UNIL | Université de Lausanne
Faculté des lettres

LALDIM

CUSO
CONFÉRENCE UNIVERSITAIRE
DE SUISSE OCCIDENTALE

LA CONSTITUTION ET LE TRAITEMENT DES GRANDS CORPUS:

JOURNEES D'ETUDE
DE L'ECOLE DOCTORALE EN SCIENCES DU LANGAGE
MODULE 4 : ANALYSE DU DISCOURS

Jeudi 15 et Vendredi 16 mai 2008

Université de Lausanne

organisé par le LALDiM (Laboratoire d'analyse linguistique des discours médiatiques) et l'EDSL, et financé par la CUSO et l'EDSL

PROGRAMME

LA CONSTITUTION ET LE TRAITEMENT DES GRANDS CORPUS

Jeudi 15 mai 2008
Bâtiment Unithèque, salle 4202

08h30-08h45	Jean-Michel ADAM (Université de Lausanne) : ouverture du colloque
08h45-10h15	Jean-Marie VIPREY (Université de Franche-Comté, Laseldi) : *La linguistique textuelle à l'orée du numériquxx*
10h15-10h45	Discussion
10h45-11h00	*Pause*
11h00-12h30	Virginie LETHIER (Université de Franche-Comté, Laseldi) : *Exploration textuelle d'un discours de presse régionale du XIXe siècle*
12h30-13h00	Discussion
13h15-14h30	*Repas au restaurant de Dorigny*
14h45-15h30	Cécile GRIVAZ (Université de Genève) : *Annotation de relations causales*
15h30-16h00	Discussion
16h00-16h15	*Pause*
16h15-17h00	Loïse BILLAT (Universités de Lausanne et Neuchâtel) : *Un « discours néolibéral » ? Quantifier et classer les spécificités lexicales et logico-syntaxiques d'un corpus néoclassique*
17h00-17h30	Discussion
17h45	Vin d'honneur offert par le Rectorat de l'Université de Lausanne

Vendredi 16 mai 2008
Bâtiment Unithèque, salle 4202

09h00-10h30	Damon MAYAFFRE (Université de Nice) : *Pour une lecture alpha-numérique des corpus textuels*
10h30-11h00	Discussion
11h00-11h45	Simone WIDLER (Université de Lausanne) : *Le traitement informatique du corpus : l'exemple des contes en prose de Perrault*
11h45-12h15	Discussion
12h30-13h45	*Repas au restaurant universitaire de Dorigny*
14h00-14h45	François BAVAUD (Université de Lausanne) : *Sur la simplicité des modèles et leur adéquation aux corpus (partie 1)*
14h45-15h15	Discussion
15h15-15h30	*Repas au restaurant universitaire de Dorigny*
15h30-16h15	Aris XANTHOS (Université de Lausanne) : *Sur la simplicité des modèles et leur adéquation aux corpus (partie 2)*
16h15-16h45	Discussion
16h45-18h00	Table ronde et synthèse des observateurs : Jean-Michel ADAM et Marcel BURGER (Université de Lausanne)
18h15	Fin du colloque

Journées d'étude de l'Ecole doctorale en Sciences du langage
Module 4 : Analyse du discours

La constitution et le traitement des grands corpus

Thème. La linguistique du texte et la linguistique du discours se posent de manière récurrente la question de la constitution et du traitement de grands corpus textuels. Cette question centrale rejoint celle, tout aussi importante, des bases de données permettant de traiter les unités linguistiques textuelles. De fait, le développement des sciences informatiques et la numérisation des textes et des corpus, modifie sensiblement la manière de concevoir ces problématiques jusqu'à imposer une redéfinition même des notions de textualité et de discours

La problématique du traitement des grands corpus est d'actualité comme en témoignent plusieurs conférences et colloques internationaux récents à l'étranger comme en Suisse (Albi 2006, JADT 2004 et 2006, IPrA 2007, DICOEN 2007, par exemple ; voir aussi la venue de Douglas Biber, Arizona State University, pour la VALS/ASLA 2006, pour la SSL novembre 2008). Plus modestement, ces deux journées d'études de l'Ecole doctorale CUSO seront l'occasion de recevoir deux équipes de recherches françaises proches des préoccupations lausannoises et dont les travaux sont d'intérêt pour tout chercheur dans le domaine de l'analyse des discours.

Conférenciers invités

Prof. Jean-Marie Viprey du Laboratoire LASELDI de l'Université de Franche-Comté. Animateur du pôle « *Archives, Bases, Corpus* » de la Maison des Sciences de l'Homme de Franche-Comté. Spécialisé dans l'analyse informatique des données textuelles. Le LASELDI place au centre de ses travaux une question qui concerne les doctorants dans le domaine de l'analyse des discours et qui pourrait être abordée lors de la première journée d'études de l'Ecole doctorale: « Qu'est-ce qu'un texte à l'ère numérique ? » (possibilité d'inviter aussi Margareta KASTBERG SJÖBLOM qui fait partie de ce même labo).

Prof. Damon Mayaffre, Université de Nice Sophia Antipolis, responsable de l'équipe *« Bases, Corpus et Langage »*, également spécialisé dans l'analyse informatique des données textuelles. Les recherches niçoises sont d'intérêt pour les doctorants notamment pour ce qui concerne les travaux consacrés au discours politique : traitement numérique des discours de Thorez, Blum, Tardieu et Flandin entre 1930 et 1939 ; discours de la cohabitation Jospin-Chirac entre 1997 et 2002 ; et discours présidentiels sous la Vème République. Ces travaux ont été publiés notamment dans les « Journées d'Analyse des Données Textuelles » 2002, 2004 et 2006 (JADT 2002, 2004, 2006).

Intervenants. Six autres exposés sont prévus qui témoignent d'un ancrage intellectuel diversifié. En effet, outre le point de vue des linguistes et analystes du discours, ces journées d'études sont l'occasion de nouer le dialogue avec des chercheurs dans le domaine des méthodes quantitatives et mathématiques, ainsi que des chercheurs dans le domaine du traitement informatique du langage.

Les intervenants suivants sont confirmés, chercheurs confirmés et doctorant-es confondus: Virginie Lethier (Université de Fraqnche-Comté), Cécile Grivaz (Université de Genève), Loïse Bilat (Université de Lausanne), Simone Widler (Université de Lausanne), François Bavaud (Université de Lausanne) et Aris Xanthos (Université de Lausanne).

RESUMES DES EXPOSES ET NOTICES BIOBIBLIOGRAPHIQUES DES INTERVENANTS

JEAN-MARIE VIPREY

Université de Franche-Comté
E-mail : jean-marie.viprey@univ-fcomte.fr

La linguistique textuelle à l'orée du numérique

La linguistique textuelle s'impose comme le champ conflictuel des théories de la production co(n)-textuelle de sens. Il est dès lors tentant de « définir » le *texte* comme l'objet de la linguistique textuelle. On commencera néanmoins par un travail définitionnel plus exigeant, à partir des notions qui forment notre programme *Archive/s Bases Corpus*, pour fonder l'idée que l'*Analyse Textuelle du Discours (ATD)* nous invite à atteindre la *langue* DANS le *discours* PAR les *textes*, la *langue* comme fait de discours donc, ainsi que l'ont montrée le plus audacieusement Renée Balibar, Jacques Guilhaumou entre autres, mais telle qu'attestée ou *attestable* en corpus, récolté par les sciences humaines. Notre programme implique le refus d'un hiatus confirmé entre les dimensions *logico-grammaticale* et *pragmatico-herméneutique* des sciences du texte.
Le *texte* à l'ère numérique est plus que jamais la *matérialité* du discours, entendue comme un construit réflexif, complexe et pluriel. Il est aussi l'objet d'une *philologie* renouvelée et élargie. Il est constitué de *documents* et *d'énoncés* mais ne se réduit en aucun cas ni à l'un ni à l'autre.
Dans toutes les humanités, l'analyse textuelle doit être constamment réhabilitée. Mais elle ne le sera que si elle sait historiciser ses propres exigences et procédures. Nous proposons d' »y contribuer notamment en intégrant (sans les confondre) dans un même ensemble conceptuel et logiciel l'établissement du texte (qui n'est pas le lieu d'une vérité unique mais d'un contrôle rigoureux et collectif), son exploration assistée, son interprétation, dans une alternance dont la sortie n'est jamais que provisoire, étant elle-même un discours et potentiellement une « couche » de l'ensemble textuel. Le vertige herméneutique doit être affronté sans détour, mais dans un cadre théorique, critique et opérationnel discuté aux yeux de tous.
Nous nous appuierons sur les diverses applications en cours à Besançon et en coopération : littéraires (Baudelaire, Claudel), historico-médiatiques (Presse régionale du XIXème siècle), sociologiques (commandite régionale sur les représentations de l'Europe), linguistiques (formalisation de la progression thématique et de la cohésion textuelle). La conférence permettra une sensibilisation à la Text Encoding Initiative et aux contraintes du balisage des textes, et elle donnera quelques aperçus sur les outils et méthodes informatisés développés à Besançon..

Jean-Marie Viprey est Professeur de Langue et Littérature françaises à l'Université de Franche-Comté, directeur-adjoint de l'Equipe d'Accueil 3187 *Archives, Textes, Sciences des Textes* et membre associé de l'EA 2181 *Laseldi*. Il est coordonnateur du pôle *Archive, Bases, Corpus* de la Maison des Sciences de l'Homme de Franche-Comté. Il est actuellement en délégation au Centre National de la Recherche Scientifique –CNRS- dans l'UMR *Bases, Corpus, Langages* de l'Université de Nice.

Publications récentes :

• 2008 avec Virginie Lethier (coordonnateurs) *Semen n°25*, *Le Discours de presse au XIXème siècle : pratiques socio-discursives émergentes*. – Besançon, Presses Universitaires de Franche-Comté

•2008 avec Virginie Lethier «Annotation linguistique de corpus : vers l'exhaustivité par la convialité.» in *JADT'09, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*. – Lyon, Presses Universitaires de Lyon

• 2008 avec Alpha Ousmane Barry «Approche comparative des résultats d'exploration textuelle des discours de deux leaders africains : Modibo Keita et Sékou Touré » in *JADT'09, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*. – Lyon, Presses Universitaires de Lyon

• 2006 (coordonnateur) *JADT'06 , 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. – Besançon, Presses Universitaires de Franche-Comté.

•2006 « Philologie numérique et herméneutique intégrative » in *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité* dir. Jean-Michel Adam & Ute Heidman. – Genève : Slatkine (pp. 51-68)

• 2006 « About Labbé's intertextual distance » in *Journal of Quantitatie Linguistics* vol.13 n° 2-3 Août-Décembre 2006, Routledge (pp.164-284)

• 2006 « …un de ces syntagmes qui… » in *Corpus* n°5, *Corpus et stylistique*, 2006, CNRS-UNSA.

• 2006 « Quelle place pour les sciences des textes dans l'Analyse de Discours » in *Semen* n° 21 *Catégories pour l'analyse du discours politique*, Besançon, Presses Universitaires de Franche-Comté (pp.167-182)

• 2006 « Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection 'géodésique' » in *JADT'06, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. – Besançon, Presses Universitaires de Franche-Comté.

• 2005 « Méthodes pour la lecture des corpus » in *Sémantique et corpus* dir. Anne Condamines. – Hermès

• 2005 « Structure non séquentielle du texte » in *Langages* n° 161, *Unité(s) du texte* dir. Dominique Legallois. – Paris : Larousse. (pp. 65-82)

• 2002 *Analyses textuelles et hypertextuelles des* Fleurs du mal - Champion, Paris.

• 1997 *Dynamique du vocabulaire des* Fleurs du mal - Champion, Paris. (Prix International de la Fondation Paul-Robert 1998).

VIRGINIE LETHIER

Université de Franche-Comté
E-mail : virginie.lethier@yahoo.fr

Exploration textuelle d'un discours de presse régionale du XIXe siècle

Longtemps demeurée dans l'escarcelle des historiens, la presse régionale du XIXe siècle constitue un témoignage précieux de la vie sociopolitique, économique et littéraire de ce siècle fondateur. Aussi dans le cadre du pôle *Archive, Bases, Corpus* de la MSH de Franche-Comté, en partenariat avec la Bibliothèque d'Etudes de Besançon, a été entreprise la constitution d'un fonds numérisé de presse régionale du 19ème siècle à partir de la collection du *Petit Comtois*, quotidien républicain, paru de 1883 à 1945. De cette base a été extrait un corpus de plus de 5 millions de mots, soit plus de 22500 articles, représentant une diachronie de 20 ans (1883-1903), envisagé comme un dispositif d'observation propre à révéler les mutations d'un discours journalistique dans un contexte de professionnalisation de la pratique sociale.

Une problématique centrale structure nos recherches menées sur ce corpus, ancrées dans le cadre d'une *analyse textuelle des discours* (Adam & Heidmann, 2005) assistée par la statistique textuelle: celle de l'*exploration* des multiples dimensions du texte. De l'étape déterminante de l'établissement du texte à celle de l'exploration proprement dite, l'analyse textuelle des discours (ATD) implique en effet une appréhension optimale de la matérialité du texte, palliant en cela le *déficit philologique* de l'Analyse du discours décrit par F. Rastier et J.-M. Adam. Aussi souhaiterions-nous examiner dans la présente contribution le potentiel renouveau de la lecture de la presse du XIXe siècle induite par une *nouvelle philologie numérique*, telle qu'appelée par J.-M. Viprey (à la suite de Rastier, 2001). Illustrant notre propos par la présentation de notre traitement du corpus du *Petit Comtois*, envisagé comme une succession de prises de vues sur la matérialité textuelle, nous ferons état des apports heuristiques d'une exploration dynamique telle que proposée par le logiciel Astartex-Diatag, développé à MSH de Franche-Comté.

S'appuyant sur des outils statistiques variés, notamment l'analyse factorielle de relevés lexicaux (distribution massive dans les articles, les classes d'articles, la diachronie ; distribution fine dans la cooccurrence lexicale), cet environnement offre des vues synthétiques de grands ensembles de données complexes à même d'établir et d'enrichir un dialogue interprétatif avec le corpus, tout en favorisant le *retour au texte*, puisque les résultats et toutes les ressources configurant le texte (tableaux, listes, dictionnaires) sont organisés en hypertexte expert.

Virginie Lethier est doctorante allocataire-monitrice au Centre Jacques Petit, ATST, de Besançon, France.

Publications récentes :

• Lethier, V., 2007, « Constitution d'un corpus de presse régionale du XIXe siècle : pratiques et enjeux », à paraître en ligne : archivesic.ccsd.cnrs.fr

• Viprey, J.-M, Lethier, V., 2008 (à paraître): « Lire l'archive, presse écrite du XIXe siècle », 5ème journée des linguistiques de corpus

• Viprey J.-M., Lethier, V., 2008 (à paraître) « Annotation linguistique de corpus : vers l'exhaustivité par la convivialité» in *JADT 2008, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*. (ENS SHS Lyon)

CECILE GRIVAZ

Université de Genève
E-mail : cecile@grivaz.net

Annotation de relations causales

Dans le cadre d'un projet de développement d'un programme capable de reconnaître automatiquement les expressions de relations causales dans des textes en langue naturelle, nous voulons annoter un corpus en repérant les relations causales qui y sont exprimées. Il n'existe pas à notre connaissance de ressource similaire en français ou en anglais, et ce corpus permettra de tester la qualité de notre système automatique et d'étudier des cas concrets d'expression de relations causales. Il pourra également être distribué à la communauté scientifique.

Les travaux de Inui (2005) qui a réalisé un corpus similaire en japonais, montrent que l'annotation de relations causales est diffcile et peut résulter en un accord inter annotateurs faible. Notre travail sera donc de construire un ensemble de consignes d'annotation qui servira de guide aux annotateurs afin d'augmenter leur accord. Pour obtenir un bon ensemble de consignes, nous les testerons sur des échantillons, et nous les raffinerons jusqu'à ce que nous obtenions un jeu de consignes qui permette un accord inter annotaeurs suffisamment haut. Cette méthodologie se base sur le travail de Hovy et de ses collègues (2006) qui ont décrit un système similaire permettant d'obtenir un accord inter annotateurs haut pour la tâche qui consiste à relier les mots d'un corpus aux concepts correspondants dans une ontologie. Nos consignes seront basées sur des caractéristiques théoriques ainsi qu'intuitives de la causalité. Nous avons obtenu une liste de caractéristiques intuitives grâce à des expériences dans lesquelles on demande à un sujet de justifier sa lecture causale ou non causale d'un énoncé.

Nous avons choisi un corpus bilingue qui peut être distribué librement. Il s'agit du corpus BAF (bitextes anglais-français) qui a été construit par le groupe de recherche appliquée en linguistique informatique de Montréal. Nous avons choisi un corpus bilingue aligné afin de faciliter l'exportation de nos annotations vers l'anglais, ce qui permettrais de créer rapidement une ressource similaire dans cette langue.

Cécile Grivaz : le projet de diplôme en informatique à l'EPFL (Lausanne, Suisse) de Cecile Grivaz porte sur l'extraction automatique d'entités dans des textes en langue naturelle. Elle a ensuite fait un *DEA* en linguistique à l'université de Genève lors duquel elle s'est intéressée aux analyseurs syntaxiques. Elle est actuellement en première année de thèse sous la supervision de Jacques Moeschler du département de linguistique de l'université de Genève et de Martin Rajman du laboratoire d'intelligence artificielle de l'EPFL. Son sujet de thèse est l'extraction automatique de relations causales exprimées dans des textes en français. Il s'agit de réaliser un programme informatique capable de détecter dans un texte des relations causales telles que *Max est tombé, Jean l'a poussé*

Références pertinentes

• Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw and Ralph M. Weischedel. 2006. *Ontonotes: The 90% Solution*. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, HLT-NAACL. The Association for Computational Linguistics.

• Takashi Inui. 2005. *Creating an annotated corpus for the analysis of causal relations*. COE-LKR2005.

LOISE BILAT

Universités de Lausanne et de Neuchâtel
E-mail : loise.bilat@unil.ch

Un « discours néolibéral » ? Quantifier et classer les spécificités lexicales et logico-syntaxiques d'un corpus néoclassique

De quelle manière isoler les « régularités discursives » et représenter les transformations de la pensée néoclassique en tant qu'utilisation particulière de la langue de l'après-guerre à l'aube du vingt-et-unième siècle ?
Les collocations, ainsi que les antagonismes lexicaux récurrents mis à jour chez trois auteurs significatifs (F. von Hayek en 1944 et 46; J. Buchanan en 1975 et 92; P. Bruckner en 2002), permettent de délimiter ce qui semble constituer un « discours néolibéral ». Accompagnant cette régularité, un changement syntaxique illustre bien l'acceptation progressive de ce discours. D'une majorité de séquences argumentatives dans l'ouvrage de 1944, les marqueurs explicatifs paraissent dominer chez un économiste américain en 1975.
Comment traiter statistiquement ces phénomènes ? Outre l'important problème de la représentativité d'un corpus plus large, les notions de classe-objet, de faisceau et d'éclairage de Jean-Blaise Grize pourraient s'avérer intéressantes pour classer des récurrences lexicales tout en prenant en compte leur environnement syntaxique..

Loise Bilat vient d'obtenir une licence en Lettres à l'université de Lausanne où elle a présenté son mémoire de licence sous la direction de J.M. Adam. Elle est nouvellement assistante à l'institut d'Information et communication de l'Université de Neuchâtel.

DAMON MAYAFFRE

Université de Nice
E-mail : DamonMayaffre@wanadoo.fr

Pour une lecture *alpha-numérique* des corpus textuels

La linguistique du texte ne peut être introspective et travaille nécessairement sur corpus. Le passage du papier au numérique des corpus textuels ne représente pas un changement technique de support mais une révolution culturelle et épistémologique dans la mesure où elle modifie notre approche de l'objet-texte et de l'acte-lire.
J'essayerai d'illustrer concrètement cette révolution, sur de grands corpus politiques, en croisant approche qualitative et approche quantitative des textes grâce au logiciel Hyperbase. Il sera question aussi bien d'alphabet que de chiffres, de navigation hypetextuelle ou de recherche documentaire que de traitements statistiques ou mathématiques de la textualité.

.

Damon Mayaffre est chercheur au CNRS, et Chargé de cours à l'Université de Nice, France. Il est spécialiste d'analyse du discours politique assistée par ordinateur.

Publications récentes :
• Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République, Paris, Champion, 2004.
• "Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques", in François Rastier et Michel Ballabriga (éds), Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation, Toulouse, Put, 2007, pp. 15-26.
• "Analyses logométriques et rhétoriques des discours", in Stéphane Olivési (dir.), Introduction à la recherche en SIC, Grenoble, Pug, 2007, pp. 153-180.
• "L'analyse de données textuelles aujourd'hui : du corpus comme une urne, au corpus comme un plan. Bilan sur les travaux actuels de topographie/topologie textuelle", Lexicométrica, 2007.

SIMONE WIDLER

Université de Lausanne
E-mail : simone.widler@unil.ch

Le traitement informatique du corpus : l'exemple des contes en prose de Perrault

L'analyse lexicale des contes de Perrault par des logiciels confirme quelques pistes novatrices que des observations à la main ont partiellement signalées. Les termes intensificateurs ('un *fort* honnête homme', 'une *si* belle princesse', etc.), pour lesquels on a opté dans notre travail, se combinent et se distribuent de manière bien particulière dans le corpus. Une des plus grandes difficultés consiste à interpréter les résultats surprenants fournis par les logiciels. Comment peut-on mettre en rapport la distribution des intensifs avec l'écriture de Perrault ?
En ce qui concerne la méthode, tout praticien sait que les résultats fournis par les programmes informatiques reposent sur les décisions propres du chercheur. Sous quelle perspective pourrait-il être fructueux de considérer un corpus aussi éclaté et travaillé comme celui des contes de Perrault? Laquelle parmi les versions textuelles est la plus susceptible de favoriser la découverte de connaissances nouvelles ? Où réside l'intérêt de l'informatique de traiter un corpus de taille relativement réduite comme c'est le cas du recueil des contes ?
A la question du corpus s'ajoutent des aspects pratiques : comment peut-on automatiquement découper le recueil des contes en ses unités signifiantes (parties du conte, unités thématiques et discursives) ? Quelles sont les différentes options de balisage textuel ? En fait, celles-ci ne s'adaptent pas toujours aux besoins de l'utilisateur et s'avèrent peu compatibles d'un logiciel à l'autre.
En passant à des unités plus petites comme les lexèmes, la question du regroupement des mots se pose. Jusqu'à quel point est-il raisonnable d'associer les formes du français du XVIIe siècle afin de rendre les résultats plus pertinents ? Comment peut-on repérer la classe dite des intensifs avec le moindre effort possible?
Dans notre analyse, le choix des méthodes est motivé, mais aussi conditionné par l'outil informatique. En fait, personne n'a la vue d'ensemble des programmes accessibles sur Internet et les logiciels gratuits présentent souvent des défauts considérables. De plus, chaque logiciel est limité à quelques fonctions précises et ce n'est que dans leur complémentarité qu'il est possible d'approfondir un maximum l'analyse textuelle.

Simone Widler est chargée de recherche à l'Université de Lausanne où elle a étudié le français, l'espagnol, l'informatique et les méthodes mathématiques. En été 2007, elle a présenté son mémoire de licence sous la direction de J.M. Adam intitulé « Analyse linguistique et informatique des données textuelles : les intensifs dans les contes en prose de Perrault ».

Références pertinentes

BEAUDOUIN, Valérie (2000) : « Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle », *Texto !*, [en ligne].
Disponible sur : http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html#1 (juillet 07).
JEANNERET, Thérèse (2005) : « Consécutives intensives et mouvement du sens dans quelques contes de Perrault, Grimm et Andersen », *Le Français moderne*, n°1-73ème année.
PERRAULT, Charles : *Contes de Perrault, textes établis par Gilbert Rouger,* Paris, Ed. Garnier, 1975.
VIPREY, Jean-Marie (2006) : « Structure non-séquentielle des textes », Langages 163, septembre 2006, 71-85.
WIDLER, Simone, *Analyse linguistique et informatique des données textuelles : les intensifs dans les contes en prose de Perrault*, mémoire de licence, Université de Lausanne, 2007.

FRANÇOIS BAVAUD

Université de Lausanne
E-mail : francois.bavaud@unil.ch

Sur la simplicité des modèles et leur adéquation aux corpus

Dans une large mesure, l'activité descriptive en sciences du langage consiste à analyser des corpus, c'est-à-dire construire des modèles pour expliquer les régularités observées dans les données. Qu'on s'intéresse à la structure phonologique, syntaxique, discursive, etc., cette démarche s'effectue en principe dans le cadre d'une famille de modèles spécifique, décrivant tel ou tel type d'unités ou de relations. La tâche de l'analyste revient alors à sélectionner, parmi les modèles appartenant à cette famille, celui qui lui semble fournir la meilleure description du corpus.

Dans cette contribution jointe, nous nous interrogerons sur ce que signifie, pour un modèle, de fournir une bonne description d'un corpus. Traditionnellement, dans les sciences du langage, l'adéquation d'une analyse particulière d'un jeu de données est évaluée en termes subjectifs, comme l'élégance, la concision, ou encore le pouvoir explicatif. Il se trouve que ces aspects, dont l'intuition saisit bien l'importance, peuvent être quantifiés en termes objectifs, dépendant uniquement de la famille de modèles considérée et du corpus. Cette approche, que nous illustrerons dans le cas de l'analyse phonologique et de la statistique textuelle, est applicable à tous les domaines des sciences du langage, et à vrai dire, des sciences en général.

François Bavaud, est professeur associé de méthodes quantitatives, en faculté des Lettres à l'Université de Lausanne. Ses recherches portent sur l'analyse et la modélisation des données textuelles. Thèmes spécifiques: théorie de l'information, mécanique statistique, méthodes factorielles généralisées, classification et visualisation.

ARIS XANTHOS

Université de Lausanne
E-mail : aris.xanthos@unil.ch

Sur la simplicité des modèles et leur adéquation aux corpus (partie 2)

Dans une large mesure, l'activité descriptive en sciences du langage consiste à analyser des corpus, c'est-à-dire construire des modèles pour expliquer les régularités observées dans les données. Qu'on s'intéresse à la structure phonologique, syntaxique, discursive, etc., cette démarche s'effectue en principe dans le cadre d'une famille de modèles spécifique, décrivant tel ou tel type d'unités ou de relations. La tâche de l'analyste revient alors à sélectionner, parmi les modèles appartenant à cette famille, celui qui lui semble fournir la meilleure description du corpus.

Dans cette contribution jointe, nous nous interrogerons sur ce que signifie, pour un modèle, de fournir une bonne description d'un corpus. Traditionnellement, dans les sciences du langage, l'adéquation d'une analyse particulière d'un jeu de données est évaluée en termes subjectifs, comme l'élégance, la concision, ou encore le pouvoir explicatif. Il se trouve que ces aspects, dont l'intuition saisit bien l'importance, peuvent être quantifiés en termes objectifs, dépendant uniquement de la famille de modèles considérée et du corpus. Cette approche, que nous illustrerons dans le cas de l'analyse phonologique et de la statistique textuelle, est applicable à tous les domaines des sciences du langage, et à vrai dire, des sciences en général.
.

Aris Xanthos est premier assistant à la faculté des Lettres de l'Université de Lausanne, dans les sections d'informatique et méthodes mathématiques et de linguistique. Sa thèse de doctorat, soutenue à Lausanne en 2007 et partiellement réalisée au cours d'un séjour au département de linguistique de l'Université de Chicago, porte sur l'apprentissage automatique de la phonologie et la morphologie. Plus généralement, sa recherche est centrée sur la modélisation de l'acquisition du langage, par des humains ou par des machines.

ORGANISATION ET RENSEIGNEMENTS

Responsable :

Jean-Michel Adam et Marcel Burger
UNIL, LALDiM - Faculté des Lettres
Bâtiment Anthropole, Bureau 3024
CH-1015 Lausanne
Tél. : ++41/(0)21/692 29 48
Email : marcel.burger@unil.ch

Site internet : http://www.unil.ch/laldim et http://www.unige.ch/lettres/linguistique/edsl

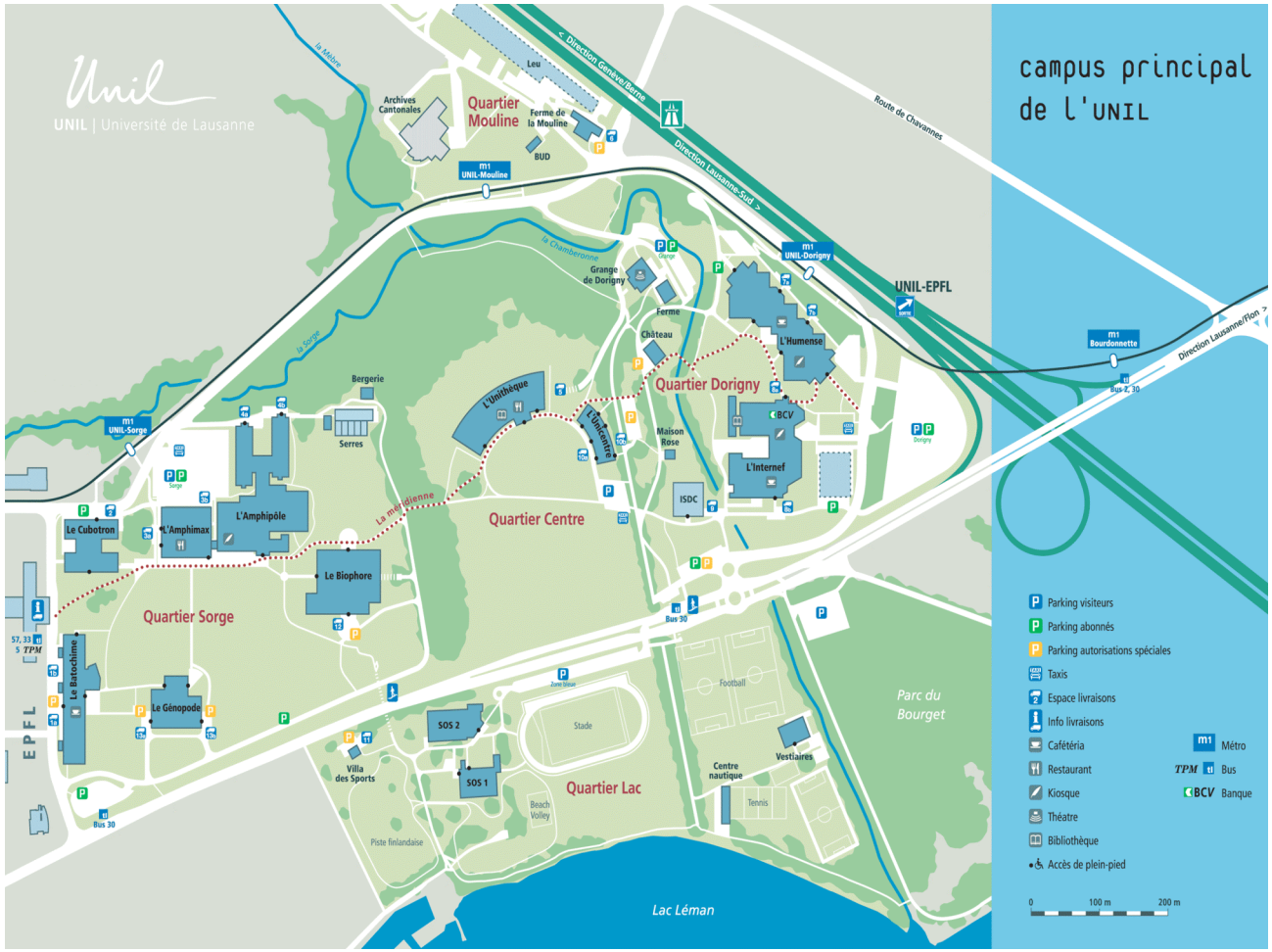
Lieux du colloque:

Université de Lausanne
Salle de conférence Unithèque 4202
CH-1015 Lausanne

Pour se rendre aux salles de conférence:

Par le train et les transports publics : Les deux salles de conférences se trouvent sur le site de l'Université de Lausanne. Pour vous rendre à l'Université, prenez le bus spécial "Métro-Bus" (MB) à la place de la Gare en direction de MONTBENON; il y en a un toutes les 6 minutes environ. Ce bus remplace l'ancien métro qui circulait d'Ouchy au centre-ville. A Montbenon, traversez la route et prenez la passerelle jusqu'à l'ascenseur du métro. Prenez ensuite le Métro-Ouest appelé TSOL jusqu'à la station "UNIL-Dorigny" (sixième arrêt). Il ne vous reste plus qu'à suivre les écriteaux pour l'Institut de droit comparé (ISDC), situé à 2 minutes, et pour le Bâtiment Humense (littéralement à côté de l'arrêt du TSOL).

En taxi : le prix approximatif du trajet de la gare CFF à l'Université s'élève à environ CHF 30.


Unil
UNIL | Université de Lausanne
campus principal de l'UNIL
Quartier Mouline
Archives Cantonales
Ferme de la Moulline
BUD
Leu
m1 UNIL-Mouline
Direction Genève/Berne
Direction Lausanne-Sud
Route de Chavannes
la Sorge
la Chamberonne
Grange de Dorigny
Ferme
Château
m1 UNIL-Dorigny
UNIL-EPFL
L'Humense
Quartier Dorigny
m1 Bourdonnette
Direction Lausanne/Flon
Bus 2, 30
Bergerie
Serres
m1 UNIL-Sorge
L'Unithèque
L'Unicentre
Maison Rose
BCV
L'Internef
ISDC
Le Cubotron
L'Amphimax
L'Amphipôle
La méridienne
Quartier Centre
Le Biophore
Quartier Sorge
Le Batochime
Le Génopode
57, 33 TPM
Bus 30
SOS 2
SOS 1
Villa des Sports
Stade
Football
Tennis
Beach Volley
Piste finlandaise
Centre nautique
Vestiaires
Quartier Lac
Parc du Bourget
Lac Léman
EPFL
Parking visiteurs
Parking abonnés
Parking autorisations spéciales
Taxis
Espace livraisons
Info livraisons
Cafétéria
Restaurant
Kiosque
Théatre
Bibliothèque
Accès de plein-pied
m1 Métro
TPM Bus
BCV Banque
0
100 m
200 m