collocationnel du SP. Les données, annotées pour le caractère animé des référents du SN et du SP, indiquent que ce facteur n'intervient pas, alors qu'il est pertinent pour l'ordre en anglais ou en allemand. En neutralisant l'effet de longueur, nous avons entrepris une étude psycholinguistique, à base de questionnaires, pour juger entre deux continuations possibles à partir de phrases de nos corpus en contexte. Les résultats concordent avec ceux des corpus.

Agnès TUTIN - LIDILEM - EA 609, Université Stendhal - Grenoble 3 (France)

L'étude et la modélisation de la phraséologie évaluative dans les écrits scientifiques : une exploitation du corpus Scientext

[...] L'objectif de notre communication est double. Dans un premier temps, nous présenterons la constitution du corpus et les types de traitements effectués, étroitement liés aux études linguistiques envisagées dans ce projet. Afin de pouvoir mettre en contraste les modes d'écriture scientifique dans les différents genres et selon les disciplines, nous avons construit un corpus diversifié comportant plusieurs types de textes [...] Dans un deuxième temps, nous souhaitons présenter une étude linguistique basée sur le corpus Scientext portant sur le lexique évaluatif en sciences humaines et sociales (linguistique, psychologie, sciences de l'éducation, économie. Ce travail s'appuie sur la phraséologie transdisciplinaire des écrits scientifiques, ces expressions récurrentes comme *résultats encourageants, problème complexe, présenter un avantage*, essentielles pour montrer son appartenance à la « tribu académique » (Swales 1990), qui traversent les disciplines et portent principalement sur les objets et la démarche scientifiques.

André VALLI - TALEP LIF UMR 6166 CNRS, Université de Provence (France)**L'importance du corpus du Web pour l'établissement d'une description satisfaisante de la valence verbale en français**

Nous avons entrepris la réalisation d'un lexique de la valence verbale du français qui présente des similitudes fortes avec l'ouvrage de Herbst et al. (2004). Nous proposons également une interface sémantique « soft », mais tout l'effort a porté, comme dans le *Valency Dictionary* sur la construction d'un lexique hautement syntaxique, avec formalisation des compléments dépendants spécifiques des verbes inventoriés ; enfin notre lexique est également « corpus based ». C'est précisément sur ce point que nous souhaiterions intervenir en montrant comment nous avons été conduits à adopter comme corpus de référence, non pas un grand corpus échantillonné comme les auteurs du *Valency Dictionary of English*, mais l'ensemble des corpus des grands dictionnaires du français (*Le Robert*, *le TLF*, *Les Verbes français*), le Lexique-Grammaire de M. Gross et enfin le corpus du Web, exploité manuellement à partir de *Google*. [...]

Elodie VIALLETON, Tim LEWIS - The Open University (Royaume-Uni)**Reconsidering authenticity: using spontaneous speech to develop listening skills in adult beginner distance learners of French**

This paper reports on a study in the use of naturally-occurring listening materials with adult beginner distance learners of French. The paper re-examines notions of authenticity, as applied to (oral) texts, context, task and purpose, in relation to the development of the skill of listening. The notion of authenticity is then confronted to the types of oral texts found in textbooks currently in use in France and in the UK. Then the paper presents an experimental study which examines the feasibility of teaching the features of spontaneous speech to learners of French.

Jean-Marie VIPREY - Archives, Textes, Sciences des Textes - EA 3187, Université de Franche-Comté, MSHE de Franche-Comté (France)**Big is beautiful, but full text online even more so**

La linguistique textuelle (Adam, Rastier) n'est pas une région des sciences du langage, mais un parti théorique qui conditionne la réflexion sur bases et corpus. Comme l'illustre « ngrams.googlelabs », les données même surabondantes sont d'un intérêt minime tant qu'elles ne sont pas disposées en plein texte dans un format d'échange universel, et implantables librement sur un système local. Nous voulons le démontrer, par l'exposition et l'approche d'une base de presse quotidienne française (5 grands titres sur 1995-2010, 500.000 articles, un milliard de mots) qu'il a fallu constituer et porter à un format TEI.

Nous nous sommes proposé d'appliquer à cette masse structurée les méthodes d'investigation lexicale liées à la *cooccurrence généralisée* (Viprey) et à l'exploration-structuration de celle-ci par les techniques d'analyse statistique multidimensionnelle (analyse factorielle des correspondances, *isotropie*). De telles analyses sont impossibles dans le cadre des accès actuellement ménagés aux ressources textuelles par les bases existantes.

Aude WIRTH-JAILLARD - LAMOP UMR 8589 CNRS, Université Panthéon-Sorbonne - Paris 1 (France) & Institut Émilie du Châtelet (France)

Des sources de données nouvelles pour le linguiste, les documents comptables médiévaux

Contrairement à ce que l'on pourrait penser, les documents comptables médiévaux ne sont pas composés de suites de chiffres uniquement : ce sont des textes rédigés, dans lesquels les recettes et les dépenses sont présentées et justifiées. Datées et localisées, originales, abondantes, ces sources constituent un matériau particulièrement intéressant. Pourtant, rares sont jusqu'à présent les linguistes à les avoir prises en considération.

Cette présentation proposera une réflexion sur les différentes possibilités d'études linguistiques offertes par ces sources et sur les moyens d'exploiter au mieux ces dernières ; elle portera notamment sur les cas de discours rapportés (accusations, moqueries, insultes, etc.) présents dans les amendes que l'on peut relever dans les recettes.

Inka WISSNER - ATILF UMR 7118 CNRS, Nancy-Université (France)

Données textuelles et diatopie : quels corpus pour l'étude de variétés diatopiques du français, et comment les exploiter ?

Dans le domaine de la linguistique descriptive, la présente communication aborde le problème de l'exploitation des données textuelles pour l'étude de variétés diatopiques du français contemporain. En partant de l'exemple d'une variété régionale en France, on proposera un relevé des ressources lexicales et des bases de données (y compris numériques) qui sont disponibles, et on se demandera comment elles peuvent être exploitées dans une optique diatopique. Pour des raisons de transparence méthodologique, l'auteure plaide en faveur d'un point de départ commun : un ensemble textuel regroupant les ressources de référence, qui portent à la fois sur la langue générale et sur les variétés diatopiques du français.

Selim YILMAZ¹, Arsun URAS YIMAZ² - ¹Université de Marmara (Turquie), ²Université d'Istanbul (Turquie)

Comment interpréter les constituants énonciatifs du turc avec les méthodes d'analyse du français oral ?

Dans cette étude, nous nous proposons d'interpréter la segmentation de l'énoncé oral en langue turque dans le cadre de la méthode d'analyse du français parlé adoptée par Blanche-

Benveniste (1997), Morel et Danon-Boileau (1998). L'objectif principal de cette présente recherche est de démontrer comment interpréter les constituants de l'énoncé oral turc dans le cadre de différents contextes et à la lumière des théories énonciatives françaises. Le turc est une langue altaïque qui fonctionne uniquement par le procédé de suffixation. L'ordre des mots enchaînés est « (S)OV+pers ». Les variations et enchaînements morphologiques constituent le système de base de la langue. Nous allons démontrer que l'énoncé oral en turc peut avoir diverses structures morphosyntaxiques. Cette analyse nous permettra de définir la segmentation et l'interprétation des opérations de *thématisation* et de *rhématisation* dans le dialogue oral en turc. Nous utiliserons la théorie de l'énonciation en relation avec la prosodie utilisée actuellement en linguistique française contemporaine.