Projet de consortium linguistique « Corpus Écrits »
Infrastructure de Recherche CORPUS

DATE REÇU	DATE EVALUÉ	DATE CRÉE

Sommaire :

1 IDENTIFICATION DU CONSORTIUM

1.1 NOM DU CONSORTIUM PROPOSÉ

« Corpus Écrits »

1.2 DISCIPLINES OU CHAMPS THÉMATIQUES PRINCIPAUX CONCERNÉS

Linguistique (section 34 du CNRS et section 07 du CNU)

1.3 PERSONNE À CONTACTER (NOM, MAILS, TELS...)

Franck Neveu, directeur de l'Institut de Linguistique Française (CNRS, FR2393), porteur du consortium.
Franck.neveu@ling.cnrs.fr
Téléphone : 01 43 13 56 45

1.4 EQUIPES MEMBRES DU CONSORTIUM (COMITÉ DE PILOTAGE OU ÉQUIPES/UNITÉS ACTIVES AU DÉMARRAGE)

	Code unité	Acronyme	Nom complet	Tutelles	URL	Contact (nom et email)	Email et tel
1	**FR 2393**	**ILF**	**Institut de Linguistique Française**	**CNRS & Université Paris Ouest, Nanterre la Défense**	**http://www.ilf.cnrs.fr/**	**Franck Neveu Franck.neveu@ling.cnrs.fr**	**01 43 13 56 45**
2	UMR 7118	ATILF (et CNRTL)	Analyse et Traitement Informatique de la Langue Française et Centre National de Ressources Lexicales et Textuelles	CNRS et Nancy Université	www.atilf.fr www.cnrtl.fr	Jean-Marie Pierrel Jean-Marie.Pierrel@atilf.fr	03 54 50 52 85
3	UMR 7597	HTL	Histoire des Théories Linguistiques	CNRS & Univ. Paris Diderot, Univ. Paris 3	http://http://htl.linguist.univ-paris-diderot.fr/index.htm	Sylvie Archaimbault sylvie.archaimbault@linguist.jussieu.fr	01 57 27 57 58
4	UMR 7187	LDI	Lexiques, Dictionnaires, Informatique	CNRS & Univ. Paris 13, Univ. de Cergy-Pontoise	http://www-ldi.univ-paris13.fr/	Céline Poudat celine@poudat.fr	01 49 40 38 57
5	UMR 7110	LLF	Laboratoire de linguistique formelle	CNRS & Université Paris Diderot	http://www.llf.cnrs.fr	Anne Abeillé abeille@linguist.jussieu.fr Clément Plancq clement.plancq@linguist.jussieu.fr	01 57 27 57 64

6	EA609	LIDILEM	Laboratoire de Linguistique et de Didactique du Français Langue Etrangère et Maternelle	Université Grenoble 3	http://www.u-grenoble3.fr/LIDILEM/	Agnès Tutin agnes.tutin@u-grenoble3.fr Marie-Paule Jacques marie-paule.jacques@ujf-grenoble.fr	04 76 82 43 68
7	UMR 5191	ICAR	Interactions, Corpus, Apprentissages, Représentations	CNRS & Univ. Lyon 2 & ENS de Lyon	http://icar.univ-lyon2.fr	Serge Heiden slh@ens-lyon.fr	04 37 37 63 12
8	EA 1339	LiLPa	Linguistique, Langues et Parole	Université de Strasbourg	http://lilpa.u-strasbg.fr/	Catherine Schnedecker cschnede@unistra.fr Amalia Todirascu todiras@unistra.fr	03 68 85 67 85
9	UMR-I 001	ALPAGE	Analyse Linguistique Profonde A Grande Echelle	INRIA & Université Paris Diderot	https://www-roc.inria.fr/alpage-wiki/tiki-index.php?page=accueil	Pascal Denis pascal.denis@inria.fr	01 57 27 57 66
10	UMR 6039	BCL	Bases, Corpus, Langage	CNRS & Université de Nice Sophia Antipolis	http://www.unice.fr/bcl/	Damon Mayaffre mayaffre@unice.fr	04 89 88 14 46

2 PRÉSENTATION GÉNÉRALE DU CONSORTIUM

2.1 CONTEXTE GÉNÉRAL DU CONSORTIUM PROPOSÉ

La création de l'Infrastructure de Recherche CORPUS (Coopération des Opérateurs de Recherche Pour un Usage des Sources numériques) a ouvert la possibilité de constituer un consortium linguistique spécialement dédié aux Corpus écrits. Partant du constat que la création de corpus numériques écrits à des fins de recherche a connu des développements considérables dans les dernières années et que des attentes fortes se font jour en matière de partage d'information, d'homogénéisation des pratiques et de mise en conformité avec des standards internationaux, le **consortium « Corpus écrits »** se donne précisément pour but de fédérer les équipes et laboratoires, les chercheurs, enseignants-chercheurs, ou ingénieurs engagés dans la production de corpus numériques écrits, quels que soient la langue et l'alphabet considérés et d'offrir la représentation la plus large possible de cette communauté, afin d'accompagner le développement des corpus écrits, d'en faire converger les pratiques et les besoins, de financer des actions répondant à ses missions.

Dimension nationale

Nous évaluons à une cinquantaine le nombre d'entités de recherche en linguistique (laboratoires CNRS et équipes universitaires) sur le territoire national, et à une trentaine les entités potentiellement concernées par le thème des corpus écrits.

Le porteur du consortium, l'ILF est une structure fédérative du CNRS qui par son statut même est en interaction permanente avec nombre de ces entités dans toute la France.

En outre, l'ATILF/CNRTL, très impliqué dans les projets du consortium, apporte à travers sa participation une plateforme commune du réseau national des MSH pour les aspects textes, lexiques et dictionnaires

Dimension internationale

Le comité de pilotage du consortium « Corpus écrits » a identifié comme un objectif hautement prioritaire l'insertion des chercheurs français dans les réseaux européens. Parmi les laboratoires représentés au comité de pilotage du consortium, certains ont déjà et depuis longtemps pris des engagements et établi des collaborations allant dans ce sens. L'ATILF, notamment, est engagé à travers

- sa participation au projet européen CLARIN d'infrastructure européenne partagée pour les SHS (Common Language Resources and Technology Infrastructure : www.clarin.eu) s'appuyant sur des centres régionaux « certifiés » dans leurs domaines respectifs. La mise en place d'un accord de partenariat entre l'ATILF et l'INIST (février 2009) a permis de positionner le CNRTL comme un des principaux centres de cette future infrastructure de recherche.

- sa fonction de centre européen support de la TEI : issue d'un partenariat entre l'ATILF, l'INIST et le LORIA, cette fonction de centre européen support de la TEI, assurée jusqu'alors par le CNRTL, est reprise pour l'avenir directement par le TGE ADONIS.

- Des collaborations directes avec des centres partenaires, en Grande-Bretagne (Université d'Oxford), en Allemagne (Centres de compétence de Trèves et de Würzburg, Université de Sarrebruck, MPI Berlin), aux Pays Bas (Université de Nimègue) et en Angleterre (Oxford).

Le comité de pilotage du consortium dépose le présent projet en vue de sa labellisation pour une durée de quatre ans.

2.2 LES MISSIONS ET OBJECTIFS QUE SE DONNE LE CONSORTIUM

Répondant aux missions générales fixées par l'IR Corpus, le consortium Corpus écrits s'accordera sur un programme pluriannuel (4 années) de numérisation et de documentation comprenant toutes les étapes de l'archivage et de la mise à disposition numérique, y compris la discussion sur les formats et standards à adopter.

Le consortium vise à répondre aux besoins d'information, de formation, de partage des bonnes pratiques et de diffusion des standards européens et internationaux en matière de création, production, développement et valorisation des corpus numériques écrits.

Dans la limite des moyens qui sont les siens, il financera des projets et actions qui lui auront été soumis et qui entreront en cohérence avec ses missions.

2.3 FONCTIONNEMENT DU CONSORTIUM

Membres du consortium

Peut être membre du consortium toute personne, dès lors qu'elle appartient à la communauté académique, qu'elle est engagée dans la production de corpus numériques écrits développés à des fins de recherche et qu'elle déclare son intérêt pour collaborer aux activités du consortium. Ainsi, la participation aux groupes de travail, aux réunions de formation et d'information sera très largement ouverte. Dans la mesure du possible le consortium pourra prendre en charge les frais de missions pour la participation aux groupes de travail et aux réunions générales du consortium à raison d'un représentant par laboratoire

Comité de pilotage

Le consortium se dote d'un comité de pilotage, constitué de 10 personnes, représentatif de la communauté scientifique concernée. Chaque équipe ou laboratoire désigne son représentant pour le comité de pilotage. Le comité de pilotage est constitué pour 4 ans. Il émet des propositions d'actions, sélectionne parmi les projets soumis au consortium et attribue les crédits.

Membres du Comité de pilotage

Sont désignés par leur laboratoire comme membres du Comité de pilotage :

Franck Neveu pour l'ILF, FR 2393
Jean-Marie Pierrel pour l'ATILF - UMR 7118 – Nancy - Université
Sylvie Archaimbault (suppléant **Bernard Colombat)** pour HTL – UMR 7597 - Université Denis Diderot - Paris 7

Damon Mayaffre (Suppléante **Mahé Ben Hamed**) pour BCL - UMR 6039 - Université Nice Sophia Antipolis
Serge Heiden pour ICAR - UMR 5191 - Université Lumière Lyon 2
Clément Plancq (suppléant **Olivier Bonami**) pour le LLF - UMR 7110 - Université Paris 7
Céline Poudat pour le LDI - UMR 7187 – Université de Paris 13
Catherine Schnedecker (suppléante **Amalia Todirascu)** pour LILPA – EA 1339 – Université de Strasbourg
Agnès Tutin (suppléante **Marie- Paule Jacques**) pour le LIDILEM – EA 609 – Université Grenoble 3
Pascal Denis pour ALPAGE – INRIA- Université Denis Diderot - Paris7

Porteur

Le porteur du consortium Corpus écrits est la fédération de recherche ILF - Institut de Linguistique Française (FR 2393 du CNRS2393), représentée par son directeur, Franck Neveu.

Ses missions sont les suivantes :

- Le porteur anime le consortium et veille à la vitalité et au bon fonctionnement des groupes de travail
- Il met en œuvre les recommandations et décisions du comité de pilotage
- Il reçoit et gère la dotation (de l'ordre de 50 000 à 60 000 € pour l'exercice 2011)
- Il est tenu de justifier le détail des dépenses pour l'exercice 2011 et s'engage à présenter un budget prévisionnel pour l'exercice 2012

2.4 MÉTHODOLOGIE

Décrire les méthodologies scientifiques et techniques existantes et à développer.

Voir fiches détaillées (section 4)

2.5 PERSONNEL AFFECTÉ AU CONSORTIUM ET SES PROJETS

Faire la liste des personnes (un tableau par partenaire) qui dans les quatre années s'engagent à participer pour une part de leur temps aux activités qui sont directement ou indirectement associées au consortium et à ses projets. Ne dressez la liste que des personnels existants. Distinguez les « permanents » (titulaires) des « temporaires » (vacations et CLD/CDD prévus dont le financement est acquis).

*Ces tableaux doivent permettre à l'IR Corpus de calculer les ressources humaines que les partenaires et leurs tutelles investissent dans le consortium**. Les chiffres pourront être ajustés annuellement**.*

2.5.1 PARTENAIRE 1 : ILF

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
NEVEU, Franck	Université Paris - Sorbonne	porteur	PR	25%	25%	25%	25%
BRISSET-FONTANA, Véronique	CNRS	Secrétaire Générale	IE	25%	25%	25%	25%
ZENDAGUI Zahia	CNRS	gestionnaire	TCN	25%	25%	25%	25%

Temporaires

Nom personnel (si connu)	Financeur du personnel	Fonction/tâche dans le cadre du consortium	Type de contrat : Vacataire, CLD…	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
GIRAULT, Stéphanie	CNRS	Ingénieur de Recherche (Linguistique & Informatique) – soutien au pilotage	CDD sur contrat DGLFLF	25%	-	-	-
À déterminer	CNRS	Ingénieur de Recherche	À déterminer	-	25%	25%	25%

2.5.2 PARTENAIRE 2 : ATILF

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
BAERMAN Michèle	CNRS ATILF	validation de ressources, balisage et codage XML/TEI	TCS	100	100	100	100
BENOIT Jean-Luc	CNRS ATILF	recommandation et normes et codage de ressources textuelles (TEI)	IE1	50	50	50	50
CLEMENT Isabelle	CNRS ATILF	balisage et codage XML/TEI	TCN	100	100	100	100
PERIGNON Jessika	CNRS ATILF	balisage et codage XML/TEI	TCN	100	100	100	100
MONTEMONT Véronique Zahia	UHP ATILF	base textuelles Frantext	TCN	25	25	25	25
PIERREL Jean-Marie	UHP ATILF	responsable du CNRTL	PR0	15	15	15	15
PETITJEAN Etienne	CNRS ATILF	responsable technique informatique	IR2	50	50	50	50
GAIFFE Bertrand	CNRS ATILF	Développement informatique et codage de corpus	IR1	100	100	100	100
SOUVAY Gilles	CNRS ATILF	Développement informatique et codage de corpus (Moyen français)	IR2	75	75	75	75
BAERMAN Michèle	CNRS ATILF	validation de ressources, balisage et codage XML/TEI	TCS	100	100	100	100

Temporaires

Nom personnel (si connu)	Financeur du personnel	Fonction/tâche dans le cadre du consortium	Type de contrat : Vacataire, CLD…	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Catherine CHEVALIER	CNRS ATILF	Numérisation et OCRisation	CDD AI sur Ressouces Frantext	100	A déterminer	A déterminer	A déterminer
Cyril PESTEL	CNRS ATILF	Ingénieur de développement informatique	CDD sur contrat	100	A déterminer	A déterminer	A déterminer
Benjamin HUSSON	CNRS ATILF	Ingénieur de développement informatique	CDD sur contrat	100	A déterminer	A déterminer	A déterminer

2.5.3 PARTENAIRE 3 : HTL

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Archaimbault Sylvie	CNRS	CP, GT Alphabets non latins, Droits des auteurs et éditeurs	DR 2	10%	10%	10%	10%
Colombat Bernard	Univ. Paris Diderot	CP, GT Etats anciens de la langue, Droits des auteurs et éditeurs	PU 1	10%	10%	10%	10%
Garcea Alessandro	Univ. Lyon 2	GT Etats anciens de la langue	PU 2	10%	10%	10%	10%
Lahaussois Aimée	CNRS	GT	IR 2	10%	10%	10%	10%
Lazcano Elisabeth	CNRS	Catalogage Normes métaD.	IEHC	10%	10%	10%	10%
Plancq Clément	CNRS	Encodage Dév.logiciel Ingénierie linguistique	IE 2	7%	7%	7%	7%

2.5.4 PARTENAIRE 4 : LDI

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Poudat Céline	Univ. Paris 13	CP, GT Traitements automatiques de corpus et outils de traitement de corpus, Usage des Corpus et droits d'auteurs ou d'éditeurs	MCF	10%	10%	10%	10%
Cartier Emmanuel	Univ. Paris 13	GT Traitements automatiques de corpus et outils de traitement de corpus, Codage de métadonnées	MCF	10%	10%	10%	10%
Issac Fabrice	Univ. Paris 13	GT Traitements automatiques de corpus et outils de traitement de corpus, Corpus en alphabets non latins et pluralité	MCF	10%	10%	10%	10%

		d'alphabets (numérisation, codage)					
Barque Lucie	Univ. Paris 13	GT Codage de données primaires, Annotation de haut niveau	MCF	10%	10%	10%	10%
Manuelian Hélène	Univ. Cergy-Pontoise	GT Corpus d'états anciens de la langue, codage de données primaires	MCF	7%	7%	7%	7%
Salvador Xavier-Laurent	Univ. Paris 13	GT Corpus d'états anciens de la langue	MCF	5%	5%	5%	5%
Grezska Aude	CNRS	GT Codage de données primaires	IR	5%	5%	5%	5%

2.5.5 PARTENAIRE 5 : LLF

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Abeillé Anne	Univ. Paris Diderot	Annotation de haut niveau, corpus arborés	PR	10%	10%	10%	10%
Plancq Clément	CNRS	Comité de pilotage, encodage, devpt logiciel	IE	10%	10%	10%	10%
Bonami Olivier	Univ. Paris Sorbonne	Comité de pilotage, BD lexicales, annotation de haut niveau	MCF	5%	5%	5%	5%
Crysmann Berthold	CNRS	Encodage, annotation de haut niveau, traitements automatiques	CR	10%	10%	10%	10%
Fon Sing Guillaume	Univ. Paris Diderot	Encodage, annotation de haut niveau	MCF	10%	10%	10%	10%

Temporaires

Nom personnel (si connu)	Financeur du personnel	Fonction/tâche dans le cadre du consortium	Type de contrat : Vacataire, CLD…	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Henri Fabiola	Univ. Lille	Encodage, annotation de haut niveau	ATER	10%	10%	10%	10%

2.5.6 PARTENAIRE 6 : LIDILEM

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
TUTIN Agnès	Université Stendhal Grenoble 3	Participation au comité de pilotage Formats d'encodage, outils d'exploitation de corpus	MCF	Pourcentage indicatif 10%	Pourcentage indicatif 10%	Pourcentage indicatif 10%	Pourcentage indicatif 10%
JACQUES Marie-Paule	Université Joseph Fourier	Participation au comité de pilotage Formats d'encodage, outils d'exploitation de corpus	MCF	Pourcentage indicatif 10%	Pourcentage indicatif 10%	Pourcentage indicatif 10%	Pourcentage indicatif 10%

Temporaires

Nom personnel (si connu)	Financeur du personnel	Fonction/tâche dans le cadre du consortium	Type de contrat : Vacataire, CLD…	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Falaise Achille	LIDILEM	Formation aux outils d'exploitation de corpus (syntaxe, structure textuelle)	Vacataire	Pourcentage indicatif 10%	Pourcentage indicatif 10%	-	-

2.5.7 PARTENAIRE 7 : ICAR

Permanents

Nom personnel	Tutelle du personnel	Fonction/tâche dans le cadre du consortium	Corps/grade	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Heiden Serge	ENS de Lyon	*Participation* au comité de pilotage *Participation Action 1* catalogage *Action 2 :* coordination de création de nouveaux corpus et annotations *Action 3 :* formation en catalogage et codage, en préparation et exploitation de corpus enrichis linguistiquement *Action 4 :* CLARIN et liaison avec les	IR2	50%	50%	50%	50%

		consortiums TEI et CCFM *Action 5 :* GT outils de TAL et outils d'exploitation de corpus (développement et diffusion de la plateforme d'exploitation et de diffusion de corpus TXM), GT corpus ancien					
Pincemin Bénédicte	CNRS ICAR	*Action 3 :* formation en exploitation de corpus enrichis linguistiquement - action 5 : GT outils de TAL et outils d'exploitation de corpus	CR	25%	25%	25%	25%
Lavrentiev Alexei	CNRS ICAR	*Action 3 :* formation en catalogage et codage, en préparation et exploitation de corpus enrichis linguistiquement - action 5 : GT corpus ancie	IR2	50%	50%	50%	50%
Guillot Céline	ENS de Lyon	*Action 3* : formation en exploitation de corpus enrichis linguistiquement *Action 5 :* GT corpus ancien développement et diffusion de la Base de Français Médiéval (BFM)	MCF	25%	25%	25%	25%

Temporaires

Nom personnel (si connu)	Financeur du personnel	Fonction/tâche dans le cadre du consortium	Type de contrat : Vacataire, CLD…	Estimation en % du temps dédié aux projets du consortium			
				Année 1	Année 2	Année 3	Année 4
Decorde, Matthieu	ENS de Lyon	*Action 5 :* GT outils de TAL et outils d'exploitation de corpus (développement de la plateforme TXM)	CDD sur contrat ANR Corptef	25%	25%	25%	25%
Rainsford, Thomas	ENS de Lyon	*Action 5 :* GT corpus ancien (développement du	CDD sur contrat ANR-DFG Srcmf	25%	A déterminer	A déterminer	A déterminer

		corpus BFM - annotation syntaxique)					

3 PROJETS ENVISAGÉS

3.1 ACTIONS PRIORITAIRES PLANIFIÉES POUR 2011-2012

Pour répondre à ses missions générales, le consortium se fixe une série de tâches prioritaires, déclinées comme suit :

ACTION 1

Dresser un état des lieux, large et fiable, des corpus existants, accessible à tous à travers des métadonnées normalisées et, au vu de ce bilan et des besoins de la communauté scientifique, proposer des priorités pour compléter l'existant. Cette action devra être lancée dès l'automne 2011 et ses résultats accessibles dans le courant de 2012
Il s'agit d'identifier les corpus écrits existants ainsi que leur état d'achèvement au regard des normes et standards internationaux.
Les corpus et enrichissements devront être librement accessibles à la communauté, ou, à défaut, proposer un échantillon de la ressource décrire les restrictions d'accès et afficher les coordonnées d'une personne de référence à contacter, si les données ne sont pas libres de droit.

Mise en œuvre : ATILF-CNRTL et/ou ICAR (à déterminer)

Solution ATILF-CNRTL :

Le fait de prévoir un hébergement de ces métadonnées sur le site du CNRTL permettrait de donner une visibilité internationale au résultat de cette action. En effet le CNRTL, en tant que centre CLARIN, est moissonné par l'infrastructure de recherche européenne CLARIN.

Solution ICAR :

ICAR propose un catalogage collaboratif de type wiki.

Le catalogue pourra être compatible avec celui de CLARIN. Un système de mise en relation avec certaines informations serait alors être mis en place.

Le wiki de la liste 'corpus-ecrits' au CRU pourrait servir à spécifier cela (structure, workflow...).
(https://listes.cru.fr/sympa/sso_login/federation_cru/Shibboleth.sso/wayf?target=https://listes.cru.fr/wiki/corpus-ecrits)

ACTION 1.1 (en relation de dépendance avec l'action 1)

Solution ATILF-CNRTL :

Les équipes et laboratoires auront la responsabilité de renseigner directement et individuellement ce catalogue.

Clarin propose à cet effet un éditeur libre de droit. Le consortium organisera la formation à cet outil de saisie des métadonnées et mettra en place un support technique pour accompagner les laboratoires qui en feront la demande.

Le consortium veillera à ce que cette activité soit compatible avec celles équivalentes du consortium Open Language Archives Community (OLAC : http://www.language-archives.org)

Solution ICAR :

Tout membre du consortium pourrait ajouter un nouvel élément et sa description au wiki et l'édition/vérification de ces informations suivrait un principe collaboratif.

ACTION 1.2 (en relation de dépendance avec l'action n°1)

Le CP du consortium envisagera le recrutement en CDD d'un ingénieur chargé de vérifier l'intégrité des métadonnées compilées.
Cette tâche pourrait aussi être confiée au CNRTL.

En 2006 le CNRS a impulsé la création de centres de ressources (www.cnrs.fr/inshs/recherche/centres-ressources-numeriques.htm) dont le CNRTL (www.cnrtl.fr) pour les ressources textuelles, lexicales et dictionnairiques. Adossé à l'ATILF (Analyse et Traitement Informatique de la Langue Française, UMR CNRS - Nancy Université : www.atilf.fr), son objectif initial était de réunir, au sein d'un portail unique, le maximum de ressources informatisées et d'outils de traitement pour l'étude, la connaissance et la diffusion de la langue française.
Pour faciliter la mutualisation de telles ressources, nous avons choisi de doter ce centre d'une visibilité spécifique, au travers de l'acquisition des droits du domaine cnrtl.fr, offrant ainsi :

- une boîte contact : contact@cnrtl.fr, qui aujourd'hui recueille en moyenne une quinzaine de messages par semaine ;
- un site web : www.cnrtl.fr , vecteur principal pour présenter le CNRTL, diffuser les ressources et permettre aux utilisateurs de proposer ou déposer des ressources auprès du CNRTL.

Le CNRTL est soutenu par le TGE Adonis et, au cours des dernières années, pour assurer sa mise en place, il a été contractualisé dans le cadre du CPER Lorrain et a bénéficié d'un soutien du FEDER lorrain (Fonds Européen de Développement Economique des Régions). De plus le CNRTL est devenu en 2009 un centre CLARIN.
A noter que l'appui sur le CNRTL pourrait donner une remarquable visibilité à ces métadonnées, en effet le seul portail lexical du CNRTL fait l'objet chaque jour, hors périodes de vacances, de plus de 300 000 requêtes venant du monde entier (cf. www.cnrtl.fr/aide/stat/).

ACTION 2

Au cours des années 2012-2014, le consortium apportera son soutien technique et éventuellement financier pour

I. aider à la mise en forme (normes et standards) de corpus existants mais non mutualisables en raison de leur format spécifique.

II. aider à leur diffusion chez un partenaire ou dans une infrastructure

III. conseiller les producteurs de corpus quant aux aspects juridiques

ACTION 3

Assurer, de la façon la plus large possible, la formation et l'information des acteurs

a. Formation/information aux acquis du projet d'infrastructure européenne CLARIN (www.clarin.eu)

Organisateurs : CNRTL ATILF

Format : workshop
Date ou périodicité : Automne 2011
Intervenants : Membre du projet CLARIN (Français et Étrangers)
Nb de Participants : de 20 à 100
Public visé : ensemble de la communauté
Budget : 10 000 € par session

b. Formation aux normes et recommandations internationales pour les métadonnées permettant le catalogage des corpus

Organisateurs : ATILF CNRTL

Format : ateliers avec TP
Date ou périodicité : 2 ou 3 jours en automne 2011 et une seconde session au printemps 2012
Intervenants : Membre du projet CLARIN (Français et Étrangers)
Nb de Participants : 20 par sessions avec comme objectif de former un ou deux correspondants par laboratoire impliqué dans le consortium
Public visé : Enseignants-chercheurs, chercheurs, ITA et IATOS
Budget : 10 000 € par session

c. Formation aux normes et recommandations de codage et d'annotation de corpus

Organisateur : ICAR (sous réserve)

Format : workshop à Lyon
Date ou périodicité : 2 ou 3 jours en automne 2011 et une seconde session au printemps 2012
Intervenants : ICAR, ATILF, LDI, LLF, BCL, LIDILEM, LiLPa, ALPAGE
Nb de Participants : 20 par sessions avec comme objectif de former un ou deux correspondants par laboratoire impliqué dans le consortium
Public visé : Enseignants-chercheurs, chercheurs, ITA et IATOS
Budget : 10 000 € par session

d. Formation/information sur les aspects juridiques

Organisateur : HTL (sous réserve)

Format : journée à Paris
Date : printemps 2012
Intervenants : HTL, ATILF, LIDILEM, ICAR

Nb de Participants : 20 - 30
Public visé : Enseignants-chercheurs, chercheurs, ITA et IATOS
Budget : 5 000 €

e. Information sur les activités du consortium à travers l'organisation d'une réunion générale (1 ou 2 jours) annuelle de bilan d'action du consortium (bilan des actions lancées et des réflexions de groupes de travail) et de prospectives.

Pour cela, le consortium pourra s'appuyer sur les infrastructures de formation continue de ses partenaires (type école thématique CNRS ou formation continue)

Organisateur : ILF pour 2011

Format : 1 journée à Paris en 2011 puis 2 journées ou plus les années suivantes
Date : décembre 2011
Intervenants : Comité de pilotage
Nb de Participants : 60
Public visé : communauté
Budget : 10 000 € pour 2011

ACTION 4

Favoriser l'insertion des acteurs français dans des réseaux internationaux

Pour remédier à la sous-représentation française dans les séminaires et groupes de travail européens et internationaux, le consortium soutiendra la participation d'experts français dans ces initiatives en finançant par exemple leur participation à des groupes de travail à l'étranger.

Nbre d'experts mandatés : 5 missions en 2011, 10 missions en 2012 et les années suivantes
Budget : 5 000 € pour 2011 ; 10 000 € pour 2012

ACTION 5

Soutenir la mise en place de groupes de travail.

Le consortium préconise la mise en place de différents groupes de travail, dès l'automne 2011, pour répondre aux besoins les plus urgents :

a. Usage des Corpus et droits d'auteurs ou d'éditeurs (aspects juridiques)
b. Corpus d'états anciens de la langue (numérisation ; codage)
c. Corpus en alphabets non latins et pluralité d'alphabets (numérisation, codage)
d. Traitements automatiques de corpus et outils de traitement de corpus
e. Codage des métadonnées

Et prévoit également la création, au cours de l'année 2012, de groupes de travail sur les thèmes suivants, qui lui apparaissent d'intérêt général pour la communauté :

f. Corpus d'écrits modernes et prise en compte de nouveaux modes de communication (SMS, mail, blog, etc)

g. Encodage et annotation de plus haut niveau : syntaxe, sémantique, références (annotations collaboratives)
h. Enrichissement linguistique de corpus (segmentation lexicale, description morphosyntaxique et lemmatisation, …)

3.2 RÉCAPITULATIF POUR LE BUDGET PRÉVISIONNEL

Lignes budgétaires	2011	2012	Descriptif
Action 1-	15 000	-	Catalogage de l'ensemble des corpus écrits produits par la communauté scientifique ; Création d'une base de données accessible et interrogeable en ligne à travers un portail dédié.
Action 2	-	15 000	Mise aux normes des corpus existants
Action 3a	10 000	10 000	Organisation d'un workshop sur les acquis du projet CLARIN
Action 3b	10 000	20 000	Ateliers de formation aux normes internationales recommandées pour les métadonnées 2 sessions par an
Action 3c	10 000	20 000	Formation aux normes et recommandations de codage et d'annotation de corpus 2 sessions par an
Action 3d	-	5 000	Formation/information sur les aspects juridiques
Action 3e	10 000	20 000	Organisation d'une réunion générale annuelle d'information à la communauté (réunion sur 2 jours en 2012)
Action 4	5 000	10 000	Insertion des acteurs français dans les réseaux internationaux
Action 5	12 500	80 000	Organisation de groupes de travail (5 groupes en 2011 ; 8 en 2012)
Fonctionnement	3 000	6 000	Fonctionnement général (organisation des réunions du CP et frais divers)
Total	**75 500**	**186 000**	

4 DESCRIPTION DE CHAQUE PARTENAIRE ET DE LEURS FONDS D'ARCHIVES DÉJÀ DISPONIBLES EN NUMÉRIQUE, AINSI QUE LES MÉTHODES DE TRAVAIL

En une page, décrire chaque partenaire du consortium. A partir de la seconde page, décrivez dans le tableau les fonds qui sont déjà disponibles et qui ne font pas l'objet des travaux du consortium. Recopiez la section pour autant de partenaires que nécessaires.

4.1 PARTENAIRE 1 : INSTITUT DE LINGUISTIQUE FRANÇAISE (ILF)

Descriptions générales	
Numéro du partenaire	1. Porteur du consortium
Description sommaire du partenaire	Créé le 1er janvier 2001, l'Institut de Linguistique Française est une Fédération de recherche du CNRS : la FR 2393. Une fédération de recherche est une structure opérationnelle de recherche qui regroupe des unités et structures diverses relevant du CNRS ou d'autres organismes en vue de favoriser la coordination de leur activité scientifique et la mise en commun de tout ou parties de leurs moyens. Les entités qui participent à une telle structure conservent leur individualité propre. L'Institut de Linguistique Française regroupe la plupart des unités du CNRS et des équipes universitaires travaillant en linguistique française.
Problématiques principales de recherche du partenaire	La Fédération n'a pas vocation à proposer un programme de recherche qui lui est propre ; sa mission est de favoriser l'activité scientifique des unités de recherches qui la constituent
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	Projet de corpus de référence du français (en cours d'élaboration) Projet national interministériel
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant ou non	Scanner A4/A3 à chargement automatique (bizhub C353 Konika Minolta) Résolutions de numérisation supportées: 200 ppp, 300 ppp, 400 ppp, 600 ppp
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	Stéphanie GIRAULT (IR en CDD pour le partenariat avec la DGLFLF) pour l'interface avec les producteurs (information, appui technique), la mise aux normes des données (primaires et secondaires) et la gestion du portail Corpus de la Parole
Expérience en numérisation	OUI
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	Expérience et utilisation des normes préconisées par la TEI, métadonnées au format Dublin Core enrichi des extensions OLAC
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	Utilisation d'Omnipage avec entraînement et paramétrage du moteur de reconnaissance associé pour l'application à des langues peu décrites.
Expérience en correction d'épreuves et logiciels utilisés (correcteur	OUI

orthographique...)	
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	Unicode/UTF-8 XML
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	OUI
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	OUI
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	Portail Corpus de la Parole http://corpusdelaparole.in2p3.fr/
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	Pour la numérisation : jusqu'ici, par les laboratoires eux-mêmes ou par l'intermédiaire du LACITO Pour l'archivage et l'hébergement : TGE-ADONIS (CINES + IN2P3)
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	Pas de corpus écrits mais des corpus oraux en français et Langues de France, en partie transcrits (éventuellement traduits) et accessibles en ligne librement (sous licence Creative Commons) Volumétrie : 1To (1600 heures) http://corpusdelaparole.in2p3.fr/
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	-
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles	-

(pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	-
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	-

4.2 Partenaire 2 : ATILF UMR 7118 & CNRTL (Centre National de Ressources Textuelles et Lexicales)

Descriptions générales	
Numéro du partenaire	2
Description sommaire du partenaire	Laboratoire d'Analyse et Traitement Informatique de la Langue Française : son centre de gravité est constitué par le lexique du français, des langues romanes et d'autres langues (dans leur comparaison avec le français) sous toutes ses facettes (et appréhendé selon une multitude d'approches théoriques) : sémantique, morphologie et combinatoire lexicales ; lexicologie et lexicographie théoriques et pratiques, synchroniques et diachroniques, monolingues et plurilingues ; acquisition et apprentissage du lexique ; utilisation du lexique dans l'interaction verbale ; exploitation du lexique en TAL ; constitution et normalisation de ressources facilitant les études lexicales (corpus écrits et oraux, éditions de textes).
Problématiques principales de recherche du partenaire	- Linguistique historique française et romane - Lexique - Apprentissage et acquisition des langues - Discours - Ressources : normalisation, annotation et exploitation
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	Projets ANR Sourcencyme, DETCOL + 3 en cours de préparation Projet ERC « Histoire du vocabulaire politique » Projet PCRD - CLARIN (Common Language Ressources and Technology

	Infrastructure : infrastructure européenne de recherche pour les SHS www.clarin.eu - IMPACT IMProving ACcès to Textr www.impact-project.eu/ - Projet CPER lorrain « Langues, Textes et Documents et CNRTL
Moyens et plate-forme technologiques utilisés : LE CNRTL	
Matériel de numérisation existant ou non	Oui : scanner à plat, Scanner de type digibook, Scanner de numérisation de microfiches et microfilms
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	• M. Baerman (TCS à 100%) : pour des aspects de validation manuelle de ressources, balisage et codage XML/TEI • JL Benoit (IE2 à 50 %) : aspects recommandation et normes et codage de ressources textuelles (TEI) • Clément (TCN à 100%) : pour des aspects de validation manuelle de ressources, balisage et codage XML/TEI • B. Gaiffe (IR Informaticien, 100 %) Corpus et outils • Montémont (MCF IUF) : base textuelles Frantext • J. Pérignon : pour des aspects de validation manuelle de ressources, balisage et codage XML/TEI • JM Pierrel (PR, Informatique et linguistique, 15 %) : responsable du CNRTL • E. Petitjean (IR informaticien, 50 %) : développement informatique • G. Souvay (IR informaticien : 75%) Corpus et dictionnaire moyen français • C. Chevalier (CDD AI à 75 %) : pour les aspects de numérisation et OCRisation de corpus • B. Husson et C. Pestel CDD IE à100%) : développement d'outils informatique
Expérience en numérisation et standard ou normes utilisés pour les formats numériques (Tiff, Jpeg, Exif...)	OUI
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	METS, EAD, TEI, Dublin Core, et format de métadonnées CLARIN
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	Fine Reader, Omnipage
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	relecture et correction manuelle assistée par ordinateur
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	XML/TEI LMF Unicode METS
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	TEI, XCES, RDF...) outils TAL utilisés éventuels TreeTagger, Cordial.
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de	FLEMM, POMPAMO, DERIF diffusés avec une licence GPL LGeRM : Lemmatisation des mots en moyen français va être diffusé sous licence GPL + mise en place d'un web service (cela nécessite préalablement environ 6

diffusion des logiciels)	mois de travail)
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	Cf. CNRTL www.cnrtl.fr onglets “Corpus”, “Lexiques” et “dictionnaires” Diffusion de corpus avec une licence Créative Commons
Expériences d'externalisation	
Expériences d’externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	Oui, mais pas très concluante
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	- Frantext : par abonnement pour l’ensemble de la base (plus de 4000 Textes, 250 Millions de mots, en téléchargement libre au format XML/TEI pour les textes libres de droits, www.frantext.fr et www.cnrtl.fr/corpus/frantext/ - Est Républicain : corpus journalistique, 600 Millions de Mots, format XML/TEI, licence Créative Commons, www.cnrtl.fr/corpus/estrepublicain/ - Un corpus d'articles issus de la revue Sciences Humaines : un partenariat avec cette revue nous autorise à diffuser ces articles sous une licence Creative Commons (attribution du texte à l'auteur, pas d'utilisation commerciale, rediffusion aux mêmes conditions). www.cnrtl.fr/corpus/shs/ - TCOF Traitement de Corpus Oraux en Français (TCOF) projet « Traitement de Corpus Oraux en Français » (TCOF) est né de la volonté de conserver des corpus oraux constitués dans les années 80-90 à des fins de recherches. L’équipe de l’ATILF (UMR CNRS 7118) a élaboré l’architecture d’une première base de données de corpus alignés texte/son avec Transcriber. Celle-ci s’est progressivement enrichie à partir des années 2000 grâce à la collaboration d’autres (enseignants-)chercheurs, d’ITA et d’étudiants en Sciences du langage de l’université de Nancy. Aujourd'hui, l’équipe met à disposition de la communauté scientifique une partie de ses ressources. www.cnrtl.fr/corpus/tcof/
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	**Portail lexical** (www.cnrtl.fr/portail), base de connaissances lexicales du français qui a pour vocation de valoriser et de partager un ensemble de données issues des travaux de recherche sur le lexique français menés à l’ATILF ou au sein de partenaires du CNRTL (Académie française, ARTFL Chicago, CLEE et IRIT Toulouse, CRISCO Caen, Laboratoire Informatique de Tours, etc.). Projet évolutif, cette base permet d’obtenir à partir d’une forme lexicale, des informations morphologiques, lexicographiques et étymologiques, des informations de synonymie, d’antonymie et de proximité sémantique (proxémie) et une concordance utilisant le corpus de textes libres de droits de FRANTEXT . Il permet d’exporter les résultats du concordancier au format XML/TEI et, à notre connaissance, c’est le seul site permettant à un utilisateur d’importer dans un format normalisé un concordancier français d’une telle importance. Ce portail sert en moyenne chaque jour plus de 350 000 requêtes (www.cnrtl.fr/aide/stat/) et est intégré sous forme d’extension aux navigateurs Firefox et Chrome. **Des lexiques** avec, entre autres : - MORPHALOU : lexique ouvert des formes fléchies du français (540.000

	formes fléchies, 68.075 lemmes), respectant les propositions de normalisation pour les ressources lexicales de l'ISO (TC37/SC4). Il est en accès libre tant en consultation qu'en téléchargement. - PROLEX : issue d'un projet piloté par le Laboratoire d'informatique de l'université de Tours, cette base fournit des connaissances sur les noms propres qui constituent, à eux seuls, 10% des textes journalistiques, à travers une plate-forme comprenant un dictionnaire électronique multilingue de noms propres (Prolexbase), des systèmes d'identification des noms propres et de leurs dérivés, des grammaires locales, etc. **Un ensemble de dictionnaires français informatisés** - Dictionnaires modernes : *TLFi* : *Trésor de la Langue Française informatisé* et sa version XML, *Dictionnaire de l'Académie française* (8ème et 9ème éditions), Dictionnaire d'expressions idiomatiques français-portugais / portugais-français. www.cnrtl.fr/dictionnaires/modernes/ - Dictionnaires anciens du XVIe au XIXe siècle : *Dictionnaire de l'Académie française*, 1ère (1694), 4ème (1762), 5ème (1798), et 6ème (1835) éditions, *Dictionarium latinogallicum* de Robert Estienne (1552), *Thresor de la langue françoyse* de Jean Nicot (1606), *Dictionnaire historique et critique* de Pierre Bayle (1740), *Dictionnaire critique de la langue française* de Jean-François Féraud (1787-1788), Encyclopédie de Diderot et d'Alembert. www.cnrtl.fr/dictionnaires/anciens/
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	Cf. site du CNRTL www.cnrtl.fr, du TLFi www.atilf.fr/tlfi et de frantext www.frantext.fr FLEMM, POMPAMO, DERIF diffusés avec une licence GPL
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	Site du CNRTL www.cnrtl.fr : accès libre A noter qu'un Equipex est en cours de préparation avec la participation du LPL, du CRDO Aix, de Modyco et du LLL
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Oui : fort de l'expérience acquise en particulier à travers notre participation au projet CLARIN
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	oui

4.3 PARTENAIRE 3 : HTL (HISTOIRE DES THÉORIES LINGUISTIQUES, UMR 7597)

Descriptions générales	
Numéro du partenaire	3
Description sommaire du partenaire	HTL (Histoire des Théories Linguistiques), UMR 7597 CNRS/Paris Diderot et Paris 3 Sorbonne Nouvelle
Problématiques principales de recherche du partenaire	Histoire des théorisations et descriptions du langage et des langues Développement et diffusion de ressources dans ce domaine
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	2 ANR Corpus closes : CGL (pilote Alessandro Garcea) et DETCOL (pilote Bernard Colombat)
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant ou non	Numériseur de microfiches microfilms Scanner professionnel A4/A3
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	- 1 IE CNRS Développement logiciel - 1 IE CNRS Documentation
Expérience en numérisation	
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	Unimarc, TEI, Dublin Core
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	Fine Reader, Omnipage
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	OUI
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	Unicode, HTML, XML, TEI
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	Développement logiciel (Module Xquery)
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP,	Expérience de diffusion J2EE, HTTP, SVN (en interne)

FTP, SVN...)	
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	- Numérisation ROC, Maison Orient Méditerranéen, Lyon - Softexeperience, Devt. logiciel - Garnier Numérique : Devt. logiciel
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)**1) CTLF, Corpus Textes Linguistiques Fondamentaux** , 811 textes, 145 volumes, 30 000 pages, outils de recherche, accès ouvert, http://ctlf.ens-lyon.fr/ en lien avec l'ANR DETCOL. **2) CGL, Corpus Grammaticorum Latinorum,** Accès aux sources grammaticales de la latinité tardive**,** 104 textes (50Mo), accès ouvert, URL : http://htl2.linguist.jussieu.fr:8080/CGL/ en lien avec l'ANR CGL. **3) Grand Corpus des grammaires françaises, des remarques et des traités sur la langue, XIVe-XVIIe siècles,** 48 textes, 15 051 pages, 3 ressources constitutives et outils de recherche, accès payant ; éditeur Classiques Garnier Numérique. **4) Digital Teveram**, Collection de 800 hymnes à Shiva, Données co-produites avec l'Institut Français de Pondichéry, 100 Mo de données textuelles (dont une concordance du Tevaram en 200 000 entrées) accompagnées de données audio et de données graphiques, accès ouvert, URL http://www.ifpindia.org/ecrire/upload/digital_database/Site/Digital_Tevaram/INDEX.HTM **5) Revue HEL**, Histoire, Epistémologie, Langage , Archives ; Environ 9000 pages (en mode image [1979-2002] ou en mode texte [2002-2004]), c'est-à-dire environ 500 articles par 400 auteurs, Données cataloguées HTML, http://kaali.linguist.jussieu.fr/HEL_public_domain/HEL_archive.htm
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	-
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel	-

(gratuit, commercial)	
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	2 ingénieurs déjà mentionnés : IE CNRS Dév. informatique et Documentation
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Ensemble des personnes du laboratoire engagées dans la constitution ou la réalisation de corpus

4.4 PARTENAIRE 4 : LDI (LEXIQUES, DICTIONNAIRES, INFORMATIQUE, UMR 7187)

Descriptions générales	
Numéro du partenaire	4
Description sommaire du partenaire	Le LDI : « Lexiques, Dictionnaires, Informatique » construit des ressources lexicales et dictionnairiques en s'appuyant sur des outils informatiques. Dans cette perspective, les corpus jouent un rôle crucial tant pour l'extraction que pour la validation des ressources.
Problématiques principales de recherche du partenaire	Élaboration et analyse de dictionnaires Construction de ressources Développement d'outils de traitements automatiques de corpus
Si le partenaire est une UMS ou USR, quelles sont les UMR ou EA auxquelles elle est associée ?	
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	[en cours] ANR Créalscience Projet CNRS PICS avec Barcelone (Figement dans 4 langues) Projet Aupelf-Uref Métalangue (discours lingusitique)
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant ou non	Oui, 2 scanners
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	- Céline Poudat (MCF): construction, annotation (balisage et codage XML-TEI) et analyse de corpus - Emmanuel Cartier (MCF): développement d'outils de corpus, annotation et encodage de corpus - Fabrice Issac (MCF): développement d'outils de corpus, annotation et

	encodage de corpus - Lucie Barque (MCF): annotation linguistique de corpus - Hélène Manuelian (MCF): numérisation, annotation de corpus - 1 bibliothécaire: Jordane Raisin
Expérience en numérisation et standard ou normes utilisés pour les formats numériques (Tiff, Jpeg, Exif...)	oui
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	LMF, RDF, TEI
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	Fine reader, Omnipage
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	oui
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	Docbook, Unicode, HTML, PDF, PDF texte, PDF image, XML, XML-TEI
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	- LMF, RDF, TEI - Etiqueteurs morphosyntaxiques (TreeTagger, TnT, Cordial…)
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	- Corpindex (analyse linguistique de corpus) - RSS corpus builder (aspiration fils RSS + zonage et conversion XML) - Telanaute (aspiration web) - Textbox (analyse linguistique de corpus)
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	Expérience de diffusion HTTP, FTP, SVN (en interne)
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	Aucune
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources –	- Corpus Droits de l'homme - archive 10 ans du Monde annotés (utilisation interne) - Corpus Métalangue (articles scientifiques de linguistique, français/arabe) - Corpus Francophonie (en cours de constitution, collab. Blumenthal) - diverses archives journalistiques (aspiration RSS)

adaptation, enrichissement...)	
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	- Morfetik (dictionnaire morphosyntaxique du français contemporain), format LMF - Base Adjectifs (accès restreint) - Dictionnaire des verbes de mouvement, description morphosyntaxique et syntactico-sémantique (accès restreint) - [à compléter]
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	[licences Recherche et commerciale] - Corpindex (analyse linguistique de corpus) - RSS corpus builder (aspiration fils RSS + zonage et conversion XML) - Telanaute (aspiration web) - Textbox (analyse linguistique de corpus)
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	Site interne du LDI http://intranet-ldi.univ-paris13.fr
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Oui Poudat, Cartier, Issac
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Ensemble des personnes du laboratoire engagées dans la constitution ou la réalisation de corpus

4.5 PARTENAIRE 5 : LABORATOIRE DE LINGUISTIQUE FORMELLE (LLF)

Descriptions générales	
Numéro du partenaire	5
Description sommaire du partenaire	LLF, UMR 7110, CNRS et Université Paris Diderot. Membre du labex EFL.
Problématiques principales de	À travers l'analyse formelle des unités traditionnelles du langage---le mot, la

recherche du partenaire	phrase, l'énoncé ou le discours---et l'analyse d'un ensemble de langues très diversifié, les chercheurs du Laboratoire de Linguistique Formelle explorent le système cognitif du langage dans son entier.
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	Labex EFL ANR Elico (clos)
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant ou non	Non
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	Clément Plancq (IE2, dvpt logiciel)
Expérience en numérisation	
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	TEI, Dublin Core, RDF
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	Omnipage
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	Unicode, TEI, HTML
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	TEI, RDF + XML documenté pour le FrenchTreebank TreeTagger
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	Utilisation de Tgrep, IMS Corpus Workbench, TigerSearch, NLTK. Dvpt modules XQuery, classes Java
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	Diffusion HTTP, J2EE pour interrogation de corpus, SVN en interne
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	

Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	FrenchTreebank (voir http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php) ELICO (voir http://elico.linguist.univ-paris-diderot.fr/)
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	-
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	-
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Clément Plancq : exploitation de corpus, dvpt outils
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Intérêt pour l'encodage, l'annotation et outils d'exploitation de corpus

4.6 PARTENAIRE 6 LIDILEM

Descriptions générales	
Numéro du partenaire	
Description sommaire du partenaire	Laboratoire de linguistique spécialisé dans la linguistique française, didactique des langues, linguistique de corpus et traitement automatique des langues
Problématiques principales de recherche du partenaire	Linguistique de corpus, linguistique appliquée et didactique des langues Dans le cadre du consortium, le LIDILEM souhaite participer à la réflexion sur l'encodage des corpus, en particulier sur les annotations linguistiques, y compris pour les applications didactiques, et la réflexion sur les outils d'exploitation. Nous pouvons apporter une expertise sur les annotations linguistiques, étiquetage et annotations syntaxiques, annotation sémantique.
Si le partenaire est une UMS ou USR, quelles sont les UMR ou EA auxquelles elle est associée ?	
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	Projets passés liés aux corpus : - ANR Corpus Scientext (2006-2010) : un corpus et des outils pour étudier le positionnement et le raisonnement de l'auteur dans les écrits scientifiques (http://scientext.msh-alpes.fr) - PPF (2003-2007) : « Développement et exploitation de ressources linguistiques pour la didactique du français à l'aide d'outils de TAL. Etude des marqueurs linguistiques de la subjectivité et de la polyphonie. » - Constitution de corpus avec annotations anaphoriques (en collaboration avec le XRCE). Corpus diffusé par ELRA. (http://catalog.elra.info/product_info.php?products_id=634&language=fr) - Projet CARMEL Technolangue (2003-2005). Alignement de corpus multilingues. Projets en cours : - ANR franco-allemande EMOLEX (lexique des émotions dans cinq langues européennes : sémantique, syntaxe et dimension discursive) http://emolex.eu/ - Manuscrits de Stendhal : http://www.manuscrits-de-stendhal.org/ - Projet corpus de SMS : http://www.alpes4science.org/ - Participation au GDR « Production écrite : Apprentissage et Expertise » pour l'axe Constitution, édition et annotation de corpus de textes scolaires multilingues, multi-genres
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant ou non	Logiciels d'OCR type omnipage
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	Pas de personnel dédié. Vacataires dans le cadre de projets financés
Expérience en numérisation et standard ou normes utilisés pour les formats numériques (Tiff, Jpeg, Exif...)	
Expérience en catalogage et	TEI P5

standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	XML TEI P5
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	- XML TEI P5, XCES - Expérience d'utilisation de logiciels et d'annotation à l'aide de ces outils : étiquetage morpho-syntaxique (Tree Tagger, Cordial, …), logiciels d'annotation syntaxique de dépendance (Syntex, Xerox Incremental Parser, Connexor). - Expérience d'alignement multilingue, au niveau des paragraphes, des phrases, des mots
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	- Expérience de développement d'interface exploitant des corpus annotés syntaxiquement et structurés TEI - Outil d'alignement (Alinéa) - Concordancier utilisant des expressions linguistiques complexes (ConcQuest)
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	- Diffusion et interrogation en ligne du corpus Scientext (http://scientext.msh-alpes.fr). Contrat Creative Commons. (Logiciels utilisés : PHP/AJAX/HTML) - Interrogation en ligne de textes multilingues alignés http://w3.u-grenoble3.fr/kraif/ConcQuest/concquest.php - Diffusion d'outils d'alignement multilingue (Alinea) et d'outil de recherche d'expressions complexes (ConcQuest)
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	Non
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	- Corpus Scientext : corpus d'écrits scientifiques annotés structurellement (TEI P5) comprenant 4,8 m de mots (8 disciplines et 3 genres). Interrogeable en ligne sur http://scientext.msh-alpes.fr par requêtes syntaxiques et sémantiques. Corpus gratuitement diffusé sur demande pour la partie libre de droits. - Corpus SMS en cours de constitution. Disponibilité prévue pour les chercheurs. - Corpus avec annotations anaphoriques (élaboré en collaboration avec le XRCE) diffusé par ELRA. - Corpus PPF Subjectivité (annotation du lexique des émotions, discours rapporté, passages entre guillemets). Annoté en XML. A mettre aux

	normes pour diffusion.
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	- Lexique des émotions classé sémantiquement.(à mettre aux normes)
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	- Logiciel Alinéa (alignement multilingue). Logiciel Freeware. - Logiciel ConcQuest (concordancier pour expressions complexes, y compris syntaxe). Logiciel Freeware. - Outil d'exploitation de corpus annotés syntaxiquement et structurellement en préparation. Il pourra être mis à la disposition de la communauté.
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	- Formation aux questions d'encodage et exploitation de corpus annotés syntaxiquement (en dépendance). - Annotation linguistique (sémantique, phénomènes énonciatifs, collocations, co-référence)
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	- Formation aux outils de catalogage et encodage. - Formation aux outils d'exploitation de corpus.

4.7 PARTENAIRE 7 : ICAR UMR 5191

Descriptions générales	
Numéro du partenaire	8
Description sommaire du partenaire	Le laboratoire ICAR, associé au laboratoire DDL au sein du labex Aslan, se caractérise par des activités scientifiques pluridisciplinaires focalisées sur l'analyse multidimensionnelle des usages de la langue dans l'interaction et dans le texte, appréhendée de manière outillée sur de grands corpus de données orales interactives et textuelles.
Problématiques principales de recherche du partenaire	Les domaines scientifiques concernés sont la linguistique interactionnelle, les approches pluridisciplinaires de l'interaction, la linguistique de corpus, le

	traitement automatique des corpus écrits et oraux, l'étude de l'acquisition, de l'apprentissage et de la didactique des langues et des sciences, la linguistique française.
Si le partenaire est une UMS ou USR, quelles sont les UMR ou EA auxquelles elle est associée ?	
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	- projet ANR Corptef 2008-2011 : corpus représentatif des premiers textes français http://corptef.ens-lyon.fr - projet ANR-DFG Srcmf 2009 - 2012 : Syntactic Reference Corpus of Medieval French https://listes.cru.fr/wiki/srcmf/index - projet ANR Textométrie 2007-2010 : Fédération des recherches et développements en textométrie autour de la création d'une plateforme logicielle ouverte http://textometrie.ens-lyon.fr - projet ANR Vecmas 2008-2011 : Valorisation et édition critique des manuscrits arabes subsahariens http://vecmas-tombouctou.ens-lyon.fr - projet région Rhône-Alpes Cluster 13 - 2009 : Edition numérique interactive de la Queste del saint Graal http://textometrie.risc.cnrs.fr/txm - projet ILF 2009- : Grande grammaire historique du français - consortium CCFM 2004 : Consortium international pour les corpus de français médiéval http://ccfm.ens-lyon.fr
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant ou non	CopyBook et scanners d'appoint (couleur et avec chargeur)
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	- A. Lavrentiev : mise en place et suivi de chaine de numérisation (avec prestataires, vacataires et CDD), encodage XML-TEI, enrichissement linguistique, catalogage, intégration dans les outils d'exploitation - C. Guillot : suivi de vacataires en numérisation et annotation linguistique, catalogage. - S. Heiden : mise en place et suivi de projets de développement d'outils d'exploitation de corpus, enrichissement linguistique, intégration dans les outils d'exploitation - Tom Rainsford : annotation linguistique, documentation sur les outils d'exploitation - M. Decorde : développement d'outils d'encodage, d'annotation, d'enrichissement linguistique et d'exploitation de corpus
Expérience en numérisation et standard ou normes utilisés pour les formats numériques (Tiff, Jpeg, Exif...)	Oui, usage de JPEG
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	- TEI, catalogage CCFM, export Dublin Core pour OAI-PMH - documentation des pratiques en ligne, http://bfm.ens-lyon.fr/rubrique.php3?id_rubrique=112 sous licence Creative Commons - gestion de catalogue en SGBDR
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	Fine Reader
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	Les correcteurs utilisent Word, OpenOffice et Oxygen
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI,	- encodage en XML-TEI, Unicode, CCFM - avec Oxygen et TXM - embauche de vacataires pour l'océrisation, la relecture et le balisage des textes.

METS, Docbook, HTML, PDF image, PDF Texte...)	- documentation de la pratique XML-TEI publique et en ligne, http://bfm.ens-lyon.fr/rubrique.php3?id_rubrique=112 sous licence Creative Commons
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	- annotation en XML-TEI - usage de TreeTagger, TnT, Cordial, MATE Tools - développement du jeu d'étiquette CATTEX pour l'ancien français
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	- développement de la plateforme TXM (analyse de grands corpus textuels compatibles XML-TEI). Impulsion initiale par le projet ANR Textométrie, puis contrats successifs pour amélioration et maintenance (Univ. Lyon 3, CNRS, DGLFLF, ANR, Equipex...) - usage de Tiger Search
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	- diffusion de la BFM avec le logiciel Weblex (CGI) : http://weblex.ens-lsh.fr/wlx [gestion de chartes utilisateurs, inscriptions, négociation avec éditeurs commerciaux...] - diffusion de la BFM avec la plateforme TXM (J2EE), à partir d'octobre 2011 : http://textometrie.risc.cnrs.fr/txm
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	- commandes de reproductions à des bibliothèques - numérisation et océrisation d'éditions critiques par une entreprise dans le cadre d'un projet ANR - prestations de retouche d'images de manuscrits
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	* Base de Français Médiéval (BFM) http://bfm.ens-lyon.fr : - la Base de Français Médiéval comporte des textes intégraux écrits entre le IXe et la fin du XVe siècle (près de 1 500 000 occurrences-mots). Elle se distingue par son empan diachronique et par la diversité de ses données (variation géographique, des genres textuels, vers / prose). - accès gratuit contre signature d'une charte : accès variable aux textes selon les contraintes négociées avec les éditeurs commerciaux ou accès total aux ressources libres - accès en ligne (voir le site web) - corpus complémentaire de Frantext * Edition numérique interactive du ms. Lyon BM p.a. 77 : images numériques du manuscrit, édition « multi-facettes » du texte en ancien français, étiquetage morphosyntaxique, traduction en français moderne, accès gratuit en ligne http://txm.risc.cnrs.fr/txm TEI URL provisoire), sources XML) disponibles sous licence Creative Commons
Ressources linguistiques disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel	- Modèle morphosyntaxique du français médiéval (XIe - XVe s.) pour le logiciel TreeTagger, http://bfm.ens-lyon.fr/article.php3?id_article=324 (aligné avec Multext) - accès gratuit en ligne sous licence Creative Commons

(gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	- plateforme open-source TXM - analyse de grands corpus textuels compatibles XML-TEI - diffusée gratuitement pour Windows, Mac OS X et Linux : https://sourceforge.net/projects/textometrie , diffusée également sous forme de portail web pour la mise en ligne de corpus, les sources sont disponibles sous licence GPL v3 pour un développement communautaire
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	- portail TXM en cours de mise en place : http://textometrie.risc.cnrs.fr/test, accès ouvert ou restreint (selon le type d'hébergement souhaité par les projets)
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	- A. Lavrentiev & C. Guillot : catalogage et encodage XML-TEI (TEI, XSLT...) - S. Heiden & B. Pincemin : préparation et analyse de corpus XML-TEI enrichis par des outils de TAL (Unicode, formats, TEI, TAL, textométrie, TXM...)
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	OUI

4.8 PARTENAIRE 8 : LILPA

Descriptions générales	
Numéro du partenaire	8
Description sommaire du partenaire	UR 1339 Linguistique, Langues et Parole (LiLPa), Université de Strasbourg
Problématiques principales de recherche du partenaire	Linguistique synchronique et diachronique, phonetique, linguistique de corpus, traitement automatique de langues, didactique des langues
Projets impliquant des corpus écrits dans lesquels le partenaire est impliqué (PCRD, ANR Corpus, CPER...)	Projet PEPS MC4 Modélisation Contrative et Computationnelle des Chaînes de Coréférence (2011-2012) ; Projet CLASSYN (programme Procope, Egide) (janvier 2010-decembre 2011) ; Projet CAP : A Hierarchy-Based Lexical Function (collaboration avec le projet CLARIN, dans le cadre du appel a projet en SHS du mai 2009-mars 2011) Projet AUF LTT « Collocations en contexte : étude et analyse contrastive » (juin 2006 – avril 2008) ;
Moyens et plate-forme technologiques utilisés	
Matériel de numérisation existant	

ou non	
Personnel assigné aux activités de numérisation, de catalogage, d'annotation, de développement d'outils ou d'exploitation de corpus	Des vacations ont été prévues pour divers projets
Expérience en numérisation	
Expérience en catalogage et standard ou normes utilisés pour les métadonnées (Unimarc, METS, EAD, TEI, Dublin Core, RDF...)	TEI, Dublin Core
Expérience en océrisation et logiciels utilisés (Fine Reader, Omnipage...) ou techniques de saisie (double saisie, en aveugle...)	
Expérience en correction d'épreuves et logiciels utilisés (correcteur orthographique...)	Oui
Expérience en encodage et standard ou normes utilisés pour les données (Unicode, XML, TEI, METS, Docbook, HTML, PDF image, PDF Texte...)	Unicode /UTF-8, XML, TEI
Expérience en annotation et standard ou normes utilisés pour les données (TEI, XCES, RDF...) et outils de TAL utilisés éventuels (TreeTagger, Cordial...)	TEI, XCES TreeTagger, Cordial, Syntex et parseur de Bohnet
Expérience en développement d'outils d'exploitation de corpus et logiciels développés en interne (précisez le type de licence de diffusion des logiciels)	Outils d'annotation de chaînes de références (RefGen) (en partenariat avec l'entreprise RBS), développement des ressources pour l'étiquetage (utilisés actuellement par l'étiqueteur TTL)
Expérience en diffusion de corpus (plateforme de diffusion en production, serveur de téléchargement...) et logiciels et matériels utilisés (J2EE, HTTP, PHP, FTP, SVN...)	Pas pour les corpus écrits : le projet Danok (www.danok.eu) a eu comme objectif la numérisation de documents textes, vidéos, sons de l'histoire culturelle du Rhin Supérieur.... Logiciels : HTTP, logiciel de base de données propriétaire
Expériences d'externalisation	
Expériences d'externalisation pour la numérisation, le catalogage, l'encodage, l'annotation, le développement d'outils ou d'exploitation de corpus	
Corpus et ressources disponibles	
Corpus écrits disponibles (pour chaque corpus : donner une description sommaire, le type d'accès - ouvert, restreint..., le mode de diffusion – hébergement en ligne, copie de CD... ainsi que les relations éventuelles avec d'autres corpus ou sources – adaptation, enrichissement...)	1) corpus parallèle aligné au niveau lexical FR-EN; EN-RO (1000 phrases), format XCES (acces libre, copie de CD); 2) Corpus annoté en relations de coréférence (10000 tokens), en format XML(acces restreint, copie de CD); 3) Corpus français de textes scientifiques-textes de vulgarisation (1000000 tokens), analysé syntaxiquement avec l'analyseur syntaxique de Bohnet (2009), format CONLL (acces libre, copie de CD);
Ressources linguistiques	1) dictionnaire multilingue de collocations, en format XML (français-roumain,

disponibles : lexique de mots simples ou composés, modèle linguistique d'étiqueteur ou de parseur... (pour chaque ressource : donner une description sommaire incluant les standard ou normes linguistiques utilisés (LMF, Multext, ISOCat...) et le type d'accès - ouvert, restreint...) – et le coût éventuel (gratuit, commercial) et le mode de diffusion – hébergement en ligne, copie de CD...)	250 entrées); 2) prototype pour un dictionnaire bilingue pour la traduction français-espagnol (Transverb) (format SQL) 3) corpus français étiqueté et lemmatisé corrigé manuellement (environ 900000 tokens), jeu d'étiquettes Multext (acces restreint, copie de CD)
Logiciels d'encodage, de traitement ou d'exploitation de corpus écrits ou de ressources linguistiques disponibles (pour chaque outil : donner une description sommaire et le type de diffusion – open-source (licence), non open-source – et le coût éventuel (gratuit, commercial)	Outils d'annotation de chaînes de références (RefGen) (en partenariat avec l'entreprise RBS)
Plateformes de diffusion de corpus en ligne disponibles (adresse, type d'accès...)	-
Capacités de formation aux standard et outils	
Personnel disponible pour former sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	-
Intérêt pour la formation aux standard et outils	
Personnel intéressé par des formations sur les standard et outils liés aux corpus écrits : numérisation, catalogage, encodage, annotation, développement d'outils ou exploitation de corpus	Oui

4.9 PARTENAIRE 9 : ALPAGE

Descriptions générales	
Numéro du partenaire	
Description sommaire du partenaire	Equipe projet Alpage, UMR INRIA-Paris 7

Problématiques principales de recherche du partenaire	Traitement automatique des langues, linguistique informatique
Moyens et plate-forme technologiques utilisés	
Formats et standards habituellement utilisés *Formats : MySQL, XML, PostgreSQL,…* *Standards : DC, RDF, TEI, EAD, METS…*	XML, TEI
Serveur d'hébergement existant	https://gforge.inria.fr/
Outils développés en interne	outils de traitement automatique du francais (segmentation, tokenisation, etiquetage morpho-syntaxique, analyse syntaxique): voire https://www-roc.inria.fr/alpage-wiki/tiki-index.php?page=Logiciels

Tableau des fonds déjà disponibles

Nom du fonds	Description sommaire	Volumétrie	URL (du fonds ou des méta-données)
French TimeBank	corpus temporel du francais annote a la norme ISO-TimeML		https://gforge.inria.fr/projects/fr-timebank/