PROGRAMME SCIENCES HUMAINES ET SOCIALES
EDITION 2007

CORPUS ET OUTILS DE LA RECHERCHE EN SCIENCES HUMAINES ET SOCIALES

APPEL A PROJETS

Destiné aux différentes disciplines des sciences humaines et sociales

FORMULAIRE

Date limite d'envoi des dossiers :
30 mars 2007 à 16H par voie électronique
30 mars 2007 à minuit par courrier postal

Les dossiers doivent être envoyés par courrier électronique (en utilisant le présent formulaire et au format ".doc")
Corpus-anr@ens-lsh.fr
et
par courrier postal en 3 exemplaires (un original et 2 copies) cachet de la poste faisant foi

ENS LSH
Programme ANR Corpus et outils de la recherche en sciences humaines et sociales
15 Parvis René Descartes
BP 7000
69342 Lyon cedex 07

RENSEIGNEMENTS ADMINISTRATIFS ET FINANCIERS
Corpus2007-anr@ens-lsh.fr

RENSEIGNEMENTS SCIENTIFIQUES
pierre-olivier.pin@agencerecherche.fr

FORMULAIRE DE SOUMISSION ET RENSEIGNEMENTS
http://www.agence-nationale-recherche.fr/
http://unitesupportanr.ens-lsh.fr

I - FICHE D'IDENTITE DU PROJET

Titre du projet *(maximum 120 caractères)*

Corpus représentatif des premiers textes français

Acronyme ou titre court *(12 caractères maxi) : CORPTEF*

Mots-clés

Linguistique diachronique, français médiéval, linguistique de corpus, linguistique variationnelle, typologie des textes, TAL

Résumé du projet en 5000 caractères maximum (*Note : ce résumé est susceptible d'être publié en cas de financement du projet par l'ANR)*

1- contexte scientifique et objectifs du projet
2- description du projet, méthodologie
3- résultats attendus

1. Notre projet de recherche vise à rendre disponible à la communauté scientifique comme au grand public un corpus représentatif des textes les plus anciens écrits en français (IXe – XIIe siècles).
Ce projet s'appuie sur l'acquis d'une base textuelle existante, la Base de Français Médiéval (BFM : http://bfm.ens-lsh.fr/) et sur sa méthode de développement et d'exploitation. Utilisée par une communauté de 300 chercheurs environ, français et étrangers, cette base jouit d'une reconnaissance internationale. Elle fait partie des plus grandes bases de français médiéval dans le monde (environ 3 millions d'occurrences-mots), et constitue l'une des assises principales du Consortium international pour les corpus de français médiéval (CCFM, présidé par le Professeur Pierre Kunstmann, de l'Université d'Ottawa, Canada). La BFM comprend d'ores et déjà la quasi totalité des textes antérieurs au XIIe siècle (ils sont peu nombreux au total), ainsi qu'une trentaine de textes du XIIe, une quarantaine du XIIIe, et une trentaine de textes de moyen français (XIVe- fin du XVe siècle). Notre projet, qui vise à compléter et étendre cette base par l'ajout d'un ensemble de textes du XIIe siècle, permettra la diffusion d'un corpus fondamental pour l'ensemble de la période médiévale. Ce corpus sera entièrement lemmatisé, étiqueté en morpho-syntaxe et fera l'objet d'une annotation syntaxique.
Les principaux objectifs que nous souhaitons atteindre sont au nombre de trois :
- développer et diffuser le corpus de français médiéval le plus vaste et le plus diversifié qui soit; d'où l'intérêt de l'enrichir afin, en particulier, de parvenir à un volume de données important (1,7 millions de mots environ) et diversifié pour la période la plus ancienne (IXe-XIIe siècles), ces données étant par ailleurs lemmatisées, étiquetées en morpho-syntaxe et annotées ;
- favoriser le développement de recherches diachroniques sur le français, en particulier parmi les membres de notre équipe et plus largement au sein de la communauté internationale des médiévistes travaillant sur le français ;
- élaborer un cadre méthodologique qui puisse être exploité par d'autres que nous, en premier lieu au sein de la communauté internationale des médiévistes regroupés dans le CCFM.

2. Le projet s'appuie sur la méthodologie de corpus, élaborée dans le cadre de la « linguistique de corpus ». Ce cadre méthodologique doit garantir l'exploitation future des données mises à disposition en nous permettant de maîtriser, diriger et décrire la diversité typologique des documents intégrés au corpus. Il permettra ainsi de définir le degré de représentativité des données exploitées et d'évaluer le degré de généralité des résultats obtenus.
Le corpus que nous souhaitons réaliser s'appuiera sur l'expérience et le savoir-faire acquis depuis près de vingt ans. Nous disposons en effet d'une chaîne éprouvée de traitement des textes (numérisation, relecture, encodage, mise en ligne) et d'un réseau de relecteurs spécialisés. Par ailleurs, nos pratiques d'encodage et

de description des méta-informations textuelles, synthétisés dans un ensemble de documents de référence publiés sur les sites de la BFM et du CCFM, ont été maintes fois exposées et discutées dans le cadre d'échanges internationaux. Elles font appel aux normes et aux formats les plus récents et les mieux partagés dans le monde (format XML, balises TEI). Enfin, le logiciel d'interrogation dont nous disposons est très performant et souple ; en constante évolution, il saura s'adapter à de nouveaux types d'exploitation (son développement est soutenu par le projet « Textométrie » financé par l'ANR).

3. Outre la mise en ligne d'un ensemble particulièrement important de textes anciens, les résultats attendus de ce projet sont de plusieurs types. Ils concernent en premier lieu les recherches linguistiques rendues possibles par le corpus. De par son empan chronologique (IXe-début du XVIe siècle), la diversité typologique et l'équilibrage relatif des différents types de documents qui le composent, ce corpus sera tout à fait exceptionnel pour l'histoire du français. La diffusion d'une part importante des données les plus anciennes disponibles à ce jour rendra possible un ensemble de recherches sur le passage du latin au français et sur le très ancien français (IXe-XIIe). Elle permettra également la recherche des attestations les plus anciennes, ce qui constitue une avancée essentielle pour les recherches portant sur l'évolution de la langue, et en particulier sur l'origine des mots du lexique et de la grammaire, des constructions, et des notions et concepts du français.
Les recherches menées par les membres de notre projet, qui concernent les différents niveaux de l'analyse linguistique (lexicologie, sémantique lexicale, syntaxe, sémantique grammaticale, analyse discursive et socio-linguistique), exploiteront ces données nouvelles. Ce corpus sera également le support d'une grande grammaire historique du français en cours de préparation dans un cadre national. Il favorisera les recherches portant sur la langue française et contribuera au renouveau que connaît dans le monde depuis une vingtaine d'années la linguistique diachronique.

Abstract *(Do not exceed 5000 car.)*

Scientific background and objectives
Description of the project, methodology
Expected results
Title: Corpus representative of the Early French Texts (before 1200)

1. The aim of this project is to create a representative corpus of the early French texts (written between the 9th and the 12th centuries) and to make it available to the scholarly community as well as to a general public. The project is based upon a major online text(ual?) database, the "Base de Français Médieval" – Old French Corpus (BFM, http://bfm.ens-lsh.fr) and the technology developed for its operation. This database is used by a community of over three hundred researchers from France and other countries worldwide, and is therefore internationally recognized. It counts among the world's largest Medieval French databases (3,000,000 occurrences) and constitutes one of the pillars of the International Consortium for Medieval French Corpora (CCFM, chaired by Prof. Pierre Kunstmann, University of Ottawa). The BFM already contains nearly all available texts written before 1100, as well as about thirty 12th century texts, some forty 13th century texts and nearly thirty Middle French texts (14th and 15th centuries). This project intends to complete and extend this collection by adding a considerable number of 12th century texts and thus form a comprehensive text corpus covering the entire Middle Age period . This corpus will be completely lemmatized, morphologically tagged and syntactically parsed.
The proposed project should achieve the following three goals:
- To compile and publish online the largest and most diversified corpus of Medieval French texts. This will require an enlargement of the existing corpus in order to create an important volume of diversified text data (approx. 1,700,000 words) for the oldest historical period (9th – 12th centuries); these texts will also be lemmatized and morphologically and syntactically tagged.
- To enhance diachronic linguistic research on the French language, in particular the work of the members of our team, and more extensively, that of the international community of medievalists working on French.
- To work out a methodological framework which could be used by other researchers, in particular the members of CCFM.

2. The project is based on the methodology of corpus analysis developed in corpus linguistics. This methodological framework will guarantee the future use of the textual data since it will allow us to describe and manage the typological diversity of the documents included in the corpus. It will therefore be possible to measure the representativeness of the available data and to evaluate the possibility of generalizing our results. The corpus will be elaborated with the twenty-year experience and know-how of the BFM team. We have presently established a workflow of text processing (digitizing, proofreading, tagging and online publishing) and formed a team of qualified proofreaders and encoders. Our practice of encoding texts and metatextual data (descriptors) is explained in a number of documents published on the BFM and CCFM websites and has been presented and discussed at numerous international meetings and conferences. This practice is based on the most recent and generally recognized formats and standards (XML format, TEI encoding scheme). We have at our disposal some powerful and flexible software developed in our laboratory for text queries and analysis, in particular Weblex, developed by Serge Heiden. As this software is continuously developed and updated, it will be no doubt possible to adapt it to new uses (thanks in particular to the ANR supported "Textométrie" project).

3. Besides the creation of an online extensive corpus of old texts, the proposed project intends to produce a number of results of various sorts. The first category of results pertains to linguistic research that will be carried out thanks to this corpus. On account of its chronological expanse (9th to early 16th century), its typological diversity and a relative balance between the various types of documents included in the database, the extended BFM corpus will be an outstanding source of data on the history of French. Access to the greater part of the earliest existing textual data will allow considerable advances in the research on the transition from Latin to French and on "very-old" French (9th – 12th centuries). It will also facilitate the search for the earliest occurrences of forms or constructions, which would be a significant progress in the research on the history of French, and in particular on the origin of lexical items, of morphology, syntax and cognitive aspects of French. The research carried out by the members of our team is focused on different levels of linguistic analysis (lexicology, lexical and grammatical semantics, syntax, discourse analysis and sociolinguistics) and we will all benefit from this new textual database. The corpus will also serve as a background for a new detailed description of the historical grammar of French which is currently in preparation. Generally speaking, this corpus will foster research on the French language and will contribute to the revival of diachronic linguistics that has been taking place for a few decades.

Coordinateur du projet (Partenaire 1)

Civilité	Nom	Prénom	Discipline	Laboratoire (nom complet)	Type (établissement public, fondation,

					association, entreprise)
Mme	GUILLOT	Céline	Sciences du langage	UMR 5191 Interactions, Corpus, Apprentissages, Représentations	Etablissement

Nom des responsables scientifiques des autres partenaires

	Civilité	Nom	Prénom	discipline	Laboratoire (nom complet)	Type (établissement public, fondation, association, entreprise)
Partenaire 2						
Partenaire 3						
Partenaire 4						
Partenaire 5						

Nombre de personnes impliquées dans ce projet (en équivalent temps plein : ETP)[1]:

- Personnels permanents ou déjà recrutés :
Chercheurs et enseignants-chercheurs permanents : 1,95
Post-doctorant(s) et doctorant(s) déjà recruté(s) : 0,25
Ingénieurs et techniciens 0,15

- Personnels à recruter : 0,86

Durée du projet : 36 mois[2]

[1] Quelle que soit la catégorie de personnel, il s'agit ici, pour chaque personne impliquée dans le projet, de multiplier son temps de recherche par le pourcentage de temps qu'il consacrera à ce projet.
[2] La durée du projet peut être de 24, 36 ou 48 mois

Dimensionnement total du projet

Coût complet du projet : **€** Reporter le total indiqué au tableau (a) du récapitulatif global (section D du formulaire)

Aide financière demandée : **€** Reporter le total indiqué au tableau (b) du récapitulatif global (section D du formulaire)

Effort en personnel demandé : **homme.mois** Reporter le total indiqué au tableau (c) du récapitulatif global (section D du formulaire)

Je déclare exactes toutes les informations contenues dans ce document et m'engage à envoyer une copie de ce dossier à chacun des établissements ou organismes de rattachement de mon laboratoire

Visa du directeur de laboratoire

Nom, Prénom
Date et signature du
coordinateur du projet
précédé de la mention
« Lu et approuvé »

Nom prénom et signature du directeur du laboratoire

En cas de recouvrement thématique avec d'autres appels à projets (AAP) lancés par l'ANR, le coordinateur du projet est invité à choisir l'AAP le mieux adapté au projet. Les personnes impliquées dans plusieurs AAP de l'ANR en 2007 devront le mentionner dans le tableau « demandes de contrats en cours d'évaluation » (Section D du document).

II - PRESENTATION DETAILLEE DU PROJET

A - Identification du coordinateur et des autres partenaires du projet[3]

Acronyme ou titre court du projet : CORPTEF

A-1 – Partenaire 1 = Coordinateur du Projet

Un coordinateur, responsable scientifique du projet, doit être désigné par les partenaires.

*** champ obligatoire**

Civilité *	**Nom** *	**Prénom** *	
Madame	GUILLOT	Céline	
Grade *	MCF	**Employeur** *	ENS-LSH
Mail *	**cguillot@ens-lsh.fr**		
Tél *	**04 37 37 63 15**	**Fax : 04 37 37 62 65**	

Laboratoire *(nom complet)* *	
Interactions, Corpus, Apprentissages, Représentations	
Code Unité *(s'il existe)* Ex : UMR 5232, EA 567	UMR 5191
Adresse complète du laboratoire *	
Ecole Normale Supérieure Lettres et sciences humaines 15, Parvis René Descartes	
Code postal * **69007**	**Ville** * : Lyon
Etablissements de tutelle *(indiquer le ou les établissements et organismes de rattachement, et en n°1 l'établissement susceptible d'assurer la gestion du projet)* **:** **1.** ENS-LSH **2.** CNRS **3.** Université Lyon2 **4.** ENS Lyon 5. INRP	

[3] Les informations personnelles transmises dans ces formulaires sont obligatoires et seront conservées et seront conservées en fichiers par l'ANR et la structure support mandatée par elle pour assurer la conduite opérationnelle de l'évaluation et de l'administration des dossiers.
Conformément à la loi n°78-17 du 6 janvier 1978 modifiée, relative à l'Informatique, aux Fichiers et aux libertés, les personnes concernées disposent d'un droit d'accès et de rectification des données personnelles les concernant.
Les personnes concernées peuvent exercer ce droit en s'adressant à la structure support (voir coordonnées dans le texte de l'appel à projets) ou l'ANR (212, rue de Bercy, 75012 Paris).

Principales publications :

Liste des 10 principales publications ou brevets de l'équipe partenaire 1 (définie tableau ci-dessous) au cours des cinq dernières années, relevant du domaine de recherche couvert par la présente demande dans l'ordre suivant : Auteurs (en soulignant les auteurs faisant effectivement partie de la demande), Année, Titre, Revue, N°Vol, Pages. N'indiquez pas les publications soumises.

- Carlier, A. (2001), « La genèse de l'article *un* », in : *Langue française,* 130, 65-88.
- Guillot, C. (2006), « Démonstratif et déixis discursive : analyse comparée d'un corpus écrit de français médiéval et d'un corpus oral de français contemporain », in : *Langue française,* 152, 56-69.
- Lavrentiev, A. en collab. avec S. Heiden (2004), « Ressources électroniques pour l'étude des textes médiévaux : approches et outils », in : *Revue française de la linguistique appliquée*, 1, 99 – 118.
- Lusignan, S. (2004), *La langue des rois au Moyen Âge. Le français en France et en Angleterre*, Paris, Presses Universitaires de France, 296 p.
- Marnette, S. (2005), *Speech and Thought Presentation in French Concept and Strategies*, Amsterdam – New York, John Benjamins.
- Möhren, F. (2005), « *Le DEAF - Base d'un atlas linguistique de l'ancien français ?* », in : M.-D. Gleßgen et A. Thibault éd., *La lexicographie différencielle du français et le* Dictionnaire des régionalismes de France, Strasbourg, Presses Universitaires de Strasbourg, 99-113.
- Pignatelli, C. (2006), « Italianismes, provençalismes et autres régionalismes chez Jean d'Antioche traducteur des Otia imperialia », in : C. Galderisi et J. Maurice éd., « Qui tant savoit d'engin et d'art ». Mélanges de philologie médiévale offerts à Gabriel Bianciotto, Poitiers, Université de Poitiers/Centre d'Etudes Supérieures de Civilisation Médiévale avec la collaboration du Centre d'étude et de recherche « Éditer / Interpréter » de l'Université de Rouen, 367-377.
- Prévost, S. (2001), *La postposition du sujet en français aux 15ème et 16ème siècles : une approche sémantico-pragmatique*, Paris, Editions du CNRS.
- Schøsler, L. (2004), « Historical corpora. Problems and methods », in : A. Bozzi, L. Cignon et, J.-L. Lebrave éd., *Linguistica computazionale XX-XXI, Digital technology and philological disciplines*, Pisa, Roma, Istituti editoriale e poligrafici internazionali, 455-472.
- Trotter, D. (2005), *Albucasis, Traitier de Cyrurgie: Edition de la traduction en ancien français de la Chirurgie d'Abū'l Qāsim Halaf Ibn ʿAbbās al-Zahrāwī du manuscrit BNF, français 1318*, in : *Zeitschrift für romanische Philologie, Beiheft* 325, Tübingen, Niemeyer.

Ce projet fait-il partie des projets labellisés (ou en cours de labellisation) par un pôle de compétitivité (ou par plusieurs, en cas de projet interpôle) ? ~~**OUI**~~**/NON**

Si oui, nom du pôle ou des pôles :

Publication d'informations relatives au projet

Si le projet est retenu pour financement, l'ANR se réserve la possibilité de rendre publiques les informations suivantes : le nom du coordinateur du projet et son adresse électronique, les noms des responsables scientifiques et techniques des partenaires du projet, les dénominations des partenaires qu'ils soient des entreprises ou qu'ils appartiennent à un organisme de recherche.

Toutefois, pour un projet de recherche partenariale organisme de recherche / entreprise retenu pour financement, l'ANR ne rendra pas publiques ces informations pour les personnes ou les partenaires pour lesquels la demande lui en est faite. En cas de refus de publication d'un ou de plusieurs de ces éléments, remplacer la mention "OUI" par "NON" dans le tableau ci-après :

	Partenaire n°1 (partenaire coordinateur)	**Partenaire n°2**	**Partenaire n°3**
Nom du responsable scientifique /coordinateur :	**Céline Guillot**	**OUI**	**OUI**
Adresse électronique du coordinateur :	**cguillot@ens-lsh.fr**		

Dénomination du partenaire (si NON, celle-ci sera remplacée, dans le texte publié, par la mention générique « Entreprise » ou « Organisme de recherche »).	**ENS-LSH**	**OUI**	**OUI**

	Partenaire n°4	**Partenaire n°5**	**Partenaire n°6**
Nom du responsable scientifique /coordinateur :	**OUI**	**OUI**	**OUI**
Dénomination du partenaire (si NON, celle-ci sera remplacée, dans le texte publié, par la mention générique « Entreprise » ou « Organisme de recherche »).	**OUI**	**OUI**	**OUI**

	Partenaire n°7	**Partenaire n°8**	**Partenaire n°9**
Nom du responsable scientifique /coordinateur :	**OUI**	**OUI**	**OUI**
Dénomination du partenaire (si NON, celle-ci sera remplacée, dans le texte publié, par la mention générique « Entreprise » ou « Organisme de recherche »).	**OUI**	**OUI**	**OUI**

Partenaire 1 = Coordinateur du Projet

	Nom *	Prénom *	Emploi actuel	Discipline	% de temps de recherche consacré au projet	Rôle/Responsabilité dans le projet 4 lignes max
exemple	*MARTIN*	*Charlotte*	*Professeur*		*30%*	
Coordinateur	GUILLOT	Céline	MCF	Sciences du langage	50%	Coordination de l'ensemble du projet, experte de syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus), experte de philologie romane (volet description des textes)
Membres de l'équipe						
	CARLIER	Anne	MCF		20%	Experte de syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus)
	LAVRENTIEV	Alexis	ATER		25%	Expert pour l'encodage et l'enrichissement linguistique (volet description des textes : outils, méthodes et analyse), expert de syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus)
	LUSIGNAN	Serge	Professeur		10%	Expert de socio-linguistique (volet description des textes et exploitation du corpus)
	MARNETTE	Sophie	MCF		15%	Experte de la syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus), étude pragmatique des textes (volet description des textes)
	MÖHREN	Frankwalt	Professeur		15%	Expert de philologie romane (volet description des textes) et de lexicologie (volet exploitation du corpus)
	PIGNATELLI	Cinzia	MCF		15%	Experte de philologie romane (volet description des textes et exploitation du corpus)
	PRÉVOST	Sophie	CR1		10%	Experte de la syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus)
	SCHØSLER	Lene	Professeur		30%	Experte de la syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus) et de philologie romane (volet description des textes)
	STEIN	Achim	Professeur		15%	Experte de la syntaxe et sémantique médiévales (volet enrichissement linguistique et exploitation du corpus)
	TITTEL	Sabine	Ingénieur		15%	Experte de philologie romane (volet description des textes et

						exploitation du corpus)
	TROTTER	David	Professeur		15%	Expert de philologie romane (volet description des textes) et de lexicologie (volet exploitation du corpus)

Achim Stein

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Stein
Prénom : Achim
Age : 45 ans
Situation professionnelle : Professeur à l'Université de Stuttgart (Institut für Romanistik, doyen de la Faculté de Sciences humaines)
Cursus :
1992 : Doctorat de Philologie à l'Université de Stuttgart (*Nominalgruppen in Patentschriften*)
2000 : Habilitation à l'University de Cologne (*Semantik und Syntax italienischer Verben. Lexikalische Beschreibung mit Konzepthierarchien*)

B/ Autres expériences professionnelles
depuis 1999: assistant à l'Université de Cologne
depuis 1994-2000 : assistant à l'Université de Stuttgart
depuis 1989 : chercheur à l'Institut für Romanistik de l'Université de Stuttgart

C/ Publications
Publications les plus significatives
- Stein, A. (1998, réed. 2005): *Einführung in die französische Sprachwissenschaft*, 2. Aufl., Stuttgart, Weimar, Metzler.
- Stein, A. (2003), « Étiquetage morphologique et lemmatisation de textes d'ancien français », in : P. Kunstmann *et al*. éd., *Ancien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours*, 273-284.
- Stein, A. en collab. avec H. Schmid (1995), « Etiquetage morphologique de textes français avec un arbre de décisions », in : *Traitement automatique des langues*, 36/1-2, 23-35.
- Stein, A. (1993), *Nominalgruppen in Patentschriften. Komposita und prädikative Nominalisierungen im deutsch-französischen Vergleich*, Tübingen, Niemeyer.
- Stein, A. en collab. avec M. D. Gleßgen (2005), « Resources and Tools for Analyzing Old French Texts », in : J. Kabatek, C. Pusch, Claus & W. Raible éd., Tübingen, Narr, 135-145.

Publications les plus récentes
- Stein, A. (2005), *Semantische Repräsentation italienischer Verben. Automatische Disambiguierung mit Konzepthierarchien*, Tübingen, Niemeyer.
- Stein, A. (2003), « Zur Repräsentation von Selektionsrestriktionen », in : A. Blank & P. Koch éd., *Kognitive romanische Onomasiologie und Semasiologie*, Tübingen, Niemeyer, 173-188.
- Stein, A. (2003), « Wörterbücher und Textkorpora für Französisch und Italienisch », in : W. Dahmen *et. al*. éd., *Romanistik und neue Medien. Romanistisches Kolloquium XVI*, Tübingen, Narr, 107-124.
- Stein, A. (2003), « Lexikalische Kookkurrenz im afrikanischen Französisch », in : *Zeitschrift für französische Sprache und Literatur*, 113, 1-17.
- Stein, A. (2002), « Valence sémantique et définition lexicographique », in : *Syntaxe & Sémantique*, 4, 161-178.

Alexei Lavrentiev

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Lavrentiev
Prénom : Alexei
Age : 34 ans
Situation professionnelle : ATER à l'ENS LSH (Section sciences du langage)
Cursus :
1999 : Candidat ès lettres (équivalent du doctorat du 3e cycle), thèse *Analyse typologique de la catégorie des cas en russe* soutenue auprès du Jury spécialisé à l'Université de Tomsk (Russie)
1995 : Diplôme d'études approfondies en sciences du langage, Université Paris III – Sorbonne Nouvelle
1993 : Diplôme d'études supérieures en langue et littérature russe, Université d'état de Novossibirsk

B/ Autres expériences professionnelles
1998 – 2004 : Chercheur dans le secteur de la langue russe en Sibérie, à l'Institut de philologie de la branche sibérienne de l'Académie des Sciences de Russie ;
2002 – 2003 : Maître de conférences au département de linguistique générale et russe de l'Université d'Etat de Novossibirsk (Akademgorodok), cours enseigné : morphologie du russe ;
2001 – 2002 : Fulbright visiting fellow, département de français et d'italien de l'Université de Princeton (Etats-Unis), participation au Projet *Charrette* (édition hypertextuelle de la tradition manuscrite du roman *Chevalier de la Charrette* de Chrétien de Troyes)

C/ Publications
Publications les plus significatives
- Lavrentiev, A. (sous presse), « Typologie textuelle pour l'étude linguistique de manuscrits français médiévaux », in : A. Lavrentiev, C. Marchello-Nizia et J.-P. Perrot éd., *Systèmes graphiques de manuscrits médiévaux et incunables français : ponctuation, segmentation, graphies.* (Actes de la Journée d'étude de Lyon, ENS LSH, 6 juin 2005), Chambéry, Presses de l'Université de Savoie.
- Lavrentiev, A. en collab. avec S. Heiden (2004), « Ressources électroniques pour l'étude des textes médiévaux : approches et outils », in : *Revue française de la linguistique appliquée*, 1, 99 – 118.
- Lavrentiev, A. (2003), « Analyse formalisée des conditions syntaxiques de l'emploi des signes de ponctuation dans les manuscrits français médiévaux », in : *Goumanitarnye naouki v Sibiri*, 4, 76 – 81 (en russe).
- Lavrentiev, A. (2001), *Catégorie des cas et typologie linguistique*, Novossibirsk, Izdatel'stvo NGU, 216 p. (en russe)
- Lavrentiev, A. (2000), « A propos de la ponctuation dans l'*Image du monde* », in : *La Licorne*, 52, 23 – 37.

Publications les plus récentes
- Lavrentiev, A. en collab. avec C. Marchello-Nizia (à par.), « Le développement du déterminant démonstratif CE dans les manuscrits du roman *Lancelot ou Le Chevalier à la Charrette* de Chrétien de Troyes », E. Thorington, , G. Greco & D. Long éd., *Medieval Philologies : Tradition and Innovation in* Charrette *Studies*.
- Lavrentiev, A. (à par.), « Base de français médiéval et transcriptions de manuscrits : recherche de complémentarité » in : *Actes du XXIV CILPR*.
- Lavrentiev, A (à par.), « Pour une méthodologie d'étude de la ponctuation médiévale basée sur une approche typologique », in : *Verbum*.
- Lavrentiev, A (2004), « Linguistique de corpus : idéologie, méthodologies, technologies », in : *Sibirskij Filologičeskij Žurnal*, 3-4, 121-134 (en russe).
- Lavrentiev, A (2005), « Représentation de transcriptions diplomatiques de manuscrits français médiévaux en XML-TEI », in : J. Kabatek, C. Pusch et W. Raible éd., *Romanistische Korpuslingustik II: Korpora und diachrone Sprachwissenschaf, Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*, Tübingen, Gunter Narr Verlag, (ScriptOralia ; 130), 109 – 121.

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Carlier
Prénom : Anne
Age : 46 ans
Situation actuelle : Maître de Conférences à l'Université de Valenciennes et du Hainaut Cambrésis
Cursus (depuis 1993)
2006 : Habilitation à diriger des recherches (Université de Lille III, sous la dir. de D. Van de Velde)
1992 : Doctorat de Philosophie et Lettres (Université de Leuven, sous la dir. de L. Melis)
1984 : DEUG de Sciences économiques (Université de Leuven)
1984 : Agrégation de l'enseignement secondaire supérieure (Université de Leuven)
1983 : Licence – Maîtrise de Philologie romane (Université de Leuven (Belgique))

B/ Autres expériences professionnelles
1991-1993 : Assistante à l'Université d'Anvers (Belgique)
1984-1990 : Assistante aux Facultés universitaires de Namur (Belgique)

C/ Publications
Publications les plus significatives
- Carlier, A. en collab. avec M. Goyens (1998), « De l'ancien français au français moderne : régression du degré zéro de la détermination et restructuration du système des articles », in : *Cahiers de l'Institut de Linguistique de Louvain-la-Neuve,* 24, 3-4 , 77-112.
- Carlier, A. (2001), « La genèse de l'article *un* », in : *Langue française,* 130, 65-88.
- Carlier, A. (2004), « Sur les premiers stades de développement de l'article partitif », in : *Scolia,* 18, 117-147.
- Carlier, A. en collab. avec L. Melis (2006), « L'article partitif et les expressions quantifiantes contiennent-ils le même *de* ? », in : G. Kleiber, C. Schnedecker et A. Theissen éd., *La relation partie-tout,* Louvain, Peeters, 449-464.
- Carlier, A. (2007), « From Preposition to Article : The Grammaticalization of the French Partitive », in : *Studies in Language,* 31, 1-49.

Publications les plus récentes
- Carlier, A. (2000), « Les articles *du* et *des* en synchronie et en diachronie : une analyse de leur résistance à l'interprétation générique », in : *Revue romane*, 35/2, 177-206.
- Carlier, A. (2002), « Les propriétés aspectuelles du passif », in : *Cahiers Chronos,* 10, 41-63.
- Carlier, A. (2005), « L'argument davidsonien : un critère de distinction entre les prédicats 'stage level' et les prédicats 'individual level' ? », in : *Travaux de linguistique,* 50, 13-35.
- Carlier, A. (2004-2005), « 'Ce sont des Anglais' : un accord avec l'attribut ? », in : *L'Information grammaticale,* 103, 13-18 et 104 , 4-14.
- Carlier, A. en collab. avec W. De Mulder (2006), « Du démonstratif à l'article défini : le cas de *ce* en français moderne », in : *Langue française,* 152, 96-113.

Céline Guillot

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Guillot
Prénom : Céline
Age : 37 ans
Situation actuelle : Maître de Conférences à l'ENS-LSH
Cursus
2003 : Thèse de doctorat en Sciences du Langage (« Le rôle du démonstratif dans la cohésion textuelle au XV^{ème} siècle. Eléments de grammaire textuelle »)
1994 : Diplôme de Conservateur du Patrimoine (Ecole nationale du Patrimoine)
1992 : Diplôme d'Archiviste-paléographe (Ecole nationale des chartes, Thèse de l'ENC)
1992 : DEA de Sciences du Langage à l'EHESS (Direction : O. Ducrot, J.-Cl. Anscombre)

B/ Autres expériences professionnelles
1999- août 2004 : ATER à l'ENS-LSH (détachement de mon corps d'origine)
1998-1999 : Conseillère pour les archives à la Direction Régionale des Affaires Culturelles de Rhône-Alpes
1994-1998 : Conservateur adjoint aux Archives Départementales de la Loire

C/ Publications
Publications les plus significatives
- Guillot, C. (2006), « Démonstratif et déixis discursive : analyse comparée d'un corpus écrit de français médiéval et d'un corpus oral de français contemporain », in : *Langue française,* 152, 56-69.
- Guillot, C. (2004), « *Ceste parole* et *ceste aventure* dans la *Queste del Saint Graal,* marques de structuration discursive et transitions narratives », in : *L'Information grammaticale*, 103, 29-36.
- Guillot, C. (2003), « Grammaticalisation et système de la référence : *celui, icelui, cest, cestui* et *ledict* dans un texte du début du XV^{ème} siècle », in : B. Combettes, C. Marchello-Nizia, S. Prévost éd., *La grammaticalisation en français*, (Actes du colloque DIACHRO-I organisé à Paris en janvier 2002, *Grammaticalisations en français*), numéro spécial de *Verbum,* XXV/3, 369-379.
- Guillot, C. en collab. avec S. Heiden (2003), « Capitalisation des savoirs par le Web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de français médiéval » in : P. Kunstmann, F. Martineau, D. Forget éd., *Ancien et moyen français sur le Web : enjeux méthodologiques et analyse de discours* (Actes du colloque organisé à Ottawa en octobre 2002), 77-92.
- Guillot C. en collab. avec S. Heiden et A. Lavrentiev (à par.), « Typologie des textes et des phénomènes linguistiques pour l'analyse du changement linguistique avec la Base de Français Médiéval », à paraître dans les actes du colloque international *Corpus et questionnements du littéraire*, (Université de Paris X, novembre 2005).

Publications les plus récentes
- Guillot, C. (dir.) (2006). *Le démonstratif en français, Langue française*, 152.
- Guillot, C. (2006), « Anaphores résomptives démonstratives et relations partie/tout en discours », in : G. Kleiber, - C. Schnedecker, A. Theissen éd., *La relation partie-tout*, Louvain-Paris, Peeters (Bibliothèqe de l'Information Grammaticale), 289-302.
- Guillot, C., Heiden, S. et Prévost, S. (dir.) (2006). *A la quête du sens : études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia*, Lyon, ENS Editions.
- Guillot, C. en collab. avec C. Marchello-Nizia et A. Lavrentiev, (à par.), « La Base de Français Médiéval (BFM) : états et perspectives », à paraître dans les actes du colloque *Le nouveau corpus d'Amsterdam* (Lauterbad, Allemagne, 23-26 février 2006).
- Guillot, C. (à par. 2007), « Entre anaphore et deixis : l'anaphore démonstrative à fonction résomptive », à paraître aux éditions Niemeyer dans les Actes du XXIV^{ème} Congrès de Linguistique et de Philologie Romanes organisé à Aberystwyth (août 2004).

Cinzia Pignatelli

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Pignatelli
Prénom : Cinzia
Age : 46 ans
Situation actuelle : Maître de conférences en Philologie Romane et Linguistique Française à la Faculté des Lettres et Langues de l'Université de Poitiers (depuis 1998)
Cursus
1997 : Doctorat ès Lettres de l'Université des Sciences Humaines de Strasbourg (Edition critique de la traduction des Otia imperialia de Gervais de Tilbury par Jean d'Antioche dans le ms. BNF f. fr. 9113)

B/ Autres expériences professionnelles
1997-1998 : Chargée de cours de Linguistique et Histoire de la Langue Italienne à l'Université des Sciences Humaines de Strasbourg
1997-2000 : Correction de copies d'Ancien Français pour la préparation au CAPES externe de Lettres Modernes pour le CNED
1996-1997 : ATER de Linguistique Diachronique du Français à l'Université des Sciences Humaines de Strasbourg
1988-1992 : Chargée de cours de Linguistique Diachronique du Français à l'Université des Sciences Humaines de Strasbourg
1985-1996 : Enseignement de l'italien langue étrangère
1980-1998 : Interprétation et traduction français-italien/italien-français

C/ Publications
Publications les plus significatives
- Pignatelli, C. (2006), « Italianismes, provençalismes et autres régionalismes chez Jean d'Antioche traducteur des Otia imperialia », in : C. Galderisi et J. Maurice éd., « Qui tant savoit d'engin et d'art ». Mélanges de philologie médiévale offerts à Gabriel Bianciotto, Poitiers, Université de Poitiers/Centre d'Etudes Supérieures de Civilisation Médiévale avec la collaboration du Centre d'étude et de recherche « Éditer / Interpréter » de l'Université de Rouen, 367-377.
- Pignatelli, C. en collab. avec D. Gerner (2006), Les traductions françaises des Otia imperialia de Gervais de Tilbury par Jean d'Antioche et Jean de Vignay : édition de la troisième partie, Genève, Droz.
- Pignatelli, C. (2002), « La Lemmatisation du texte du Chevalier de la Charrette (Lancelot) ou le nécessaire retour aux manuscrits », in : C. Pignatelli et M. Robinson éd., Chrétien de Troyes, Le Chevalier de la Charrette (Lancelot). Le "Projet Charrette" et le renouvellement de la critique philologique des textes, Tübingen, Gunter Narr, Oeuvres et critiques XXVII/1, 52-69.
- Pignatelli, C. (2001), « Les glossaires bilingues médiévaux : entre tradition latine et développement du vulgaire », in : Revue de Linguistique Romane, 65, 75-111.
- Pignatelli, C (à par.), « Le moyen français dans les traductions de Jean d'Antioche », à paraître dans C. Galderisi et C. Pignatelli éd., *La traduction vers le moyen français*. Actes du 2e Colloque de Moyen Français (Poitiers 27-29 avril 2006), Turnhout, Brepols.

Publications les plus récentes
- Pignatelli, C. (2006), « Une approche de la tradition textuelle du Chevalier de la Charrette : la quantification des phénomènes régionaux", in : C. Arrignon et al. éd., Cinquante années d'études médiévales. A la confluence de nos disciplines. Actes du Colloque organisé à l'occasion du Cinquantenaire du CESCM (Poitiers 1er-4 septembre 2003), Turnhout, Brepols, 741-752.
- Pignatelli, C. (2005), « Le « Projet Charrette » à Poitiers : un état des lieux », in : Cahiers de Civilisation Médiévale, 48, 227-232.
- Pignatelli, C. (2004), « Un traducteur qui affiche ses croyances : l'ajout d'exempla au corpus des Otia imperialia de Gervais de Tilbury dans la traduction attribuée à Jean d'Antioche », in : M. Colombo et C. Galderisi éd. ,« Pour acquérir honneur et pris ». Mélanges de Moyen Français offerts à Giuseppe Di Stefano, Montréal, Ed. CERES, 47-58.
- Pignatelli, C. (2004), « I Vocabula Magistri Gori de Aretio della British Library», in : Per Alberto Nocentini. Ricerche linguistiche, Firenze, Alinea, 189-218.
- Pignatelli, C. (à par.), « *Comme firent les Scipions a Romme* : figures d'analogie dans le *Quadrilogue invectif* d'Alain Chartier », à paraître dans les actes du XIIe colloque international sur le Moyen Français, *Le langage figuré (*Montréal, McGill Univ., 4-6 octobre 2004).

Lene Schøsler

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Schøsler
Prénom : Lene
Age : 61 ans
Situation actuelle : Professeure à l'Université de Copenhague
Cursus
1984 : thèse de doctorat (« La déclinaison bicasuelle de l'ancien français »)
1989 – 1990 : chercheure attachée à Eurotra-DK, responsable du dictionnaire électronique danois – italien
1997 : Professeure des langues romanes, Institut d'Etudes Romanes, Université de Copenhague.
1999 -2002 : Directrice de l'Institut d'Etudes Romanes, Université de Copenhague
2000-2002 : Directrice d'un projet de recherche sur le français parlé.
2001- 2006 : Directrice de plusieurs projets de recherche sur la grammaticalisation.

B / Autres expériences professionnelles
Conférencière et professeure invitée dans un grand nombre d'universités et d'institutions étrangères : Allemagne, Angleterre, Belgique, Canada, Estonie, Etats-Unis, Finlande, France, Islande, Italie, Norvègue...

C/ Publications
Publications les plus significatives
- Schøsler, L. (2006), « Vivre ça me fout la trouille. Mourir plus encore, mais vivre ça me fait vraiment peur. A diachronic perspective on support verb constructions », in : Eksell, Kerstin & Thora Vinther éd., *Change in Verbal Systems. Issues on Explanation*, Frankfurt a/M, Peter Lang, 177-198.
- Schøsler, L. (2004), « Historical corpora. Problems and methods », in : A. Bozzi, L. Cignon et, J.-L. Lebrave éd., *Linguistica computazionale XX-XXI, Digital technology and philological disciplines*, Pisa, Roma, Istituti editoriale e poligrafici internazionali, 455-472.
- Schøsler, L. (2001), « From Latin to modern French: Actualization and markedness », in : H. Andersen éd., *Actualization. Linguistic Change in Progress. Papers from a workshop held at the 14th Int. Conf. on Historical Linguistics. Vancouver, B.C., 14 Aug. 1999*, (Current issues in linguistic theory; 219), 169-185.
- Schøsler, L. en collab. avec P. van Reenen et S. C. Herring (2000), « On textual parameters in older languages - theoretical overview », in : S.C . Susan C Herring, Pieter van Reenen & Lene Schøsler éd., *Textual parameters in older languages*, Amsterdam, John Benjamins Publ. Co., 1-31.
- Schøsler, L. (1984), *La déclinaison bicasuelle de l'ancien français, son rôle dans la syntaxe de la phrase, les causes de sa disparition,* Odense, Etudes romanes de l'Université d'Odense, 19, 321 p., Thèse de doctorat.

Publications les plus récentes
- Schøsler, L. (2006), « Grammaticalisation et dégrammaticalisation. Etude des constructions progressives en français du type Pierre va / vient / est chantant », in : E. Labeau, C. Vetters & P. Caudal éd., *Sémantique et diachronie du système verbal français,* Cahiers Chronos, 16, 91-119.
- Schøsler, L. (2004), « Scribal variations: When are they genealogically relevant - and when are they to be considered as instances of 'mouvance? », in : P. van Reenen *et al*. éd*., Studies in Stemmatology*, 2, 207-226.
- Schøsler, L. (2003), « Les verbes supports dans une perspective diachronique. Le cas de *garde*, noyau prédicatif », in : P. Kunstmann, F. Martineau et D. Forget éd., *Ancien et moyen français sur le web. Enjeux méthodologiques et analyse du discours.* (Colloque tenu en oct. 2002 à l'Úniversité d'Ottawa), 221-271.
- Schøsler, L. (2003), « Le rôle de la valence pour une classification sémantique des verbes », in : P. Blumenthal et J.-E. Tyvaert éd., *La cognition dans le temps. Etudes cognitives dans le champ historique des langues et des textes,* 145-160.
- Schøsler, L. (2002), « Je le pince au nez - je lui pince le nez - je pince son nez - Jean lève la main. La possession inaliénable : perspectives synchroniques et diachroniques », in : D. Lagorgette et P. Larrivée éd., *Représentations du sens linguistique* (*LINCOM Studies in theoretical linguistics;* 22), 331-348.

D/ Prix et distinctions
2001 : Décoration des "Palmes académiques" ; 2001 : Prix linguistique de "Einar Hansens Forskningsfond" ; 2002 : Décoration de "l'ordre du mérite" ; 2005 : Membre de l'Académie Royale des Sciences et Lettres du Danemark ; 2007 : By-Fellow à Churchill College, Cambridge.

Serge Lusignan

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Lusignan
Prénom : Serge
Age : 63 ans
Situation actuelle : Professeur d'histoire du Moyen Âge à l'Université de Montréal
Cursus
S. Lusignan est né le 22 octobre 1943. Il a terminé son doctorat à l'Université de Montréal en 1971. Le titre de la thèse est : *La logique dans le* Speculum doctrinale *livre III de Vincent de Beauvais*. Il est au service de l'Université de Montréal comme enseignant depuis 1972. Il est maintenant professeur titulaire d'histoire du Moyen Âge au département d'Histoire de cette université. Il a été : maître de conférence associé à l'Université de Paris XII en 1990-1991 et professeur associé à la même université en 1991-1992 ; professeur invité à l'Université de Paris XIII à l'automne 1998 ; Mundus Visiting Fellow à University of Saint Andrews et à l'Université de Perpignan d'octobre à décembre 2005. Il est également chercheur associé UMR 8589, Laboratoire de Médiévistique Occidentale de Paris, Université de Paris I et CNRS.

C/ Publications
Publications les plus significatives
- Lusignan, S. (2004), *La langue des rois au Moyen Âge. Le français en France et en Angleterre*, Paris, Presses Universitaires de France, 296 p.
- Lusignan, S. (1986, 2e éd. 1987), *Parler vulgairement. Les intellectuels et la langue française aux XIIIe et XIVe siècles*, Paris-Montréal, Vrin-Presses de l'Université de Montréal, 204 p.
- Lusignan, S. en collab. avec S. Brazeau (2004), « Jalon pour une histoire de l'orthographe française au XIVe siècle : l'usage des consonnes quiescentes à la chancellerie royale », in : *Romania*, 122, 444-467.
- Lusignan, S. (1999), « Langue française et société du XIIIe au XVe siècle », in : J. Chaurand éd, *Nouvelle histoire de la langue française*, Paris, Seuil, 91-143.
- Lusignan, S. (1999), « L'usage du latin et du français à la chancellerie de Philippe VI », in : *Bibliothèque de l'École des Chartes*, CLVII, 509-521.

Publications les plus récentes
- Lusignan, S. en collab. avec O. Guyotjeannin et avec le concours des étudiants de l'École nationale des chartes et la collaboration d'E. Frunzeanu (2005), *Le formulaire d'Odart Morchesne dans la version du ms BnF fr. 5024*, in : *Mémoires et documents de l'École des chartes*, 80, Paris, École des chartes, 480 p.
- Lusignan, S. (1999), *« Vérité garde le roy. » La construction d'une identité universitaire en France (XIIIe -XVe siècle)*, Paris, Publications de la Sorbonne, 332 p.
- Lusignan, S. (2005), « La résistible ascension du vulgaire : persistance du latin et latinisation du français dans les chancelleries de France et d'Angleterre à la fin du Moyen Âge », in : *Mélanges de l'École française de Rome, Moyen Âge*, 117/2, 471-508.
- Lusignan, S. (2004), « *François est profitable et latin est prejudiciable* : l'enjeu linguistique d'un conflit entre le village de Saint-Albain et le chapitre de Mâcon », in : *Retour aux sources. Textes, études et documents d'histoire médiévale offerts à Michel Parisse,* 795-801.
- Lusignan, S. en collab. avec A.-I Tardif (1999),. « Des druides aux clercs : quelques lectures françaises de Jules César aux XIIIe et XIVe siècles », in : *Revue Historique*, CCCI/3, 435-462.

D/ Prix et distinctions
Serge Lusignan est membre de la Société Royale du Canada depuis 1989. En 2002-2004, il a bénéficié de la bourse de recherche Killam, la bourse la plus prestigieuse dans le domaine de la recherche universitaire au Canada.

Sophie Marnette

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Marnette
Prénom : Sophie
Age : 37 ans
Situation actuelle : Maître de conférences en français médiéval, Université d'Oxford, Royaume-Uni et Directrice d'étude (tutorial fellow) en langues modernes, Balliol College, Royaume-Uni
Cursus
1996 : Doctorat ès lettres (linguistique française), Université de Californie, Berkeley, USA
1991 : Licence en Philologie Romane, Université Libre de Bruxelles, Belgique, Grande distinction
1989 : Candidature en Philologie Romane, Université Libre de Bruxelles, Belgique, Grande distinction

B/ Autres expériences professionnelles
2001-2003 : Harvard University, USA; Chercheuse invitée
2001-2003 : Université d'Oxford, Royaume-Uni; Zaharoff Teaching Fellow, chargée de cours en linguistique française
1997-2001 : Université de Cambridge, Royaume-Uni; chargée de recherche, Gonville & Caius College
1999-2000 : Université Libre de Bruxelles, Belgique; suppléance du prof. B. Cerquiglini
1996-1997 : Université de St. Andrews, Royaume-Uni; chargée de cours et de recherche
1996 : University de Californie à Los Angeles, USA; chargée de cours invitée

C/ Publications
Publications les plus significatives
- Marnette, S. (à par.), « La Ponctuation du discours rapporté dans quelques manuscrits de romans en prose médiévaux », in : *Verbum*.
- Marnette, S. (2006), « La Signalisation du discours rapporté en français médiéval », in : *Langue française*, 149, 31-47.
- Marnette, S. (2006), « Experiencing Self and Narrating Self in Medieval French Chronicles », in : V. Greene éd., *The Medieval Author in Medieval French Literature. Palsgrave-Mc Millan, Studies in Arthurian and Courtly Cultures*, 115-34.
- Marnette, S. (2005), *Speech and Thought Presentation in French Concept and Strategies*, Amsterdam – New York, John Benjamins.
- Marnette, S. (1998), *Narrateur et points de vue dans la littérature française médiévale : Une approche linguistique*, Bern, Peter Lang.
- Marnette, S. (1996), « Réflexions sur le discours indirect libre en français médiéval », in : Romania, 114, 1-49.

Publications les plus récentes
- Marnette, S. (2006), « Je vous dis que l'autocitation c'est du discours rapporté », in : J. M. Lopez-Muñoz, S. Marnette & L. Rosier éd., *L'Autocitation*, *Travaux de linguistique,* 52, 25-40.
- Marnette, S. (2002/2003), « Sources du récit et discours rapportés : L'art de la représentation dans les chroniques et les romans français des 14^{e} et 15^{e} siècles », in : *Le Moyen Français,* 51-52-53, 435-459.
- Marnette, S. (2002), « The Evolution of Reported Discourse in Medieval French: An Overview », in : W. Bennett & R. Sampson éd., *Interpreting the History of French. A Feitschrift for Peter Rickard on the occasion of his eightieth birthday,* Amsterdam-New York, Rodopi, 3-34.
- Marnette, S. (2002), *« Je dis que ... Je pense que …* Le *je* narrateur, auteur, témoin et personnage des chroniques », in : *LYNX,* 32, 271-84.
- Marnette, S. (1999), *«* Il le vos mande, ge sui qui le vos di : Les stratégies du dire dans les chansons de geste », in : *La Revue de linguistique romane,* 63/251-252, 387-417.

Frankwalt Möhren

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Möhren
Prénom : Frankwalt
Age : 64 ans
Situation actuelle : Außerplanmäßiger Professor de la Faculté des Lettres de l'Université de Heidelberg (depuis 1995) et Chargé de recherches auprès de l'Académie des sciences de Heidelberg, directeur et rédacteur du Dictionnaire étymologique de l'ancien français (depuis le 1er octobre 1985)
Cursus
1983 : Habilitation à l'Université de Heidelberg; Venia legendi pour 'Philologie romane: linguistique' (thèse publiée : *Wort- und sachgeschichtliche Untersuchungen an französischen landwirtschaftlichen Texten, 13., 14. und 18. Jahrhundert (Seneschaucie, Menagier, Encyclopédie))*
1975 : Promotion au doctorat à l'Université Laval, Québec (thèse publiée : *Le renforcement affectif de la négation par l'expression d'une valeur minimale en ancien français,* mention : summa cum laude)
1963-1968 : Études de Philologie romane (Langue et Littérature) et d'Anglistique (Langue et Littérature) à l'Université de Heidelberg

B/ Autres expériences professionnelles
1989-1990 : Professeur titulaire invité à la Heinrich-Heine-Universität de Düsseldorf
1983-1994 : Privat-Dozent de la Faculté des Lettres de l'Université de Heidelberg
1975-1985 : Wissenschaftlicher Assistent au Romanisches Seminar de l'Université de Heidelberg ; Séminaires et Travaux pratiques en Philologie française et italienne (Linguistique)
1968-1975 : Enseignement de Philologie française et de Phonétique (Professeur adjoint) à l'Université Laval, Québec
1968-1975 : Directeur de bureau et rédacteur du *Dictionnaire étymologique de l'ancien français* à l'Université Laval, Québec
1964 -1967 : Enseignant de Grammaire espagnole à l'École de Langues et d'Interprètes du Englisches Institut, Heidelberg

C/ Publications
Publications les plus significatives
- Möhren, F. (2006), « *L'importance de la critique des sources en étymologie »*, in : E. Buchi éd., *Actes du Séminaire de méthodologie en étymologie et histoire du lexique*, Nancy (ATILF), 7 avril 2006, en ligne.
- Möhren, F. (2005), « *Le DEAF - Base d'un atlas linguistique de l'ancien français ? »*, in : M.-D. Gleßgen et A. Thibault éd., *La lexicographie différencielle du français et le* Dictionnaire des régionalismes de France, Strasbourg, Presses Universitaires de Strasbourg, 99-113.
- Möhren, F. (2004), « Seme und Sachen », in : F. Lebsanft et M.-D. Gleßgen éd., Historische Semantik in den *romanischen Sprachen*, Tübingen, Niemeyer (Ling. Arb. 483), 71-77.
- Möhren, F. (2003), « Le Godefroy, une source encore valable au XXIe siècle ? », in : F. Duval éd., Frédéric Godefroy. Actes du Xe Colloque international sur le moyen français, Paris, Ecole nat. des chartes (Mém. et Doc. 71), 279-294.
- Möhren, F. (2000), « Onefold lexicography for a manifold problem? », in : D.A. Trotter éd., Multilinguism in later medieval Britain, Woodbridge, Boydell & Brewer, 157-168.

Publications les plus récentes
- Möhren, F. (2006), « *Les débuts de l'écriture française de la géométrie au XIIIe siècle »*, in *: C. Thomasset, L'écriture du texte scientifique : des origines de la langue français au XVIIIe siècle, Paris, Presses Universitaires de l'Université Paris-Sorbonne,* 93-116.
- Möhren, F. (2005), « Englisch standard. Ein Beispiel französisch-englischer Wort- und Sachgeschichte », in : W. Dahmen et al., Englisch und Romanisch, Tübingen, Narr, 53-75.
- Möhren, F. (2003), « Bilanz und Perspektiven », in Th. Städtler éd., Wissenschaftliche Lexikographie im deutschsprachigen Raum, Heidelberg, Winter, 33-47.
- Möhren, F. (2000), « *Guai victis! Le problème du* gu *initial roman »*, in : Medioevo Romanzo, 24 5-81.
- Möhren, F. (1999), « *Kreuzzugsvokabular: exotisches Dekorum oder kulturelle Übernahme? »*, in : M. Bierbach et B. von Gemmingen éd., *Kulturelle und sprachliche Entlehnung: Die Assimilierung des Fremden*, Bonn, Rom. Verl. (Abh. zu Spr. u. Lit. 123), 104-118.

D/ Prix et distinctions
1983 : Prix Albert Dauzat par la Société de Linguistique romane

Sophie Prévost

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Prévost
Prénom : Sophie
Age : 39 ans
Situation professionnelle : *Chargée de Recherche 1ère classe (CR1),* UMR 8094 « Langues, Textes, Traitements Informatiques et Cognition » (LATTICE) (CNRS / ENS Paris)
Cursus :
1997 : *Thèse de doctorat (*« Les énoncés à sujet postverbal en français aux 15ème et 16ème siècles : une analyse sémantico-pragmatique »)
1993 : *D.E.A. de Linguistique Théorique et Formelle*
1991 : *Agrégation de Lettres Modernes*

B/ Autres expériences professionnelles
2006-2007 : Chargée de cours en linguistique en 2ème année de Licence à Paris-10 et en 2ème année de Master à Paris-7
1998-2001 : Chargée de Recherche, UMR 8503 « Analyses de corpus linguistiques » (CNRS / ENS-LSH Lyon)
1994-1998 : Chargée de cours/ATER en linguistique : Universités Paris-3, Paris-7, Caen
1991-1994 : Professeure de français en lycée
Co-organisation des colloques internationaux DIACHRO-1 et DIACHRO-2 (diachronie du français) en janvier 2002 et janvier 2004 et organisation du colloque DIACHRO-3 en septembre 2006

C/ Publications
Publications les plus significatives
- Prévost, S. (sous presse) « Grammaticalisation, lexicalisation et dégrammaticalisation : des relations complexes », in : *Cahiers de Praxématique*, 46.
- Prévost, S. (2003), « Les compléments spatiaux : du topique au focus en passant par les cadres », in : *Travaux de Linguistique,* 47, 51-78.
- Prévost, S. (2003), « La grammaticalisation : unidirectionnalité et statut », in : *Le Français Moderne*, LXXI (2), 144-166
- Prévost, S. (2003), « Détachement et topicalisation : des niveaux d'analyse différents », in : *Cahiers de Praxématique*, 40, 97-126.
- Prévost, S. (2001), *La postposition du sujet en français aux 15ème et 16ème siècles : une approche sémantico-pragmatique*, Paris, Editions du CNRS.

Publications les plus récentes
- Prévost, S. (à par.), « A (ce) propos de X / à ce propos / à propos : évolution du 14ème au 16ème siècle », in : *Langue Française.*
- Prévost, S. (à par.), « Encadrement temporel en français médiéval », in : *Revue Québéquoise de Linguistique.*
- Prévost, S. (2005), « Exploitation d'un corpus de français médiéval : enjeux, spécificités et apports », in A. Condamines éd., *Sémantique et corpus,* Paris, Hermès/Lavoisier (Série « Traité IC2 » ; Cognition et traitement de l'information), 147-176.
- Prévost, S. (2002), « Evolution de la syntaxe du pronom personnel sujet depuis le français médiéval : la disparition d'alternances signifiantes », in : D. Lagorgette et P. Larrivée éd., *Représentations du sens linguistique*, Munich, Lincom, Studies in Theoretical Linguistics, 22, 309-329.
- Prévost, S. en collab. avec S. Heiden (2002), « Etiquetage d'un corpus de français médiéval : enjeux et modalités », in : C.D. Pusch et W. Raible éd., *Romance Corpus Linguistics - Corpora and Spoken Language*, Tübingen, Gunter Narr Verlag, 127-136.

Sabine Tittel

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Tittel
Prénom : Sabine
Age : 35 ans
Situation actuelle : Rédactrice du *Dictionnaire étymologique de l'ancien français* à l'Académie des Sciences de Heidelberg (depuis le 1er octobre 1997)
Cursus
1991-1997 : Études de Philologie romane (Langue et Littérature) et d'Ethnologie à l'Université de Heidelberg
2003: Promotion au doctorat à l'Université de Heidelberg (thèse publiée : *Die* « *Anathomie* » *in der* « *Grande Chirurgie* » *des Gui de Chauliac. Wort- und sachgeschichtliche Untersuchungen und Edition,* mention : summa cum laude)

C/ Publications
Tittel, S. (2004), Die « Anathomie » in der « Grande Chirurgie » des Gui de Chauliac. Wort- und sachgeschichtliche Untersuchungen und Edition, Tübingen, Max Niemeyer Verlag (Beihefte zur Zeitschrift für Romanische Philologie 328).

David Trotter

A/ Nom, prénom, âge, situation actuelle, cursus
Nom : Trotter
Prénom : David Andrew
Age : 59 ans
Situation actuelle : Professeur titulaire de français et directeur du Département de Langues Européennes, University of Wales Aberystwyth (depuis octobre 1993); directeur, *Anglo-Norman Dictionary* (www.anglo-norman.net)
Cursus
1979-1985 : thèse de doctorat (D.Phil.) à Oxford, sur la littérature des croisades en ancien français

B/ Autres expériences professionnelles
1985-1993 : Lecturer in Medieval French Language and Literature, University of Exeter
1983-1985 : Laming Junior Fellow, The Queen's College, Oxford
1978-1979 : Assistant d'anglais à Strasbourg, séjours de recherches à Paris (1980/1981), recherches à Nancy (ATILF, archives) en 2001

C/ Publications
Publications les plus significatives
- Trotter, D. en collab. avec W. Rothwell et S. Gregory (2005), *Anglo-Norman Dictionary: revised edition, A-C*; *D-E*, 2 vols, London, MHRA.
- Trotter, D. éd. (2000). *Multilingualism in Later Medieval Britain,* Cambridge, D.S. Brewer.
- Trotter, D. éd. (2005), *Albucasis, Traitier de Cyrurgie: Edition de la traduction en ancien français de la Chirurgie d'Abū'l Qāsim Halaf Ibn 'Abbās al-Zahrāwī du manuscrit BNF, français 1318*, in : *Zeitschrift für romanische Philologie, Beiheft* 325, Tübingen, Niemeyer.
- Trotter, D. (1997), « *Mossenhor, fet metre aquesta letra en bon francés*: Anglo-French in Gascony , in Stewart Gregory & D.A. Trotter (éds), *De mot en mot: Essays in honour of William Rothwell* (Cardiff: MHRA/University of Wales Press, 1997), 199-222.
- Trotter, D. (1996/1997 [1998]). 'Les néologismes de l'anglo-français et le *FEW'*, *Le Moyen Français*, 39-41 (1996/1997 [1998]), 577-635.

Publications les plus récentes
- Trotter, D. (2006). ""*Si le français n'y peut aller*: Villers-Cotterêts and mixed-language documents from the Pyrenees', in D.J. Cowling (ed.), *Conceptions of Europe in Renaissance France: a Festschrift for Keith Cameron* (Amsterdam: Rodopi, 2006), 77–97.
- Trotter, D. (2006). 'Language Contact, Multilingualism, and the Evidence Problem', in Schaefer, Ursula (ed.), *The Beginnings of Standardization: Language and Culture in Fourteenth-Century England* (Frankfurt: Peter Lang, 2006), 73–90.
- Trotter, D. (2006). 'Contacts linguistiques à l'intérieur de la Romania: Langues romanes et français / occitan', *Romanische Sprachgeschichte*: *Histoire linguistique de la Romania (HSK 23)*, ed. G. Ernst, C. Schmitt, M.-D. Gleßgen, W. Schweickard (Berlin: De Gruyter, 2006), II, 1776–1785.
- Trotter, D. (2005). *Boin sens et bone memoire*: tradition, innovation et variation dans un corpus de testaments de Saint-Dié-des-Vosges (XIII^{e} et XIV^{e} siecles)', in A. Schrott / H. Völker (eds), *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen* (Göttingen: Universitätsverlag Göttingen, 2005), 269–278.
- Trotter, D. (2005). 'Diastratische und Diaphasische Variation: Normierungstendenz und Unabhängigkeit in lothringischen Dokumenten des Mittelalters', in Gärtner, K. & Holtus, G., (eds.), *Überlieferungs- und Aneignungsprozesse im 13. und 14. Jahrhundert auf dem Gebiet der westmitteldeutschen und ostfranzösischen Urkunden- und Literatursprachen. Beiträge zum Dritten internationalen Urkundensprachen-Kolloquium vom 20.-22 Juni 2001 in Trier*, Trierer Historische Forschungen, 59 (Trier: Kliomedia, 2005), 245–322.

Acronyme ou titre court du projet :

A-2 : Autres partenaires du projet (remplir une fiche par partenaire)
Un responsable scientifique de l'équipe partenaire doit être désigné

Partenaire 2 * **champ obligatoire**

Civilité *	**Nom ***	**Prénom ***
Grade*		**Employeur ***
Mail *		
Tél *	**Fax :**	

Laboratoire * *(nom complet)*	
Code Unité *(s'il existe)* Ex : UMR 5232, EA 567	
Adresse complète du laboratoire *	
Code postal *	**Ville * :**

Etablissements de tutelle *(indiquer le ou les établissements et organismes de rattachement, et en n°1 l'établissement susceptible d'assurer la gestion du projet)* **:**

1.
2.
3.
4.

Acronyme ou titre court du projet :

A-2 : Autres partenaires du projet (remplir une fiche par partenaire)
Un responsable scientifique de l'équipe partenaire doit être désigné

Partenaire 3 * **champ obligatoire**

Civilité *	**Nom ***	**Prénom ***

Grade*		**Employeur ***	
Mail *			
Tél *		**Fax :**	

Laboratoire * *(nom complet)*	
Code Unité *(s'il existe)* Ex : UMR 5232, EA 567	
Adresse complète du laboratoire *	
Code postal *	**Ville * :**
Etablissements de tutelle *(indiquer le ou les établissements et organismes de rattachement, et en n°1 l'établissement susceptible d'assurer la gestion du projet)* **:** **1.** **2.** **3.** **4.**	

Acronyme ou titre court du projet :

A-2 : Autres partenaires du projet (remplir une fiche par partenaire)

Un responsable scientifique de l'équipe partenaire doit être désigné

Partenaire 4 * **champ obligatoire**

Civilité *	**Nom ***	**Prénom ***

Grade*		**Employeur ***	
Mail *			
Tél *		**Fax :**	

Laboratoire * *(nom complet)*	
Code Unité *(s'il existe)* Ex : UMR 5232, EA 567	
Adresse complète du laboratoire *	
Code postal *	**Ville * :**

Etablissements de tutelle *(indiquer le ou les établissements et organismes de rattachement, et en n°1 l'établissement susceptible d'assurer la gestion du projet)* **:**

1.
2.
3.
4.

Principales publications :

Liste des 10 principales publications ou brevets de l'équipe partenaire 2 (définie tableau ci-dessous) au cours des cinq dernières années, relevant du domaine de recherche couvert par la présente demande dans l'ordre suivant : Auteurs (en soulignant les auteurs faisant effectivement partie de la demande), Année, Titre, Revue, N°Vol, Pages. N'indiquez pas les publications soumises.

Acronyme ou titre court du projet :

A-2 : Autres partenaires du projet (remplir une fiche par partenaire)

Un responsable scientifique de l'équipe partenaire doit être désigné

Partenaire 5 * **champ obligatoire**

Civilité *	**Nom ***	**Prénom ***

Grade*		**Employeur ***	
Mail *			
Tél *		**Fax :**	

Laboratoire * *(nom complet)*	
Code Unité *(s'il existe)* Ex : UMR 5232, EA 567	
Adresse complète du laboratoire *	
Code postal *	**Ville * :**

Etablissements de tutelle *(indiquer le ou les établissements et organismes de rattachement, et en n°1 l'établissement susceptible d'assurer la gestion du projet)* **:**

1.
2.
3.
4.

Principales publications :

Liste des 10 principales publications ou brevets de l'équipe partenaire 2 (définie tableau ci-dessous) au cours des cinq dernières années, relevant du domaine de recherche couvert par la présente demande dans l'ordre suivant : Auteurs (en soulignant les auteurs faisant effectivement partie de la demande), Année, Titre, Revue, N°Vol, Pages. N'indiquez pas les publications soumises.

Programmes SHS 2007

B - Description du projet

Acronyme ou titre court du projet : CORPTEF

Les objectifs, l'originalité du projet, la problématique, les méthodologies employées et les modalités d'accès aux terrains, le programme des travaux et ses différentes phases, la bibliographie et l'état de l'art, les modalités de valorisation des connaissances doivent être présentées. Les modalités de mise en œuvre de l'interdisciplinarité éventuelle et des diverses collaborations doivent être précisées et justifiées en accord avec l'orientation du projet. Les modalités de coordination et de travail en commun des différents partenaires doivent être décrites.

La capacité de ou des équipes doit être attestée par la qualification et les productions scientifiques antérieures de leurs membres. Leurs rôles dans les différentes phases du projet doivent être précisés et la valeur ajoutée des collaborations entre les différentes équipes sera argumentée. Les moyens demandés doivent être justifiés au regard des objectifs scientifiques du projet et du programme des travaux.

B-1 – Objectifs, contexte et état de la question, originalité

Motivations du projet et contexte dans lequel il s'insère

1) Le renouveau de la linguistique historique et des corpus de langues anciennes

Les recherches en linguistique historique, et plus spécifiquement sur l'histoire du français, connaissent actuellement un essor important. Après plusieurs décennies d'un primat très largement accordé aux recherches synchroniques, la linguistique historique a repensé son champ et ses méthodes depuis une vingtaine d'années. Elle l'a fait tout d'abord en utilisant, comme pour les mettre à l'épreuve, certains concepts essentiels des théories linguistiques contemporaines (théories cognitives et informationnelles, théories de l'énonciation...). Par ailleurs, une nouvelle approche, dite de la « grammaticalisation », s'est fait jour à la fin des années 80. Diachronique par définition, cette approche est en plein développement et jamais sans doute depuis le XIXe siècle on n'avait vu paraître autant d'études diachroniques ou historiques, et sur des langues aussi diverses.

Or un trait inhérent à la démarche historique est le recours au corpus. Par définition, les historiens de la langue travaillent sur des états de langue pour lesquels le linguiste n'a pas de compétence et pour lesquels *il n'existe plus de locuteur « natif »*, donc de jugement en dernière instance sur la grammaire de ces états de langue anciens. L'étude historique suppose donc le traitement de corpus, vastes et facilement accessibles. Outre la sécurité que procure la prise en compte et l'analyse d'un très grand nombre d'occurrences, on sait désormais qu'il est utile de repérer et de prendre en compte également les très basses fréquences, ce qui conduit à complexifier la notion de forme (ou construction) marquée, et d'examiner systématiquement aussi les absences. Que tel ordre des mots, telle forme, n'apparaissent pas (ou plus) dans un état de langue donné, est aussi important pour en décrire la grammaire que le fait qu'elle possède, attestée par des milliers d'occurrences, telle ou telle construction. Et, bien sûr, l'exploration systématique des plus anciens textes permet de situer la première apparition d'une forme ou d'une construction : on a constaté que plus on possédait de documents anciens explorables, plus la date des 'premières attestations' reculait dans le temps.

2) Utilité des corpus de langue ancienne pour les recherches en français moderne

Les chercheurs travaillant sur le français moderne ont de plus en plus recours aux corpus de français ancien. Il est très fréquent en effet que les recherches menées, dans un cadre sémantique notamment, sur des phénomènes et des catégories linguistiques du français moderne en viennent à explorer les états de langue passés. Dans bien des cas, seule une étude diachronique permet de comprendre et de décrire le fonctionnement actuel de telle ou telle forme ou catégorie linguistique, en permettant d'examiner la constance de ces catégories et de leurs traits définitoires à travers le temps. Des recherches de ce type sont en cours dans le laboratoire ICAR, et permettent de réunir autour d'un même projet, comme le projet « Evolution linguistique et corpus » (ELICO), spécialistes de sémantique, de

diachronie du français et d'informatique. Ce type de travail collaboratif, fondé sur l'exploitation d'un corpus de français ancien assez étendu pour qu'on puisse analyser les fréquences et les contextes d'apparition de telle ou telle forme, sera amené à se développer, à n'en pas douter, dans les années à venir.

3) Intérêt de la période très ancienne du français

Différentes recherches en cours ont montré la nécessité de remonter le plus en amont possible pour comprendre et décrire le fonctionnement des unités linguistiques. Outre qu'une couverture historique maximale permet d'engager des recherches sur un empan chronologique plus grand – ce qui permet de mieux cerner les étapes successives de l'évolution, du point de départ au point d'arrivée actuel – le recours aux données les plus anciennes est souvent déterminant dans notre compréhension des faits langagiers et de leur évolution. Une étude récemment menée par Christiane Marchello-Nizia (2004a, 2006) sur les démonstratifs en français a ainsi montré une évolution très nette de la fonction pragmatique et du contenu sémantique de ces déterminants entre le très ancien français (IXe- XIIe siècles) et ce qu'on a coutume d'appeler (à tort ou à raison) l'ancien français « classique » (XIIIe siècle). L'analyse du fonctionnement des démonstratifs pendant la période antérieure au XIIIe s'avère très nécessaire à la compréhension de la façon dont le français s'est peu à peu démarqué du latin. Elle permet d'envisager cette évolution sans rupture et constitue de ce fait une donnée tout à fait fondamentale pour notre connaissance des processus et des étapes par lesquels le français a évolué du latin jusqu'à nos jours.

Enfin, nous sommes souvent à la recherche des premières attestations de telle forme ou de telle construction en français. Ce sont les premières attestations qui permettent de vérifier quels sont les traits les plus anciens – les traits 'primitifs' – de la forme ou construction en question et qui permettent ainsi aussi de proposer une explication de l'apparition de cette forme ou construction. Là encore le recours à un corpus de textes très anciens s'avère tout à fait décisif.

4) Pour un développement des corpus de français médiéval existants

Les corpus de français médiéval existants ne sont ni assez vastes, ni assez diversifiés à ce jour. La question de la taille requise pour qu'un corpus soit vraiment exploitable reste toujours en suspens. Diverses voix, celle de Claire Blanche-Benveniste par exemple, plaident pour un minimum de 10 millions d'occurrences. Or les corpus de français médiéval actuellement accessibles sont loin d'approcher cette taille : la Base de français médiéval (http://bfm.ens-lsh.fr/) compte environ 3 millions d'occurrences, le corpus du Laboratoire de Français Ancien (TLFA, Ottawa, Pierre Kunstmann ; http://www.uottawa.ca/academic/arts/lfa/) en compte 3,5 millions, le corpus élaboré à Amsterdam par A. Dees et diffusé par Achim Stein (Université de Stuttgart ; http://www.uni-stuttgart.de/lingrom/stein/corpus/), qui mêle éditions critiques et éditions de manuscrits médiévaux, en compte 3 millions également. Quant à l'Anglo-Norman On-Line Hub (David Trotter, Université d'Aberystwyth ; http://www.anglo-norman.net/), il ne diffuse à ce jour qu'une dizaine de textes. A terme, ce corpus devrait atteindre 5 millions d'occurrences-mots.

Parmi toutes les autres entreprises de constitution de corpus en cours (en particulier le projet de numérisation de chartes lorraines dirigé par Martin-Dietrich Gleßgen, Université de Zurich, le projet *Khartês* de numérisation de chartes wallonnes dirigé par Marie-Guy Boutier, Université de Liège, et le projet « Modéliser le changement : les voies du français » dirigé par France Martineau, Université d'Ottawa), aucune n'a déjà atteint cette masse critique. Un accroissement des bases existantes est donc tout à fait souhaitable.

Par ailleurs, ces différents corpus ne reflètent généralement pas la diversité des données disponibles. Souvent centrés sur une période donnée (par exemple les XIIIe-XIVe siècles pour le corpus littéraire d'Amsterdam), ils ont été constitués autour d'un projet de recherche relativement précis. Le TLFA par exemple rassemble une collection très importante de miracles médiévaux, mais offre assez peu de données sur des documents scientifiques ou juridiques. Le corpus d'Amsterdam rassemble quant à lui un grand nombre d'extraits de textes littéraires, mais on y trouve très peu de textes didactiques ou religieux par exemple. Quant à l'Anglo-Norman On-Line Hub, qui ne concerne que les textes écrits en anglo-normand, il ne saurait refléter la diversité linguistique de la période médiévale puisqu'il ne donne accès qu'à un ensemble de données limité sur le plan géographique (aire anglo-normande).

Enfin, même lorsqu'un corpus dispose de données plus nombreuses et plus variées, comme c'est le cas du DMF pour la période postérieure (6 800 000 occurrences au total), la diversité typologique des textes n'est généralement ni pleinement affirmée ni vraiment mise en évidence. Elle n'est donc pas rendue directement exploitable pour les utilisateurs, les métadonnées auxquels ils ont accès ne leur permettant pas de mesurer pleinement le degré de diversité des données qu'ils utilisent.

Or un certain nombre d'outils déjà accessibles, parmi lesquels la bibliographie en ligne du *Dictionnaire étymologique de l'ancien français* (DEAF), donnent des indications très précieuses et parfois relativement fines sur les données textuelles existantes et disponibles (données bibliographiques, variations dialectales, références aux glossaires et aux revues scientifiques, etc.). L'exploitation de ces données dans le cadre des corpus existants, leur échange et leur enrichissement serait de ce fait tout à fait bénéfique pour l'ensemble de la communauté internationale des médiévistes travaillant sur le français.

Le projet que nous souhaitons mettre en œuvre doit répondre à cet état de fait. Il vise à constituer et diffuser un corpus représentatif des textes les plus anciens écrits en français (IXe-XIIe siècles), à partir de textes sélectionnés et organisés selon un ensemble de critères externes bien définis. Ce nouveau corpus, qui viendra rejoindre les textes déjà intégrés à la Base de Français Médiéval, comprendra donc un nombre important de textes du XIIe siècle, les textes antérieurs étant malheureusement assez peu nombreux. Il ouvrira des possibilités tout à fait nouvelles à la recherche linguistique et historique, le XIIe étant le premier siècle pour lequel nous disposons d'un nombre conséquent de textes, littéraires mais aussi techniques. C'est également à cette période que peuvent être identifiées les bases de ce qu'on connait des plus anciennes étapes du français. L'intérêt scientifique est donc grand de pouvoir étudier cette 'région' de l'histoire du français encore en partie inexplorée faute d'accès commode, et de mettre à disposition cet ensemble de données.

Contexte dans lequel s'insère le projet

Développement de la Base de Français Médiéval (http://bfm.ens-lsh.fr/)
Notre projet s'appuie sur une base de textes médiévaux déjà constituée, la Base de Français Médiéval (BFM), internationalement reconnue et utilisée dès à présent par une communauté d'environ 300 chercheurs français et étrangers.
Parmi toutes les bases de français médiéval que nous venons d'énumérer, la Base de Français Médiéval se distingue des autres sur trois points principaux :

- Elle constitue sans nul doute la base dont l'empan chronologique est le plus vaste, les textes représentés allant du IXe siècle au début du XVIe siècle ;
- Elle offre également la particularité qu'elle donne accès, via l'outil de recherche quantitative Weblex, à la quasi totalité des textes écrits entre le IXe et le XIe siècle et conservés jusqu'à nos jours, ainsi qu'à un grand nombre de textes du XIIe siècle, les autres bases s'intéressant davantage à « l'ancien français classique » du XIIIe siècle. Cette caractéristique constitue l'un des atouts de cette base (cf. le point 5 de la section précédente) ;
- Enfin, elle offre dès à présent une relative diversité typologique, et surtout, grâce à un travail de description des textes et à l'élaboration de différentes catégories de descripteurs, elle dispose d'un certain nombre d'outils nécessaires au développement et à l'exploitation de cette diversité (cf. le point 5 de la section précédente). On pourra à terme travailler ainsi uniquement sur les textes picards, champenois ; ou bien sur les textes de la seconde moitié du XIIe siècle, ou bien encore sur ceux d'un même auteur, etc.

Compte-tenu de ces caractéristiques, la Base de Français Médiéval semble être la base qui répond le mieux à ce jour aux exigences et aux objectifs que nous nous sommes donnés dans notre projet de recherche. Elle est donc la mieux placée pour la mise en œuvre de ce projet.

De plus, tous les textes de la BFM ont récemment été formatés et balisés selon les normes internationales les plus répandues et les plus récentes (format XML, selon les recommandations de la TEI P4X, le passage à la version P5 étant en cours), ce qui la distingue encore de plusieurs autres corpus du même type (voir Heiden et Guillot 2002, Heiden et Lavrentiev 2004, ainsi que Heiden, Guillot et Lavrentiev 2005a et b). Elle joue également un rôle de premier plan dans l'animation du Consortium international pour les Corpus de Français Médiéval (cf. ci-dessous), dont l'une des tâches principales est l'élaboration et l'échange de standards communs aux différents corpus de français médiéval.

Relations du projet avec le développement du Consortium international pour les Corpus de Français Médiéval (CCFM)

Créé en 2004 à l'initiative de l'Université d'Ottawa, de l'Ecole normale supérieure Lettres et sciences humaines de Lyon, de l'Université de Stuttgart, de l'Université de Zürich, du laboratoire ATILF (Nancy, CNRS), de l'Université du Pays de Galles et de l'Ecole nationale des chartes, ce consortium international a vocation à fédérer la communauté internationale des médiévistes ayant constitué des corpus de français médiéval ou les utilisant. Au fil de temps en effet, une communauté internationale de chercheurs s'est formée autour de la pratique d'échanges réguliers de textes, des problématiques de recherche communes sur ces textes et des questions liées à l'élaboration de ce type de bases de données. Destiné à faciliter l'échange de données, de pratiques et de méthodologies de recherche, le CCFM constitue donc un lieu de discussion et de réflexion autour de standards communs. Il veille à l'élaboration de ces standards et incite à leur utilisation. C'est pourquoi, lors de sa dernière Conférence qui se tenait en octobre 2006 à Lyon, le CCFM a tenu à inviter l'un des fondateurs de la TEI, Lou Burnard, qui a pu évaluer et conseiller nos pratiques.
Une part importante du travail mené dans le cadre du consortium concerne donc la définition de normes communes concernant l'encodage des données et les descripteurs utilisés dans la caractérisation des unités textuelles. Les descripteurs élaborés dans le cadre du projet BFM ont ainsi été présentés à l'ensemble de la communauté réunie en octobre dernier à l'ENS-LSH de Lyon. Ils ont été approuvés et adoptés, après discussion et modification de certains points, par l'ensemble des partenaires. La liste de ces descripteurs est actuellement diffusée sur le site du CCFM, hébergé à l'ENS-LSH (http://ccfm.ens-lsh.fr/), et un groupe de travail a la charge de poursuivre le travail engagé à Lyon. Les normes d'encodage des textes ont également été discutées lors de la dernière rencontre du CCFM en octobre et ces échanges ont permis de définir un ensemble de principes communs.
Notre projet s'insère donc dans le cadre de l'effort de standardisation engagé par le CCFM, entreprise à laquelle notre équipe contribue grandement et qu'elle anime en grande partie. Les résultats de ce projet auront par conséquent une influence directe sur les travaux du CCFM et ont vocation à être diffusés dans l'ensemble de la communauté internationale des médiévistes qui travaillent sur le français. Les liens qui ont d'ores et déjà été tissés avec le Centre national de ressources textuelles et lexicales (CNRTL) – Marie-Luce Demonet a participé à la dernière rencontre du CCFM organisée à Lyon – contribueront également à la diffusion sur le plan national des résultats de notre projet.

Objectifs du projet

Les objectifs que ce projet entend réaliser sont au nombre de cinq :

1) Ce corpus permettra tout spécialement de développer **notre connaissance de la langue médiévale et de l'évolution du français**, en particulier pour la période la plus ancienne. Il sera exploité par les membres du projet dans un cadre pluri-théorique, les différents partenaires impliqués étant spécialistes de lexicologie et sémantique lexicale, syntaxe, sémantique grammaticale, analyse discursive et sociolinguistique. Son exploitation rendra notamment possible une comparaison avec les états de langue postérieurs, déjà bien représentés dans la BFM, ce qui permettra de caractériser beaucoup plus finement que cela n'a été fait jusqu'ici les spécificités du très ancien français. Le projet permettra de confronter et de vérifier différentes théories sur la variation et sur le changement de la langue.

2) Plus largement, le nouveau corpus rendra possibles des recherches sur **une diachronie particulièrement étendue, allant du IXe au début du XVIe siècle,** et offrira une couverture relativement équilibrée de chacun de ces siècles. La diversité typologique du corpus garantira la généralité des résultats obtenus.

3) Par ailleurs, les données particulièrement anciennes diffusées dans ce cadre permettront aux linguistes de **pister les premières attestations** des phénomènes auxquels ils s'intéressent, ce qui rendra le corpus tout à fait unique et irremplaçable dans le monde pour l'histoire du français. Ces données seront précieuses tout spécialement pour les lexicologues impliqués dans le projet (AND, DEAF), mais aussi pour tous les linguistes et philologues médiévistes qui participent au CCFM et/ou qui exploitent des données anciennes (DMF, partie étymologique du Trésor informatisé de la langue française, etc.)

4) Les corpus fondés sur une **typologie étayée et explicite des documents** qui les composent sont encore trop peu nombreux à ce jour. La mise en place d'une méthodologie efficace et la définition de descripteurs utiles à la caractérisation externe des textes constitue une avancée importante dans les recherches menées sur corpus. Nous souhaitons donc que notre projet puisse favoriser le développement d'entreprises similaires sur des données du même type ou

différentes. En outre, le travail de **caractérisation externe des textes,** préalable à l'exploitation de ces textes, doit être complété par une réflexion, menée a posteriori et après analyse des données, sur les types définis en premier lieu. Il est tout à fait probable que l'exploitation des textes enrichira la connaissance que nous avons de la langue médiévale - c'est là notre objectif premier -, mais aussi des textes particuliers que nous exploitons, ce qui constitue également un point important. Une telle démarche devra nous conduire à un enrichissement, voire à une révision complète, des traits qui avaient été définis au départ et elle participera au développement de notre connaissance des textes médiévaux. Cet aspect du projet s'inspirera des recherches les plus récentes de la linguistique variationnelle, qui n'ont pas encore été appliquées sur des corpus diachroniques du français.

5) Enfin, notre projet permettra de **conserver, diffuser et valoriser le patrimoine culturel, linguistique et littéraire de la France**. Ce faisant, il donnera accès à un ensemble de données remarquables et utiles aussi bien au linguiste, au littéraire, à l'historien, à l'ethnologue... qui s'intéresse à cette période de notre histoire. Par ailleurs, le format choisi pour l'encodage de ces données en rendra l'exploitation d'autant plus aisée qu'il s'agit actuellement du format le plus répandu dans le monde.

Notre projet vise donc non seulement à donner la plus large diffusion possible à un corpus de données utiles à toutes sortes de recherches, mais il doit également contribuer au développement d'une méthodologie de corpus, garante de l'exploitation future des données.

B-2 – Description du projet et résultats attendus

Problématique dans laquelle s'insère le projet

a) Elaboration et utilisation d'une méthodologie de corpus

Notre projet s'insère dans le cadre général de la méthodologie de corpus et vise à ce titre à rendre possible l'exploitation des données textuelles mises à disposition à travers les outils. La notion de corpus, telle qu'elle est définie dans ce cadre, est relativement restreinte et ne recouvre pas ce que l'on peut appeler une simple collection de données réunies ici ou là : « Un corpus est une collection de données langagières qui sont sélectionnés et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » (J. Sinclair 1996, p.4). Un corpus se définit donc comme un ensemble de données dont on connaît les sources, dont la qualité est contrôlée et dont les principes de sélection et d'organisation sont raisonnés et explicites (B. Habert, A. Nazarenko, A. Salem 1997). Le respect de ces différents principes conditionne en effet l'exploitation future des données. On ne peut interpréter correctement les informations qualitatives et quantitatives fournies par le corpus si l'on ne dispose pas d'informations suffisantes sur la nature des documents qui le composent.

La définition de principes méthodologiques clairs dans la constitution, la gestion et l'exploitation des corpus textuels s'avère d'autant plus essentielle que la demande en corpus croît de façon spectaculaire en ce moment. Les possibilités nouvelles offertes par la Toile ouvrent en effet de nouvelles perspectives. Aussi constate-t-on un mouvement parallèle de réflexion méthodologique sur la façon dont on peut utiliser la Toile comme source principale de données langagières. Ce mouvement en cours illustre l'importance et la pertinence des questions méthodologiques dans la recherche fondée sur corpus.

La méthodologie utilisée dans le cadre de notre projet vise avant tout à évaluer et contrôler au mieux la représentativité des données rendues exploitables. Bien qu'on sache qu'aucun corpus ne peut représenter une langue dans sa totalité, d'autant que la notion de langue n'est qu'une abstraction couvrant une multitude de niveaux structurels, de registres stylistiques et de dialectes géographiques et sociaux, néanmoins, la linguistique de corpus cherche à maîtriser cette diversité pour pouvoir évaluer la généralité des phénomènes observés. C'est donc dans cette perspective que nous souhaitons élaborer un dispositif méthodologique permettant de sélectionner et d'organiser les textes de notre corpus au sein d'un cadre typologique défini. Un corpus organisé typologiquement doit permettre d'évaluer le degré de généralité des résultats obtenus à partir de celui-ci.

Le cadre ainsi défini aura donc une triple fonction : il guidera la sélection des textes à ajouter au corpus existant, puisqu'il permettra d'identifier les zones non couvertes par le corpus actuel. Il sera ainsi le garant de la pertinence de nos choix. D'autre part, cet ensemble de descripteurs sera un outil essentiel pour les utilisateurs de notre base. Il permettra notamment qu'on puisse sélectionner un ensemble de textes à partir des critères établis dans ce cadre, ces critères devant refléter les principaux facteurs de variabilité rencontrés dans les textes. Les utilisateurs pourront ainsi choisir de travailler sur différentes catégories de textes (les romans, les textes champenois, les textes du domaine juridique...), et les

différents critères utilisés pourront être croisés. Enfin, ce travail de description préalable garantira la validité des résultats obtenus à partir des données, puisqu'il permettra de définir le degré de généralité de ces résultats.

Grâce au cadre méthodologique que nous souhaitons appliquer, nous devrons donc être à même de répondre à **trois exigences**, qui sont nécessaires à la constitution et à l'exploitation d'un corpus de données représentatif :

- le corpus devra être composé d'un **volume de données suffisant** pour qu'on puisse en tirer des résultats significatifs ;
- il devra être **suffisamment diversifié** pour que les données soient interprétables pour des recherches en langue, l'objectif ultime étant de parvenir à un équilibre dans la représentation des données reflétant cette diversité ;

les **facteurs de variabilité des données** devront être **maîtrisés,** rendus **explicites** aux utilisateurs du corpus et pourront servir à la constitution de sous-corpus particuliers à des recherches précises.

Ce sont ces différents principes qui guident depuis plusieurs années le développement de la Base de Français Médiéval. Ils conditionnent déjà le choix des textes que nous ajoutons à la base et nos choix de description des données textuelles. Or, en l'état actuel de notre corpus, l'ajout d'un ensemble de données du XIIe siècle permettra d'atteindre un certain équilibre entre les documents des XIIe et XIIIe siècles (ces derniers constituant la majeure partie de la production écrite en français entre le IXe et le XIIIe siècle).

b) Elaboration d'un cadre typologique pour la BFM

L'essor constant de la linguistique de corpus a permis l'élaboration et l'utilisation de cadres typologiques de toutes sortes dans l'étude en corpus des langues et de leurs usages (voir en particulier les travaux fondateurs de Biber 1988, 1995, 1998, Lee 2001 et Habert 2000, ainsi que par exemple Bilger 2000, Malrieu 2004, Habert et Fuchs 2004, Adam 1999, Rastier 2001, etc.). Par ailleurs, les historiens de la langue, philologues et linguistes diachroniciens, qui travaillent depuis toujours sur des données attestées, sont particulièrement sensibles à cette question de la représentativité des données. Les variations de tous ordres, si prégnantes dans les textes médiévaux, ont fait l'objet de recherches nombreuses et toutes ces études sont particulièrement utiles pour l'établissement du cadre typologique que nous souhaitons construire. La littérature portant sur la variation selon l'usager (variation diatopique et diastratique), ou selon l'usage (variation diaphasique et diamésique), particulièrement abondante pour la langue ancienne (voir par exemple Lodge 1993 et 2004, Lusignan 2004, Trotter 2003, Glessgen et Buchi 2001, Dees 1980 et 1987...) offre donc des outils tout à fait essentiels dans l'élaboration du cadre typologique qui doit servir de fondement à notre corpus. Un certain nombre des principaux représentants de ce courant de recherche sont membres de notre projet et leur participation à la mise en place de ce dispositif descriptif est tout à fait essentielle. Il s'agit notamment de Serge Lusignan, Frankwalt Möhren, Lene Schøsler et David Trotter. Par ailleurs, Françoise Vielliard, professeure à l'Ecole nationale des chartes, a accepté également d'intervenir comme experte sur ces questions.

Les recherches menées depuis plusieurs années par l'équipe du DEAF et désormais rendues accessibles dans une grande base bibliographique (IXe-XIVe siècles) diffusée sur la Toile seront bien entendu l'une des principales sources utilisées. En outre, le travail déjà engagé dans le cadre du développement de la Base de Français Médiéval servira également de base à la mise en œuvre de cet aspect de notre projet (voir en particulier Guillot, Heiden et Lavrentiev à par., Guillot, Lavrentiev et Marchello-Nizia à par., et Lavrentiev à par.). La liste des descripteurs déjà définie dans ce cadre devra être complétée, affinée et enrichie grâce à l'apport et l'expertise scientifique de l'ensemble de nos partenaires et du CCFM, et grâce aussi aux recherches linguistiques menées en parallèle, qui permettront d'exploiter et d'éprouver sur des problématiques linguistiques précises le cadre ainsi établi (voir par exemple Marchello-Nizia 2004b et 2005, Prévost 2005, Guillot, Heiden et Lavrentiev à par., Lavrentiev à par., Guillot et Mortelmans à par.). La pertinence des taxonomies en cours d'élaboration en partenariat avec l'ensemble des membres du CCFM (pour la description des données génériques ou dialectales par exemple) sera de la sorte garantie, et ces taxonomies auront vocation à être amplement diffusées dans la communauté des médiévistes.

c) Diffusion du corpus et de son cadre typologique

Le nouveau corpus réalisé dans le cadre du projet sera interrogeable en ligne, gratuitement, grâce au moteur de recherche Weblex. Il constituera un corpus particulier et identifiable, mais les textes qui le

composent seront également intégrés à la BFM actuelle, ce qui permettra des recherches sur une diachronie beaucoup plus large.
Comme on l'a vu plus haut, la base des descripteurs dont le projet permettra d'établir une liste précise et dont les valeurs seront définies pour chaque unité textuelle devra également être rendue accessible aux utilisateurs du corpus, ce qui constituera une avancée essentielle pour la communauté de ceux qui utilisent d'ores et déjà la BFM ou qui seront amenés à le faire.
On devra également permettre l'accès à la base des descripteurs pour tous les membres du projet. Ils pourront ainsi intervenir directement sur la base et introduire de nouvelles informations ou modifier des informations existantes. L'outil à concevoir devra donc répondre à trois fonctionnalités essentielles : l'organisation des descripteurs retenus pour la caractérisation des textes à l'intérieur d'une base, un accès à distance à cette base pour les membres du projet, un accès plus restreint pour les utilisateurs du corpus qui se serviront de cet outil dans la définition de leur sous-corpus de travail.
D'autre part, on souhaite proposer à l'ensemble de la communauté du CCFM d'exploiter cette liste de descripteurs et de l'utiliser dans leurs corpus propres, ce qui constitue l'une des retombées majeures de notre projet, le travail accompli sur notre base ayant de ce fait vocation à être directement ré-exploité par la communauté des médiévistes regroupés dans le CCFM. C'est donc à terme l'ensemble des données accessibles dans le monde sur le français médiéval qui bénéficieront de la valeur ajoutée par notre projet. Nous pourrons ainsi fédérer les travaux en cours et éviter aussi que soient menées en différents endroits des entreprises similaires.

d) Evaluation et maîtrise de la qualité des données

On veillera à prendre en compte deux aspects particulièrement épineux dans la constitution et la description de corpus de données médiévales.

- Les problèmes posés par l'exploitation de sources secondaires

Notre projet de recherche vise à constituer et diffuser à grande échelle un ensemble de données secondaires sur la langue et les textes du Moyen Age français. D'autres projets parallèles ont fait le choix de ne pas travailler à partir d'éditions critiques déjà publiées mais d'éditer sous forme numérique des manuscrits médiévaux (voir en particulier « The Charrette Project », http://lancelot.baylor.edu/, projet avec lequel la BFM a toujours eu un lien fort, auquel Alexei Lavrentiev a activement participé pendant une année et dont il est toujours membre, ainsi que l'édition en ligne du manuscrit de Lyon de la *Queste del saint Graal*, roman du XIIIe siècle, réalisée par notre équipe). Il s'agit là de partis pris dont les fondements sont surtout pratiques. Le souci d'aller vite et de procurer à la communauté un ensemble de données conséquent a motivé notre choix. On parvient en effet à des résultats bien différents dans l'un et l'autre cas : sur une même période d'une quinzaine d'années, notre base s'est enrichie de 74 textes intégraux d'ancien et de moyen français ; le Projet Charrette a quant à lui réalisé l'édition d'un seul texte, le *Chevalier de la Charrete* de Chrétien de Troyes, dont les 8 manuscrits existants ont été transcrits et édités sur la Toile.
Le recours à des éditions critiques, s'il permet de rassembler assez rapidement un grand nombre de données, pose cependant un certain nombre de problèmes, en particulier en ce qui concerne la fiabilité de ces données et la possibilité de leur exploitation future. Le premier problème est celui de la **qualité de l'édition choisie**. L'authenticité des données dépend en grande partie du type d'édition choisi, ainsi que, bien sûr, de la qualité de l'édition. Dans un grand nombre de cas, la qualité du travail éditorial effectué sur ces textes autorise qu'on les utilise comme sources documentaires, les outils de balisage dont nous disposons nous permettant par ailleurs d'apporter les modifications et rectifications nécessaires. Enfin, grâce aux excellents outils bibliographiques que nous possédons pour l'ancien français (bibliographie du DEAF, *Manuel bibliographique* de Bossuat et ses compléments, *Dictionnaire des lettres françaises*...), nous sommes suffisamment renseignés sur la valeur des éditions pour éviter celles qui paraissent trop peu fiables.
La seconde difficulté que pose l'exploitation de données secondaires est d'ordre **juridique**. Les éditions critiques que nous utilisons étant généralement sous droits, elles ne peuvent être librement accessibles à tout utilisateur. Un volet important de notre projet de recherche concerne la clarification du statut juridique des données dont nous disposons et la mise en place d'une procédure claire et sécurisée pour leur diffusion et leur exploitation. Différents projets en cours, dont celui de la base textuelle FRANTEXT, connaissent la même situation et cherchent à utiliser les possibilités offertes par la législation actuelle (dans le cas de FRANTEXT, on utilise en l'étendant au maximum le principe du droit de citation pour donner accès à un contexte de 300 mots aux utilisateurs de la base). Un travail de réflexion, comparable à celui qu'ont mené la DGLFLF et Olivier Baude sur les corpus oraux (cf. *Corpus oraux. Guide des bonnes pratiques*), serait très profitable pour l'écrit. L'état imparfait de la législation entravant partiellement le développement de la science, il s'avère nécessaire d'utiliser toute la latitude permise par les textes actuels.

Ce projet permettra également de faire avancer la réflexion sur la nature de la 'valeur ajoutée' apportée par le traitement des sources utilisées dans la constitution du corpus et l'expertise que suppose ce traitement. Les textes ajoutés au corpus étant numérisés, formatés, encodés et enrichis, ils ne sont plus le reflet direct de leurs originaux imprimés et possèdent un grand nombre d'informations supplémentaires qui permettent une exploitation bien différente.

- Les problèmes de datation des textes

Le plus souvent, l'attribution d'une date précise aux textes médiévaux pose des problèmes épineux. Doit-on retenir la date de composition de l'œuvre ou bien la date du manuscrit de base suivi par l'éditeur ? Dans certains cas, l'écart entre ces deux dates est supérieur à un siècle (cf. l'exemple du *Jeu de saint Nicolas* de Jean Bodel, composé entre 1191 et 1202 et dont le seul manuscrit conservé date des années 1295, et l'exemple du premier texte écrit en français en 842, les *Serments de Strasbourg,* conservé dans un manuscrit latin daté de l'an Mil). On sait par ailleurs que des manuscrits tardifs présentent parfois des formes linguistiques plus archaïques que des manuscrits plus anciens. En outre, la date de composition d'un texte n'est souvent connue que très approximativement. La datation de certains textes se situe **assez souvent dans une fourchette d'une ou deux décennies, mais parfois d'un demi-siècle** (par exemple, la seconde moitié du X^{e} siècle pour la *Passion de Clermont*), ou même davantage (par exemple, le XIII^{e} siècle pour le *Lancelot en prose*). Enfin, il est souvent très difficile aussi de dater précisément les manuscrits médiévaux qui nous sont parvenus.

Il s'agit là de points méthodologiques essentiels, qui devront être discutés et pris en compte dans le cours de notre projet, l'élaboration d'une méthodologie claire et raisonnée pour traiter ces questions constituant une avancée en soi. Dans un premier temps, la sélection des textes qui devront composer le corpus représentatif des premiers textes français a été faite à partir de la date présumée de composition de l'œuvre, mais au cours du projet nous veillerons à ce que la date du manuscrit soit également prise en considération dans l'organisation et l'exploitation du corpus.

e) Exploitation scientifique du corpus

L'ensemble des partenaires sollicités dans le cadre de ce projet ont été sélectionnés pour la qualité de leurs recherches scientifiques mais aussi pour la complémentarité de celles-ci. L'un des objectifs scientifiques de notre projet étant de parvenir à une meilleure connaissance du plus ancien français, et plus spécifiquement de la langue du XIIe siècle, nous avons fait le choix de rassembler plusieurs représentants des différentes spécialités de la linguistique, tous étant par ailleurs particulièrement qualifiés dans le domaine de la linguistique diachronique et de l'histoire du français. L'objectif de notre projet est donc de parvenir à terme à une description de cette langue qui prenne en compte les différents niveaux de l'analyse linguistique.

Un premier volet de recherches concernera les données lexicales fournies par notre corpus. La bonne représentativité des différentes variétés dialectales du français de cette période permettra notamment des recherches sur les particularismes dialectaux présents dans le lexique des textes exploités. David Trotter et Frankwalt Möhren, qui sont à la tête de deux des principales entreprises lexicographiques en cours dans le monde sur le français médiéval (l'Anglo-norman Dictionary d'un côté et le Dictionnaire étymologique de l'ancien français de l'autre) sont avec Cinzia Pignatelli les principaux chercheurs de notre projet engagés dans ce type de problématiques. Nous profiterons également de l'expertise de Marie-Guy Boutier et Nicolas Mazziotta (Université de Liège).

Un second volet de la recherche concernera la syntaxe des textes exploités. De ce point de vue, notre projet aura la chance de bénéficier de recherches menées en parallèle par Achim Stein à l'Université de Stuttgart dans le cadre d'un projet d'annotation syntaxique (manuelle et automatique) du Nouveau Corpus d'Amsterdam. Ce projet, parallèle au nôtre, prolonge le travail déjà effectué sur le corpus d'Amsterdam, anciennement étiqueté par l'équipe d'A. Dees grâce à un ensemble de 225 étiquettes morphosyntaxiques et récemment lemmatisé par Achim Stein et Pierre Kunstmann. Cinq textes de la Base de Français Médiéval ont également été enrichis au moyen d'un jeu d'une cinquantaine d'étiquettes morphosyntaxiques (élaboré par Sophie Prévost), et notre collaboration avec Achim Stein et Serge Heiden est d'ores et déjà effective dans ce domaine (expert dans le cadre du projet). Il s'agit là de l'un des axes tout à fait essentiels pour le développement et l'exploitation scientifique de notre base, puisque tout ce travail d'enrichissement offre de nouvelles possibilités de recherche dans les textes, mais aussi pour l'enrichissement de notre connaissance de la syntaxe médiévale et de son évolution.

Un premier travail d'étiquetage morphosyntaxique et de lemmatisation sera mené sur les textes du corpus, à partir de l'outil TreeTagger et des étiquettes utilisées sur le corpus d'Amsterdam. L'annotation syntaxique des textes se fera dans un second temps, grâce à un partenariat étroit avec l'équipe de Stuttgart. Elle se déroulera en quatre phases. Une première phase manuelle concernera l'annotation

syntaxique de certains syntagmes (CP, IP, PP, NP, AP) et de certaines fonctions syntaxiques (sujet, objet, autres compléments du verbe). La phase suivante sera consacrée à l'entraînement du « chunker » (parseur partiel) qui utilise le même type de paramètres que le TreeTagger. Dans un troisième temps, on procèdera à des essais d'annotation automatique sur des textes non balisés. Enfin, nous effectuerons des corrections manuelles qui permettront, entre autres, d'améliorer le chunker. Le but ultime de cette annotation syntaxique est double : introduire un balisage des phrases/syntagmes dans le corpus et développer un nouvel outil de balisage syntaxique pour l'ancien français.
Par ailleurs, bon nombre des membres de notre projet mènent actuellement des recherches sur la syntaxe médiévale : Lene Schøsler sur l'ordre des mots et les questions de valence verbale notamment, Sophie Prévost sur l'ordre des mots également, Alexei Lavrentiev sur la ponctuation médiévale dans différents types de textes et son interaction avec l'organisation syntaxique de l'énoncé.
Différents chercheurs de notre équipe sont également engagés dans des recherches en sémantique grammaticale, en particulier sur le système des déterminants en français (Anne Carlier, Céline Guillot et Christiane Marchello-Nizia, experte dans le cadre du projet). Ces travaux en cours, qui visent à étudier par exemple comment les langues romanes en sont venues à exprimer par des morphèmes spéciaux les notions de 'définitude' et d''indéfinitude' ou comment se sont développés les indéfinis et les démonstratifs, devraient continuer à se développer à l'avenir, dans l'optique des grammaires cognitives notamment.
En outre, l'un des axes de recherche communs à plusieurs des travaux que nous venons de mentionner concerne les questions de cohésion et de cohérence textuelle ainsi que le mode d'organisation et de structuration discursive des textes médiévaux (recherches sur les déterminants, la ponctuation médiévale, l'ordre des mots notamment). C'est également dans ce cadre que se situent les travaux de Sophie Marnette, qui portent sur le discours rapporté dans les textes médiévaux et qui se situent dans perspective pragmatique, discursive et textuelle.
Par ailleurs, tout l'appareil descriptif qui aura été défini par notre projet et qui sera composé d'un ensemble de métadonnées externes dont certaines ont trait à l'usage des textes et à leur fonction sociale rendra également possible une exploitation du corpus dans une optique sociolinguistique. Serge Lusignan sera l'un des principaux chercheurs impliqué dans cet axe de recherche, qui croise également les problématiques impliquées par plusieurs des travaux précédemment cités (les recherches de Sophie Marnette, Alexei Lavrentiev et Céline Guillot notamment).
Enfin, plusieurs recherches en cours dans notre équipe sont menées dans le cadre de la théorie de la grammaticalisation, et il est très souhaitable que la constitution de ce nouveau corpus puisse favoriser les travaux de ce type. Ces recherches, qui concernent surtout Christiane Marchello-Nizia, Lene Schøsler, Anne Carlier et Sophie Prévost pour l'instant, seront sans aucun doute amenées à se développer à l'avenir, étant donné la place grandissante prise par ce nouveau cadre théorique dans le domaine de la linguistique diachronique. Un groupe de chercheurs récemment constitué grâce au GDR 2349 (CNRS) "Diachronie du français et évolution des langues" et financé par l'Institut de Linguistique française (resp. Sophie Prévost) a pour tâche d'étudier différents phénomènes de grammaticalisation en français, comme le développement des locutions conjonctives (du type de *avant que*, *dès que),* des marqueurs de topicalisation (du type de *quant à, au regard de*, *en ce qui concerne, pour ce qui touche à...),* des locutions prépositionnelles, des locutions verbales (*prendre congé* ou *chercher noise,* par exemple*)* et des adverbiaux modalisateurs *(à l'/d'/par aventure* ; *à la/de/en/par/pour vérité*…).
Ce nouveau corpus sera donc abondamment exploité dans le cadre des recherches menées par notre équipe et plus largement dans le monde de la linguistique diachronique du français. Par ailleurs, on souhaite que l'ensemble de la BFM, enrichie de ces textes les plus anciens, puisse servir de base à un projet de grande grammaire historique du français, élaboré sur le plan national sur le modèle du projet actuellement en cours pour le français moderne.

La mise en œuvre de notre projet se fera en étroite collaboration avec deux autres projets soutenus par l'ANR à partir de 2006, ELICO (resp. Lucia Tovena, Université de Paris VII) et Textométrie (resp. Serge Heiden, UMR 5191 ICAR). Le projet ELICO, qui porte sur l'évolution des déterminants sur une longue période de temps (XIIIe-XIXe siècles) occupe en effet une position orthogonale par rapport au nôtre : il concerne l'étude d'un type particulier de marqueurs linguistiques (les déterminants) sur une période temporelle étendue. Le corpus représentatif des premiers textes français couvre au contraire une période temporelle limitée, mais il concerne tous les niveaux de l'analyse linguistique de cette période. Ces deux projets partagent des principes méthodologiques communs concernant la constitution et l'organisation du corpus.
Le second projet retenu en 2006 par l'ANR, le projet Textométrie, vise quant à lui l'élaboration d'un nouvel outil d'interrogation et d'analyse des corpus linguistiques. Ce nouvel outil devra rassembler diverses fonctionnalités aujourd'hui assurées par des logiciels différents. Ces deux projets se développant en parallèle, nous pourrons à la fois faire part aux concepteurs du nouveau logiciel des

fonctionnalités les plus utiles aux futurs utilisateurs et en même temps bénéficier de ce nouvel outil. Le logiciel en cours de construction assurera ainsi une bonne diffusion et une meilleure exploitation du corpus que nous aurons élaboré.

Volume de données à numériser, liste des tâches et calendrier

Une première estimation du volume de données à numériser a été effectuée à partir des textes du XIIe siècle répertoriés dans la bibliographie en ligne du DEAF, membre du CCFM et de notre projet. Cette estimation s'appuie par ailleurs sur l'expérience acquise lors de la constitution de la BFM et tient compte de la spécificité des données à traiter. On ne peut en effet confier le travail de relecture de textes d'ancien français à des personnes totalement inexpérimentées. Bien que nous bénéficiions d'un réseau de relecteurs déjà éprouvé, cette contrainte réduit le choix des personnes impliquées dans cet aspect de notre projet.
Nous évaluons à 25 000 mots les données que nous devrons numériser pour le premier tiers du XIIe siècle (ce qui nous permettra de disposer de la quasi totalité des textes composés à cette période, compte-tenu des textes que nous possédons déjà). Le volume de données que nous souhaitons numériser pour le second tiers du XIIe siècle s'élève à 300 000 mots. Ces données viendront compléter les 200 000 mots que nous possédons déjà pour cette période. Enfin, nous souhaitons numériser pour le dernier tiers du siècle un volume de 200 000 mots, ce qui nous permettra de disposer d'un ensemble de 1 million de mots pour cette dernière période et permettra d'équilibrer cette partie du corpus avec les données déjà disponibles dans la BFM pour le XIIIe siècle.

Les différentes tâches que nous avons identifiées pour la mise en œuvre de notre projet sont présentées ci-dessous :

1. Choix des textes
   - recherche des textes sur critères typologiques (date, dialecte, domaine, forme…)
   - évaluation des éditions disponibles
   - acquisition de l'édition choisie ou copie

2. Mise à jour de la documentation :
   - protocole d'encodage
   - protocole de description
   - protocole d'étiquetage

3. Numérisation du corps du texte :
   - OCR
   - première relecture
   - balisage TEI
   - deuxième relecture
   - vérification finale
   - intégration dans Weblex

4. Description de l'unité textuelle
   - identification de l'œuvre
   - caractérisation typologique
   - identification de l'édition
   - caractérisation du manuscrit de base
   - état de numérisation

5. Enrichissement linguistique (procédure automatique)
   - étiquetage morphosyntaxique
   - lemmatisation
   - annotation syntaxique
   - évaluation de la qualité de l'enrichissement

6. Mise au point des outils d'exploitation
   - élaboration du cahier des charges
   - conception de l'outil de gestion des descripteurs
   - transfert des données de la base existante

- test de l'outil et mise au point de la version finale
- organisation du travail collaboratif à distance
- formation des membres du projet à l'outil

7. Exploitation du corpus
   - requêtes et analyses
   - journée d'études (bilan intermédiaire)
   - colloque
   - publication des actes

Ces tâches seront réalisées selon le calendrier suivant :

Année 1
Choix des textes (mois 1-6)
Mise à jour de la documentation (mois 1-6)
Réunion préparation du corpus (mois 3)
Description des unités textuelles du corpus (mois 3-12)
Numérisation (1600 pages, mois 1-12)
Mise au point des outils (mois 6-12)

Livrables :
1600 pages de textes numérisés, relus et balisés, intégrés BFM1
Protocole de description des textes
Outil d'exploitation des descripteurs
Publication de nouvelles taxonomies
Page web du projet (sur le site de la BFM)

Année 2

Numérisation (1600 pages, mois 13-24)
Enrichissement linguistique (mois 13-24)
Formation aux outils (mois 13)
Exploitation du corpus (13-24)
Journée d'études (mois 24)

Livrables
1600 pages de textes numérisés, relus et balisés, intégrés BFM1
Textes lemmatisés, étiquetés en morphosyntaxe, procédure automatique
Création du corpus représentatif des premiers textes français
Mise en ligne des travaux de la journée d'études

Année 3

Exploitation du corpus (mois 25-36)
Enrichissement linguistique, correction (mois 25-30)
Organisation du colloque et colloque (mois 25-36)

Livrables
Rapport sur la qualité de l'enrichissement
Proposition d'extension des normes de description et d'encodage des textes dans le cadre du CCFM
Actes du colloque
Participation à des colloques ; articles

Résultats attendus

Outre les livrables que nous avons mentionnés, et qui constituent autant de documents et productions élaborés à différents stades du projet, les principaux résultats finaux que nous escomptons sont de trois type. Ils concernent à la fois les aspects méthodologiques et technologiques du projet de recherche, son exploitation scientifique, et plus largement son apport culturel et patrimonial.

Tous ces résultats seront rendus possibles par un souci constant de diffusion du nouveau corpus. La diffusion des données s'appuiera sur des outils existants et éprouvés (sites web, comme celui de la BFM, du CCFM, du CNRTL, etc. ; outils de recherches spécialisés, comme le logiciel d'interrogation Weblex), et sur des outils et productions spécifiques réalisés dans le cadre de ce projet (outil informatique, publications et colloque).

Résultats méthodologiques et technologiques

Comme on l'a vu déjà, une part importante du projet concerne la définition de principes généraux et de normes concernant l'encodage des textes et la description des métadonnées textuelles. Ce travail de normalisation, qui sera mené dans le cadre offert par le consortium international de la Text Encoding Initiative, sera également diffusé dans ce consortium, auquel nous appartenons d'ores et déjà. L'utilité et la pertinence des critères définis dans le cadre de ce travail de normalisation auront par ailleurs été garanties par les recherches menées en parallèle sur le corpus par les membres du projet.

En outre, tout le travail de description des textes sera exploité par la communauté des médiévistes réunis dans le CCFM. Tous les corpus participant au CCFM pourront ainsi s'enrichir des normes mises en place dans notre projet, mais aussi des données définies pour chaque texte. A terme, ce travail rendra possible la mise en place d'un point d'accès central de diffusion des données bibliographiques, documentaires et typologiques pour tous les corpus de français médiéval réunis dans le CCFM.

L'outil informatique de mise en ligne et d'interrogation de la base de descripteurs qui sera élaboré dans le projet pourra également être utilisé dans ce cadre plus général. Il s'agit là d'une extension 'naturelle' et capitale pour nous du travail engagé par notre équipe. Les retombées en seront aussi importantes pour les utilisateurs des corpus de français médiéval, disposant ainsi d'outils performants et d'informations exhaustives sur l'ensemble des données disponibles dans le monde, que pour les équipes ayant en charge la constitution et la diffusion des corpus de français médiéval.

Résultats scientifiques

Le projet permettra tout d'abord la mise en ligne d'un corpus représentatif des plus anciens textes français d'une taille de 1,7 millions de mots environ. De plus, ce corpus sera enrichi grâce à un étiquetage morphosyntaxique, une lemmatisation et une annotation syntaxique.

Ce corpus sera exploité par les linguistes et philologues participant au projet. Ces recherches seront exposées lors d'une journée d'étude organisée à mi-parcours et d'un colloque international organisé à Lyon à la fin de la durée du projet. Le colloque donnera lieu à la publication d'Actes et les travaux exposés lors de la journée d'études seront également mis en ligne.

L'exploitation de ce corpus par des utilisateurs multiples et variés bénéficiera également de deux atouts majeurs de notre équipe :

- notre participation à divers cercles de linguistes spécialistes de la diachronie du français (réseau de l'ancien GDR 'Diachronie et évolution des langues', réseau du colloque international Diachro, qui a lieu tous les deux ans et dont l'une des prochaines réunions pourra se tenir à Lyon, autres projets de recherche en cours qui concernent le français médiéval, comme le projet ELICO).
- l'utilisation d'un logiciel d'interrogation en ligne performant et souple, Weblex, qui saura également s'adapter à de nouveaux types d'exploitation (cf. le projet Textométrie soutenu par l'ANR)

Résultats culturels et patrimoniaux

Le projet permettra la conservation et la diffusion à grande échelle d'un patrimoine particulièrement ancien et important pour l'histoire de notre pays et de sa langue. Ce corpus sera d'autant plus utile et précieux que les éditions qui seront exploitées dans le corpus sont parfois relativement anciennes et difficiles à trouver. Et la pérennité du corpus informatisé sera garantie par les formats utilisés dans le projet.

Enfin, la diffusion d'un volume important de données anciennes et diversifiées favorisant les recherches menées dans le monde sur le français, le corpus représentatif des premiers textes français participera directement aux efforts menés actuellement pour promouvoir la langue française dans le monde. C'est à ce titre que la Base de Français Médiéval a pu bénéficier précédemment d'un financement de la Délégation générale à la langue française et aux langues de France ainsi que d'un financement de l'Institut de linguistique française.

B-3 – Bibliographie

Adam, J.-M. (1999), *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

Baldinger, K. (éd) (1993), *Dictionnaire étymologique de l'ancien français. Complément bibliographique rédigé par F. Möhren*, Tübingen, Niemeyer, [version plus complète en ligne : http://www.deaf-page.de/].

Baude, O. (éd.) (2006), *Corpus oraux : guide des bonnes pratiques*, Paris, CNRS Editions.

Biber, D. (1988), *Variation across Speech and Writting,* Cambridge, Cambridge University Press.

Biber, D. (1995), *Dimensions of register variation: a cross-linguistic comparison*, Cambridge, Cambridge University Press.

Biber, D., Conrad, S., Reppen, R. (1998), *Corpus Linguistics. Investigating Language Structure and Use,* Cambridge, Cambridge University Press.

Bilger, M., éd. (2000), *Corpus : méthodologie et applications linguistiques*. Paris, Champion.

Bossuat, R. (1954 -), *Manuel bibliographique de la littérature française du Moyen âge*, Paris, Librairie d'Argences.

Bossuat, R. et al. (1964 1ère édition), *Dictionnaire des Lettres françaises. Le Moyen Age*, tome 1, Paris, Fayard.

Condamines, A. (éd.) (2005), *Sémantique et corpus*, Paris, Hermès science publ.

Dees, A. (1980), *Atlas des formes et constructions des chartes françaises du 13e siècle*, Tübingen, Niemeyer.

Dees, A. (1987), *Atlas des formes linguistiques des textes littéraires de l'ancien français*, Tübingen, Niemeyer (Beihefte zur Zeitschrift für romanische Philologie 212).

Glessgen, M. et Buchi, E. (2001), « Variétés locales et suprarégionales dans la genèse des langues romanes standard », in : J. François éd., *Les langues de communication : quelles propriétés structurales préalables ou acquises ?*, Mémoire de la Société de Linguistique de Paris, nouvelle série, tome 11, 65-86.

Guillot, C. et Mortelmans, J. (à par.), « Clarté ou vérité, *ledit* dans la prose de la fin du Moyen-âge », à paraître dans les *Mélanges Bernard Combettes*.

Guillot, C., Heiden, S. et Lavrentiev, A. (à par.), « Typologie des textes et des phénomènes linguistiques pour l'analyse du changement linguistique avec la Base de Français Médiéval », à paraître dans les actes du colloque international *Corpus et questionnements du littéraire*, (Université de Paris X, novembre 2005).

Guillot, C., Marchello-Nizia, C. et Lavrentiev, A. (à par.), « La Base de Français Médiéval (BFM) : états et perspectives », à paraître in P. Kunstmann et A. Stein éd., *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*.

Habert, B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? », in : M. Bilger éd., *Linguistique sur corpus. Etudes et réflexions*, Perpignan, Presses Universitaires de Perpignan (n°31 des Cahiers de l'université de Perpignan), 11-58.

Habert, B. (2005), *Instruments et ressources électroniques pour le français*, Gap/Paris, Ophrys (Collection L'essentiel français).

Habert, B. et Fuchs, C. (2004), « Bilan et perspectives méthodologiques", in : *Le français moderne,* 72/1, 88-97.

Habert, B., Nazarenko, A. et Salem, A. (1997), *Les linguistiques de corpus*, Paris, Armand Colin.

Heiden, S. et Guillot, C. (2002), « Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval », in : P. Kunstmann *et al.* éd., *Ancien et moyen français sur le Web, enjeux méthodologiques et analyse du discours*, Ottawa, Les éditions David, 77-92.

Heiden, S. et Lavrentiev, A. (2004), « Ressources électroniques pour l'étude des textes médiévaux : approches et outils », in *: Revue française de la linguistique appliquée*, 1, 99 – 118.

Heiden, S., Guillot, C. et Lavrentiev, A. (2005a), « Manuel d'encodage BFM / XML-TEI, Version 2.1 », *BFM - Base de Français Médiéval* [En ligne], Lyon : UMR ICAR / ENS-LSH <http://bfm.ens-lsh.fr/IMG/pdf/Manuel_Encodage_TEI.pdf>.

Heiden, S., Guillot, C. et Lavrentiev, A. (2005b), « Consignes pour le balisage des textes de la Base de Français Medieval, Version 3.2 », *BFM - Base de Français Médiéval* [En ligne], Lyon : UMR ICAR / ENS-LSH < http://bfm.ens-lsh.fr/IMG/pdf/Consignes_BFM.pdf>.

Kunstmann, P. et Stein, A. (éd.) (2007), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, Stuttgart, Steiner.

Kunstmann, P., Martineau, F. et Forget, D. (2003), *Ancien et moyen français sur le Web. Enjeux méthodologiques et analyse du discours,* Ottawa, Editions David.

Lavrentiev, Alexei (à par.), « Typologie textuelle pour l'étude linguistique de manuscrits français médiévaux », in A. Lavrentiev éd., *Systèmes graphiques de manuscrits médiévaux et incunables français : ponctuation, segmentation, graphies,* Actes de la journée d'étude (Lyon, 6 juin 2005), Chambéry, Presses de l'Université de Savoie.

Lee, D. (2001), « Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle », in : *Language Learning & Technology*, 5, 37-72.
Lodge, R. A. (1993), *French: from dialect to standard*, London/New York, Routledge.
Lodge, R. A. (2004), *A sociolinguistic history of Parisian French*, Cambridge, Cambridge University Press.
Lusignan, S. (2004), *La langue des rois au Moyen Age. le français en France et en Angleterre*. Paris, PUF.
Mac Enery, T. et Wilson, A. (2001,1ère édition 1996), *Corpus linguistics*, Edinburgh, Edinburgh University Press.
Malrieu, D. (2004), « Linguistique de corpus, genres textuels, temps et personnes »**,** in : *Langages,* 153, 73-85.
Marchello-Nizia, C. (2004a), « Deixis and subjectivity: the semantics of demonstratives in Old French (9th-12th century) »**,** in : *Journal of Pragmatics,* 37/1, 43-68.
Marchello-Nizia, C. (2004b), « Linguistique historique, linguistique outillée : les fruits d'une tradition », in : *Le Français moderne,* 72/1, 58-70.
Marchello-Nizia, C. (2005), « A NLP-driven approach to historical linguistics »**,** in : J. Kabatek, C. Pusch, W. Raible éd., *Romanistische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaf, Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*, Tübingen, Gunter Narr Verlag (ScriptOralia ; 130), 11–30.
Marchello-Nizia, C. (2006), « From personal to spatial deixis : The semantic evolution of demonstratives from Latin to French »**,** in : M. Hickman et S. Robert éd., *Space in languages, linguistic systems and cognitive categories*, Amsterdam, Benjamins Publishing Company**:** chapitre 5.
Prévost, S. (2005), « Exploitation d'un corpus de français médiéval : enjeux, spécificités et apports », in : A. Condamines éd., *Sémantique et corpus,* Paris, Hermès/Lavoisier (Série « Traité IC2 » ; Cognition et traitement de l'information), 147-176.
Mellet, S. (2003), « Prenons nos distances pour comparer des textes, les analyser et les représenter », en collaboration avec J.-P. Barthélémy et X. Luong, in : Corpus, 2, 5-18.
Luong, X et Mellet, S. (*2003), «* Mesures de distance grammaticale entre les textes *»,* in *: Corpus*, 2, 141-166.
Rastier, F. (2001), *Arts et sciences du texte*, Paris, PUF.
Sinclair, J. (1996), *Preliminary recommendations on Corpus Typology*, Rap. tech., EAGLES (Expert Advisory Group on Language Engineering Standards), CEE.
Trotter, D. A. (2003), « L'anglo-normand : variété insulaire ou variété isolée ? », in : *Médiévales*, 43-54.
Woledge, B. et Clive, H.P. (1964), *Répertoire des plus anciens textes en prose française depuis 842 jusqu'aux premières années du XIIIe siècle*, Genève, Droz.

B-4 – Collaborations internationales

Elles seront centrales dans le projet, la très grande majorité des membres qui y participent étant étrangers. Les institutions concernées sont les suivantes :

- Université de Stuttgart (Achim Stein)
- Université de Copenhague (Lene Schøsler)
- Université du Pays de Galles (David Trotter)
- Université de Heidelberg (Frankwalt Möhren et Sabine Tittel)
- Université de Montréal (Serge Lusignan)
- Université d'Oxford (Sophie Marnette)

B-5 – Justification scientifique des moyens demandés pour chaque équipe partenaire impliquée dans le projet.

On présentera ici une justification scientifique des moyens demandés pour chacun des partenaires impliqués dans le projet, en distinguant les demandes en équipement, fonctionnement, personnels. Pour les demandes d'équipement, préciser si les achats envisagés doivent être complétés par d'autres sources de crédits, le montant et l'origine des crédits complémentaires qui seront utilisés.

Partenaire 1

Fonctionnement

Frais de mission
Quatre réunions plénières (préparation du corpus, formation aux outils, journée d'études et colloque) sont programmées sur la totalité de la durée du projet. On évalue ces frais de mission à 18 000 euros (420 euros en moyenne par mission pour les étrangers, à l'exception de Serge Lusignan 1000 euros). Diverses réunions qui n'engagent pas la totalité des participants seront également nécessaires, en particulier pour le travail d'enrichissement linguistique (7 500 euros, pour 3 réunions par année pour le déplacement de deux personnes).
Enfin, nous prévoyons des frais pour la participation à différents colloques (colloque de la TEI, CILPR, etc.) : 4000 euros.
Total des frais de mission : 29 500 euros

Achat d'ouvrages : 1000 euros
Nous devrons acquérir, dans la mesure du possible, les éditions imprimées utilisées dans la constitution du corpus. Dans le cas contraire, les ouvrages seront reproduits (avec autorisation).

Achat de logiciels (éditeur XML, logiciel d'OCR) : 1000 euros

Achat d'un serveur : 3000 euros
Le corpus produit sera installé sur un serveur particulier. L'installation et le maintien du serveur seront assurés par le service informatique de l'ENS-LSH.

Achat d'un PC pour le développeur : 1200 euros

Achat d'un ordinateur portable pour le CDD chargé de la mise à jour des descripteurs et qui sera amené à se déplacer dans diverses bibliothèques et institutions : 1200 euros

Fournitures (encre et papier essentiellement) : 1000 euros
Notre projet implique l'utilisation de quantités importantes de papier, la relecture des textes se faisant à la fois sur une version papier et sur la version numérique.

Achat d'une imprimante : 400 euros
Pour les raisons que nous venons d'évoquer, et compte-tenu de la sur-utilisation du matériel existant, nous aurons besoin de nous équiper d'une nouvelle imprimante.

Organisation du colloque : 5 000 euros
Le budget global du colloque est évalué à 15 000 euros. Les frais de déplacement des membres du projet participant au colloque seront assurés par ailleurs (rubrique frais de mission). Une participation sera également demandée au CNRS, à l'ENS, au Ministère des Affaires étrangères ainsi qu'au Ministère de la Culture et de la Communication (DGLFLF).

Personnels

Numérisation, relecture et balisage des textes
D'après l'expérience acquise, le coût de la numérisation, de la relecture et du balisage d'une page de texte s'élève, tout compris, à 7,20 euros environ. C'est sur cette base que nous avons établi le coût total de l'opération :
- 25 000 mots pour le premier tiers du XIIe siècle = env. 150 p. = 1 000 euros
- 300 000 mots pour le second tiers du XIIe siècle = env. 18 000 p. = 13 500 euros
- 200 000 mots pour le dernier tiers du XIIe siècle = env. 1200 p. = 9 000 euros

Total 23 500 euros
Ce total correspond à un CDD de 10 mois (2500 euros par mois, doctorant).

Mise à jour de la base de descripteurs : un CDD de deux mois (doctorant) : 5 000 euros.

Conception informatique de l'outil de gestion et d'exploitation de la base de descripteurs : 1 mois de CDD = 3000 euros.

Développement informatique de l’outil de gestion et d’exploitation de la base de descripteurs : 6 mois de CDD (cela comprend le développement et la recette, c’est à dire la livraison de l’outil, sa validation/test et la formation des utilisateurs) = 3000 euros x 6 = 18 000 euros.

Enrichissement linguistique (annotation syntaxique manuelle et automatique) : CDD de 12 mois : 35 000 euros.

Partenaire 2

Partenaire …

Propositions d'experts et confidentialité

Les membres du comité d'évaluation et du comité de pilotage sont astreints à la confidentialité.

- Possibilité de fournir une liste de 3 à 5 noms d'experts français ou étrangers (avec coordonnées complètes : adresse postale et adresse électronique) susceptibles d'évaluer le projet avec lesquels les équipes participant au projet n'ont ni conflit d'intérêt, ni collaborations en cours.
- Possibilité éventuelle de fournir une liste de 5 noms max. d'experts auxquels les participants au projet ne souhaitent pas que le projet soit envoyé s'il y a risque de conflits d'intérêts.

Experts proposés

NOM Prenom	Discipline	Adresse professionnelle / Etablissement	Tel professionnel	Mail
COMBETTES Bernard	Sciences du langage	Professeur à l'Université Nancy 2 UNIVERSITÉ NANCY 2 3 place Godefroy de Bouillon B.P. 3397 F 54015 NANCY CEDEX	03 83 96 70 68	Bernard.Combettes@univ-nancy2.fr
GALDERISI Claudio	Philologie romane	CESCM Professeur à l'Université de Poitiers 24, rue de la Chaîne BP 603 F 86 022 Poitiers cedex	01 40 15 08 63	ClaudioGalderisi@aol.com
MELLET Sylvie	Sciences du langage	Directrice du laboratoire « Bases, Corpus et Langage » UMR 6039 Université de Nice - Sophia Antipolis Faculté des Lettres 98, Bd Edouard Herriot - BP 3209 06204 Nice Cedex 3	04-93-37-53-16	mellet@unice.fr
HAVU Jukka	Sciences du langage	Professeur à l'Université de Tempere University of Tampere, FIN-33014 Tampere Finlande	358 3 35516140	jukka.havu@uta.fi
RODRIGUEZ SOMOLINOS Amalia	Sciences du langage	Professeur à l'Université de Madrid Universidad Complutense de Madrid Facultad de Filología Departamento de Filología Francesa Ciudad Universitaria 28040 - MADRID	913945494	arsomol@filol.ucm.es

Experts non souhaités

NOM Prenom	**Adresse professionnelle**

Programmes SHS 2007

C - Moyens financiers et humains demandés par chaque équipe partenaire du projet

Chaque équipe partenaire remplira une fiche de demande d'aide selon les modèles proposés ci-dessous (laboratoire public ou fondation ; entreprise ou association) en fonction de son appartenance.

Programmes SHS 2007

Fiche de demande d'aide
Laboratoire public / Fondation

Acronyme ou titre court du projet
CORPTEF

Partenaire 1 - Coordinateur (nom, prénom) : Guillot, Céline

Calcul de l'aide demandée à l'ANR et estimation du coût complet du projet pour le laboratoire du partenaire

Avant de remplir ce tableau il vous faut décider quel sera votre établissement gestionnaire (cf notes 4 et 5 en bas de page)

				Euros HT	Taux spécifiques à chaque établissement	
	Nbre Homme. mois	*Coût Homme.mois (salaire chargé)*	*Nombre de personnes impliquées*			
Dépenses de personnel(1) Professeur Chercheur/enseignant chercheur Ingénieur Postdoc etc.	 15,3 19,8 2,7 4,5	 112 409 95 772 13162 14850	 5 5 1 1	(P1) 236 193 (avec taxes)	Taux Env	425 147
Dépenses de personnel non permanent à financer par l'ANR(2) Doctorant Ingénieur etc.	 12 19		 2 3	(Q1) 84 500	80%	152 100
Equipements (>4000 €) détail § B-5				(R1)	Taux TVA non réc.	0
Petits matériels, consommables, fonctionnement, etc				(S1)	Taux TVA non réc.	13 800
Frais de missions si montant > 5% de la somme demandée, justification § B-5				(T1)	Taux TVA non réc.	29 500
Prestations de service externes, sous-contractant (3)				(U1)	Taux TVA non réc.	0
Total des dépenses de fonctionnement				(X1)= S1+T1+U1		43 300
Frais généraux (assistance, encadrement, coût de structure) (max 4 % du coût total des dépenses)						1732
Coûts éligibles à l'aide ANR						129 132
Aide demandée **≤ Z** (4)						129 132

Coût complet du projet (5)		**622 679**

(1) Il s'agit du personnel qui serait affecté au projet mais qui est présent dans le laboratoire ou l'entreprise indépendamment de la réussite de l'appel de l'agence. Salaire mensuel chargé (charges salariales et patronales). Pour les enseignants-chercheurs ne compter que la part salariale correspondant à la part recherche (Pour un enseignant chercheur consacrant tout son temps de recherche au projet, on comptera 50% du salaire).

5 grandes catégories (CDD ou CDI) : Ingénieur, chercheur, enseignant chercheur, technicien, autres. Lorsque dans une même catégorie, plusieurs personnes de salaire différent sont mentionnées indiquer la valeur moyenne. Pour les laboratoires publics ou fondations, ces données ne servent qu'à calculer le coût complet du projet.

(2) Personnel non statutaire directement affecté au projet exprimé en hommes mois. Les dépenses éligibles se limitent aux salaires et aux charges sociales. Exemple : post-doc, ingénieur d'études etc.

(3) Propriété intellectuelle, location de matériel, service, etc. Le total des dépenses de prestation de service doit être ≤ 50% du total des coûts éligibles à l'aide de l'ANR, sauf dérogation accordée par le Directeur de l'Agence, sur demande motivée du bénéficiaire.

(4) L'aide demandée doit correspondre au montant HT augmenté éventuellement de la TVA non récupérable. La TVA non récupérable est actuellement, par exemple, de 88% pour le CNRS et l'INRA, de 94% pour l'Inserm et de 100% pour les universités. En conséquence pour une demande qui sera gérée par l'INRA, le taux de TVA non récupérable est 0,88x0,196=0,1725, ce qui conduit à inscrire dans la colonne de droite pour une demande HT de 10 000 euros, 10000x(1+0,1725) soit 11 725 euros soit une demande d'aide de 11 725 euros si le partenaire veut disposer de 10 000 euros dans la réalisation de son projet.

En cas d'aide accordée par un autre financeur sur les mêmes dépenses que celles listées dans le tableau, il peut y avoir une diminution de l'aide accordée par l'ANR pour rester conforme à la réglementation.

(5) Pour le calcul en coût complet, il faut augmenter le salaire chargé d'un taux d'environnement, qui tient compte des conditions d'environnement des personnels (infrastructure, par exemple). Par exemple, ce taux est à l'heure actuelle de 1,8 pour l'Inserm et le CNRS.

Programmes SHS 2007

Fiche de demande d'aide
Entreprise / Association

Acronyme ou titre court du projet

Partenaire n°	Responsabl e scientifique (nom, prénom) :

Calcul de l'aide demandée à l'ANR et estimation du coût complet du projet pour le partenaire :

				Euros HT
	Nbre Homme. mois	*Coût Homme. mois Salaire chargé*	*Nombre de personnes impliquées*	
Dépenses de personnel (1) Ingénieur Chercheur etc.				(P)
Dépenses de personnel non permanent à recruter pour le projet (2) Ingénieur Chercheur etc.				(Q)
Amortissements des équipements (>4000 €) Nature et justification de la dépense				(R)
Petits matériels, consommables, fonctionnement, etc.				(S)
Frais de missions si montant >5% de la somme demandée, justification de la dépense				(T)
Prestations de service externes(3)**, sous-contractant**				(U)
Prestation de service interne à l'entreprise ou à l'organisme				(V)
Total frais fonctionnement				(x) =S+T+U+V
Frais généraux (assistance, encadrement, coût de structure) (4)				(Y)
Coût complet du projet			**CC= P+Q+R+X+Y**	
Taux de l'aide (Voir texte de l'AAP page 9)				**%**
Aide demandée (5) Se référer au texte de l'AAP				**Aide demandée**

(1) Il s'agit du personnel qui serait affecté au projet mais qui est présent dans le laboratoire ou l'entreprise indépendamment de la réussite de l'appel de l'agence. Salaire mensuel chargé (charges salariales et patronales). Pour les enseignants-chercheurs ne compter que la part salariale correspondant à la part recherche (50% du salaire pour 100% de temps consacré à la recherche).
5 grandes catégories (CDD ou CDI) : Ingénieur, chercheur, enseignant chercheur, technicien, autres. Lorsque dans une même catégorie plusieurs personnes de salaire différent sont mentionnées indiquer la valeur moyenne. Pour les laboratoires publics ou fondation, ces données ne servent qu'à calculer le coût complet du projet.

(2) Personnel non statutaire directement affecté au projet exprimé en hommes mois. Les dépenses éligibles se limitent aux salaires et aux charges sociales.

(3) Propriété intellectuelle, location de matériel, service, etc. Le total des dépenses de prestation de service doit être ≤ 50% du total des coûts éligibles à l'aide de l'ANR, sauf dérogation accordée par le Directeur de l'Agence, sur demande motivée du bénéficiaire.

(4) Pour les associations et TPE, les frais généraux peuvent être au maximum = 4% de R + 8% de (P+Q+S+T+U). Pour les sociétés civiles, les entreprises hors TPE, les GIE, les centres techniques, les frais généraux peuvent être au maximum de = 7% de (R+S+T+U) + 68% de (P+Q)

(5) En cas d'aide accordée par un autre financeur sur les mêmes dépenses que celles listées dans le tableau, il peut y avoir une diminution de l'aide accordée par l'ANR pour rester conforme à la règlementation .

Programmes SHS 2007

D - Récapitulatif global de la demande financière pour le projet

Acronyme ou titre court du projet : CORPTEF

a-Estimation du coût complet de cette demande

*(reporter les valeurs (**CC**) des fiches des différents partenaires)*

	Coût complet
Coordinateur (Partenaire 1)	**622 679**
Partenaire 2	
Partenaire 3	
Partenaire 4	
Partenaire 5	
Total	**622 679**

b-Total de l'aide demandée

*(reporter les valeurs (**Aide demandée**) des fiches des différents partenaires)*

	Aide demandée
Coordinateur (Partenaire 1)	**129 132**
Partenaire 2	
Partenaire 3	
Partenaire 4	
Partenaire 5	
Total	**129 132**

c- Effort en personnel demandé

(reporter les valeurs des fiches des différents partenaires)

	en homme/mois (≤ 72 mois)
Coordinateur (Partenaire 1)	**31**
Partenaire 2	
Partenaire 3	
Partenaire 4	
Partenaire 5	
Total	**31**

d- Dépenses de fonctionnement

(reporter les valeurs des fiches des différents partenaires)

	en euros
Coordinateur (Partenaire 1)	**43 300**
Partenaire 2	
Partenaire 3	
Partenaire 4	
Partenaire 5	
Total	**43 300**

Contrats publics et privés sur les trois dernières années (effectués et en cours)

Nom du membre participant à cette demande	% d'impli-cation	Intitulé de l'appel à projets Source de financement Montant attribué	Titre du projet	Nom du coordinateur	Date début - Date fin
Guillot Céline	50	ANR « Corpus », 180 000 euros ILF 3500 euros	Evolution des déterminants et corpus Evolution des démonstratifs en français	Lucia Tovena Céline Guillot	2006-2008 2006
Carlier Anne	80	ANR « Corpus », 180 000 euros ILF 3500 euros	Evolution des déterminants et corpus Evolution des démonstratifs en français	Lucia Tovena Céline Guillot	2006-2008 2006
Lavrentiev Alexei	25	ANR « Corpus », 180 000 euros	Evolution des déterminants et corpus	Lucia Tovena	2006-2008
Pignatelli, Cinzia	40 15	ANR « Corpus »	Transmedie Biblifre	Claudio Galderisi	
Prévost, Sophie	50 10	Institut de Linguistique Française 8000 euros (sur 3 ans) ANR « projets blancs » 70000 euros (sur 3 ans)	Grammaticalisation et lexicalisation : changements de classes de mots Spatial Framing Adverbials : linguistic and psycholinguistic approaches	Sophie Prévost Michel Charolles	2005- 2007 2006-2008

Demandes de contrats en cours d'évaluation [4]

Nom du membre participant à cette demande	% d'impli-cation	Intitulé de l'appel à projets Source de financement Montant demandé	Titre du projet	Nom du coordinateur
Guillot Céline	10	Cluster 13 (Région Rhône-Alpes)	Corpus représentatif des premiers textes français	Céline Guillot
Alexei Lavrentiev	20	Cluster 13 (Région Rhône-Alpes)	Corpus représentatif des premiers textes français	Céline Guillot
David Trotter	15	Cluster 13 (Région Rhône-Alpes)	Corpus représentatif des premiers textes français	Céline Guillot
Sophie Prévost	10	ANR Corpus	RHAPSODIE	Anne Lacheret

[4] *Mentionner ici les projets en cours d'évaluation soit au sein de programmes de l'ANR, soit auprès d'organisme de fondations, de l'Union européenne, etc. que ce soit comme coordinateur ou comme partenaire. Pour chacun donnez le nom de l'appel à projets, le titre du projet et le nom du coordinateur.*