



Le Master *Recherche*

spécialité

Traitement Automatique des Langues (TAL)

Année 2008-2009

URL : <http://plurital.org>

URLs « officielles » du Master *cohabilité* Paris III / Paris X / Inalco :

Paris X : http://www.u-paris10.fr/15586752/0/fiche_pagelibre/

Paris III : http://www.univ-paris3.fr/master_recherche.htm

Formation proposée par

I'Université Paris X – Nanterre

I'Université Paris III - Sorbonne nouvelle

I'INALCO

Responsables:

Marcel Cori (Paris X)
André Salem (Paris III)
Monique Slodzian (Inalco)

Commission pédagogique :

Jean-Michel Daube (INALCO)
Serge Fleury (Paris III)
Marcel Cori (Paris X)

Hypertoile du MASTER : le site pluriTAL

<http://plurital.org>

Weblog du master :

<http://tal-p3.wordpress.com/>

Sommaire

1	CONTACTS	4
2	INSCRIPTIONS PEDAGOGIQUES ET REUNIONS DE RENTREE	4
3	GROUPE PLURITAL (LISTE DE DIFFUSION)	4
4	LOCALISATIONS	5
4.1	PARIS X	5
4.2	ILPGA / PARIS 3	5
4.3	INALCO	5
5	OBJECTIFS GENERAUX	6
6	OBJECTIFS D'APPRENTISSAGE	6
7	INSCRIPTIONS – VALIDATIONS D'ACQUIS	7
7.1	INSCRIPTION EN 1 ^{ERE} ANNEE	7
7.2	INSCRIPTION EN 2 ^{EME} ANNEE.....	7
7.3	JURY	7
8	ORGANISATION PEDAGOGIQUE DU MASTER	8
9	DIRECTEURS DE RECHERCHE POUR LE MASTER 2^{EME} ANNEE	9
10	MAQUETTE DES ENSEIGNEMENTS	10
10.1	M1 (1 ^{ERE} ANNEE, 1 ^{ER} SEMESTRE) :	10
10.1.1	<i>Tronc commun</i>	10
10.1.2	<i>liste M1-TAL</i>	11
10.2	M2 (1 ^{ERE} ANNEE, 2 ND SEMESTRE)	13
10.2.1	<i>Tronc commun</i>	13
10.2.2	<i>liste M2-TAL</i>	14
10.3	M3 (2 ^{NDE} ANNEE, 1 ^{ER} SEMESTRE) :	15
10.3.1	<i>Tronc commun</i>	15
10.3.2	<i>liste M3-TAL</i> :	16
10.4	M4 (2 ^{NDE} ANNEE, 2 ND SEMESTRE) :	17
11	PLANNING DES COURS DU TRONC COMMUN MASTER TAL RECHERCHE	18
12	DESCRIPTIF ET HORAIRES DES COURS (1^{ERE} ET 2^{EME} ANNEES), TRONC COMMUN	20
12.1	GRAMMAIRES FORMELLES	20
12.2	ANALYSE SYNTAXIQUE AUTOMATIQUE.....	20
12.3	GESTION INFORMATIQUE DU MULTILINGUISME	20
12.4	TRAITEMENT DE L'ORAL / INFORMATIQUE ET PHONETIQUE	21
12.5	PROGRAMMATION ET PROJET ENCADRE (SEMESTRE 1)	21
12.6	BASES DE DONNEES POUR LINGUISTES	22
12.7	STATISTIQUE ET ANALYSE MULTIDIMENSIONNELLE	22
12.8	CORPUS PARALLELES/COMPARABLES	22
12.9	RECHERCHE ET EXTRACTION D'INFORMATION	23
12.10	SEMANTIQUE FORMELLE.....	23
12.11	SYNTAXES FORMELLES	23
12.12	DOCUMENT STRUCTURE	24
12.13	PROGRAMMATION ET PROJET ENCADRE (SEMESTRE 2)	24
12.14	MODELISATION DES LANGUES : MOTS ET GRAMMAIRE.....	24
12.15	FORMALISMES SYNTAXIQUES ET TRAITEMENT AUTOMATIQUE	25
12.16	DOCUMENT STRUCTURE : XML	25
12.17	ANALYSE DU DISCOURS ET LEXICOMETRIE	25
12.18	ÉCRITURES ET MULTILINGUISME	25
12.19	MULTIMODALITE DU DOCUMENT NUMERIQUE.....	26
12.20	SEMANTIQUE DES TEXTES MULTILINGUES	26
12.21	SYNTHESE DE CONFERENCES	26
12.22	EXPOSES D'ARTICLES.....	26

1 Contacts

Voir la page « Contact » sur la page web du MASTER.

Lien direct : <http://tal.univ-paris3.fr/plurital/contact.html>

2 Inscriptions pédagogiques et réunions de rentrée

Rentrée universitaire 2008/2009

Journée d'accueil du MASTER : (date disponible en ligne sur le site [pluriTAL](#))

Début des cours : (date disponible en ligne sur le site [pluriTAL](#))

Inscriptions pédagogiques :

PARIS 3	PARIS X	INALCO
(cf secrétariat P3/ILPGA ou sur le site pluriTAL)	(cf secrétariat PX)	(cf secrétariat INALCO)

3 Groupe PluriTAL (liste de diffusion)

Inscription **obligatoire** pour tous les étudiants devant suivre des cours du Master pluriTAL.

Voir la page « Groupe » sur la page web du MASTER (site [pluriTAL](#)).

Lien direct : <http://tal.univ-paris3.fr/plurital/groupeplurital.html>

Inscription au groupe Yahoo-PluriTAL sur cette page (avec liste de diffusion modérée)

4 Localisations

4.1 Paris X

Pour se rendre l'Université de Nanterre : RER A, Direction Saint Germain-en-Laye, station Nanterre Université. Plan de l'université voir <http://www.u-paris10.fr>. Pour se rendre au bâtiment L. **a)** sortir du RER en queue de train en venant de La Défense/Paris, sortie principale par un escalier montant, prendre sur la gauche, en descendant une rampe d'accès. **b)** en bas de la rampe, tourner immédiatement à droite à 90 degrés. **c)** S'engager à droite dans le passage pratiqué dans la haie une quinzaine de mètres plus loin. **d)** suivre le chemin goudronné qui passe sous le bâtiment (pilotis) **e)** suivre la route parsemée de brise-vitesses, qui longe le bâtiment de l'université à G. et une clôture à D. avec un bâtiment flambant neuf **f)** au bout d'un moment, à travers un cèdre, on distingue un bâtiment aux lignes volontaires. C'est le bâtiment L. Les machines à café sont sur la mezzanine, au premier étage.

4.2 ILPGA / Paris 3

ILPGA / Université Paris 3 Sorbonne nouvelle,

19 rue des Bernardins, 75005 Paris - **Tél.** : 01 44 32 05 70 **Fax** : 01 44 32 05 73

Plan du quartier : <http://voici.monplan.com/ilpga>



4.3 INALCO

Siège de l'INALCO

2 rue de Lille
75343 Paris cedex 07

Métro :

ligne 4 - station Saint Germain des Prés

ligne 12 - station Rue du Bac

ligne 17 - station Palais Royal-Musée du Louvre

Autobus : lignes 24, 27, 39, 48, 69, 95 - station Pont du Carrousel

R.E.R. : ligne C - station Musée d'Orsay

Standard : 01 49 26 42 00 - **Fax** : 01 49 26 42 99

A partir de la rentrée 2008/2009, les cours de l'INALCO (M1 et M2) auront lieu au 49bis, avenue de la Belle Gabrielle, 75012 PARIS.

5 Objectifs généraux

Le traitement automatique de la langue et des textes fait appel à des méthodes et des savoirs multiples. Il s'agit de :

1. Disposer des pré-requis en linguistique : maîtriser les manipulations débouchant sur des descriptions détaillées de faits de langue, connaître les bases des grands domaines des sciences du langage (phonétique et phonologie, morphologie, syntaxe et sémantique) ;
2. Connaître les bases de la recherche et extraction d'information, de la constitution et de la gestion de corpus et de ressources, y compris multilingues : les corpus sont des mines d'information pour une description réaliste d'emplois d'une langue, les techniques de la recherche et de l'extraction d'information permettent de rapatrier les documents ou les parties de documents jugés pertinents pour un besoin particulier ;
3. Exprimer les règles et les régularités à l'œuvre, par le biais des grammaires formelles et des traitements quantitatifs pour savoir passer d'une description linguistique à une représentation plus précise permettant son utilisation par des logiciels.

6 Objectifs d'apprentissage

Les bases solides en linguistique sont complétées par des dimensions plus spécifiques au traitement automatique des langues. La « matière » des textes, c'est-à-dire leur encodage informatique, est présentée, en particulier les solutions qui ont été offertes pour la gestion des différents systèmes d'écriture. Est également analysé le « grain de la voix » : les phénomènes propres à l'oral (marqueurs de discours, amorces et répétitions, parole superposée).

Les traitements automatiques de la langue associent des traitements s'appuyant sur des régularités constatées et des traitements reposant sur des règles explicites. Les premiers, approximatifs mais robustes, supposent des quantifications. Ce sont par exemple les principes des moteurs de recherche qui calculent une ressemblance globale entre les mots de la requête et les documents indexés. Les seconds, fins mais fragiles, explicitent le savoir des locuteurs. La formation associe alors les deux grandes familles de traitement. L'analyse statistique des données textuelles, la recherche et l'extraction d'information relèvent de la première famille. La sémantique et la syntaxe formelles de la seconde.

Les données textuelles supposent de maîtriser à la fois un « couteau suisse » permettant de faire des opérations de formatage, des nettoyages, des changements de caractères, d'ajouter des étiquettes à des mots en utilisant des dictionnaires, etc. Les langages de scripts (Python, Perl, PHP) constituent de tels couteaux suisses. Par ailleurs, les textes sont structurés (en parties, sous-parties, bibliographie, index, notes, etc.). Cette structure est de plus en plus explicite (par des balises par exemple en HTML). Il faut donc savoir expliciter la structure quand c'est nécessaire ou utiliser cette structure quand elle est présente. C'est l'ensemble des techniques et outils autour de XML.

Ces manipulations des documents nécessitent la connaissance de l'algorithmique, c'est-à-dire le découpage raisonné d'un problème en modules articulés, ainsi que celle des structures de données nécessaires (tables, arbres, graphes) ainsi que celle des bases de données, qui permettent de stocker et de manipuler des connaissances détaillées (dictionnaires).

Le travail sur des données textuelles sous forme électronique repose souvent sur l'association de compétences variées. Une part importante de la formation prépare aux projets collectifs, avec la répartition des tâches, la gestion des échéances et des rendus qu'ils impliquent. Les stages en laboratoire ou en entreprise contribuent à ce volet de la formation. Le rapport de stage et sa soutenance sont l'occasion de manifester la compréhension du collectif dans lequel s'insère le stagiaire, des problèmes qu'on lui demande de traiter, de l'apport qu'il a pu effectuer.

7 Inscriptions – validations d'acquis

7.1 Inscription en 1^{ère} année

L'étudiant devra être titulaire d'une licence dans l'une des mentions ou spécialités suivantes : « Sciences du Langage » ; « Lettres » ; « Langues, littératures et civilisations étrangères » ; « Sciences humaines et sociales » ; « Psychologie » ; « Mathématiques appliquées aux sciences sociales » ou d'une bi-licence ou encore d'une licence inter-mentions ayant une composante de Sciences du langage (ex. « Sciences du langage, civilisation européenne : langue » ; « Lettres/sciences du langage »)

7.2 Inscription en 2^{ème} année

L'étudiant devra avoir validé au minimum un M1 « Sciences du langage » à Paris III, Paris X ou dans une autre université européenne, il devra en outre avoir suivi des enseignements en informatique ou effectuer une mise à niveau.

Toute demande qui ne remplirait pas cette condition sera examinée par une commission de validation d'acquis, qui tiendra compte des ECTS en « Sciences du langage » obtenus par l'étudiant et de son parcours personnel.

7.3 Jury

Un jury se réunit en fin de premier et de second semestre de MASTER pour évaluer les résultats obtenus par les étudiants et faire organiser, le cas échéant, des sessions de rattrapage dans les matières où les étudiants auraient échoué.

Un jury d'admission statue sur les candidatures des étudiants souhaitant poursuivre leur formation en MASTER 2 - Recherche.

8 Organisation pédagogique du MASTER

La formation proposée a été conçue comme devant mener progressivement d'une formation « généraliste » en TAL à une spécialité; il s'agit au cours de ce Master de donner aux étudiants qui auront confirmé leur intérêt pour le Traitement Automatique des Langues les moyens les moyens de maîtriser les outils indispensables à l'activité de recherche dans le domaine (modélisation syntaxique, exploration de textes sur ordinateur, analyses quantitatives sur corpus de textes...).

La mutualisation des compétences permet d'offrir aux étudiants dès la première année du Master une large palette de savoirs et de savoir faire, ces étudiants pourront ensuite choisir une des spécialités offertes par une des 3 entités constituantes du Master (orientation recherche ou professionnelle).

La première année du Master est organisé autour

- d'un module "Linguistique"
- d'un module "TAL et ingénierie linguistique"
- d'un module " Informatique "
- d'une langue vivante et d'une unité libre

La seconde année est organisée autour

- d'un module "Linguistique et TAL »
- d'un module "Linguistique"
- d'un module « synthèses de conférences »
- d'un travail de recherche (avec stage de 3 mois) donnant lieu à un mémoire
- d'une unité libre
- d'une langue vivante

9 Directeurs de recherche pour le Master 2^{ème} année

Nom	Spécialité	Université/Laboratoire
M. André SALEM, Professeur salem@msh-paris.fr	lexicométrie, analyse de discours, statistique textuelle	Paris 3 - SYLED Centre de Lexicométrie et d'Analyse Automatique des Textes (CLA ² T)
M. Marcel Cori mcori@u-paris10.fr	syntaxe formelle	Paris X - Modyco
M. Sylvain Kahane Sk@ccr.jussieu.fr	syntaxe formelle, lexique formel	Paris X
Anne Lacheret	phonétique, synthèse de la parole, prosodie	Paris X
Bernard Laks	phonologie formelle	Paris X
M. Monique Slodzian mslodz@inalco.fr		INALCO - CRIM

(liste provisoire)

10 Maquette des enseignements

10.1 M1 (1^{ère} année, 1^{er} semestre) :

10.1.1 Tronc commun

Code P3	Code PX	Unités d'enseignement	crédits
L7TIL	LMSUFT12	TAL et ingénierie linguistique 1	12
L7T01	LMSDL189	Grammaires formelles	3
L7T02	LMSDL184	Analyse syntaxique automatique	3
L7T03	LMSDL182	Gestion informatique du multilinguisme	3
L7T52	LMSDL188	Informatique et phonétique	3
L7LIN	LMSUFT11	Linguistique 1 9 crédits à prendre dans la liste M1-TAL (infra)	9
L7INF	LMSUCT13	Informatique 1	6
L7T04	LMSDL180	Programmation et projet encadré	3
L7T05	LMSDL186	Bases de données pour linguistes	3
L7LVE	LMSUCT14	Langue vivante	3

10.1.2 liste M1-TAL

Code P3	Code PX	Code Inalco	SITE Paris X	crédits
L7L01	LMSDL151	-	<i>Acquisition des langues et psychologie</i>	6
L7L02	LMSDL162/168	-	<i>Linguistique du texte littéraire et sémantique des genres</i>	6
L7L03	LMSDL118	-	<i>Syntaxe et sémantique</i>	6
L7L04	LMSDL176	-	<i>Pragmatique et approches communicatives</i>	6
Code P3	Code PX	Code Inalco	SITE Paris III ¹	crédits
LYM01	-		<i>Eléments de Phonologie</i>	3
LYM02	-		<i>Théories sollicitées dans l'enseignement/apprentissage des langues</i>	3
LYM04	-		<i>Quelques problèmes récurrents dans la description des langues africaines</i>	3
LYM05	-		<i>Méthodes de collecte des données dialectales, linguistique de terrain</i>	3
LYM06	-		<i>Questions d'interface syntaxe et sémantique</i>	3
LYM08	-		<i>Diversité des langues : approches typologiques en morphosyntaxe</i>	3
LYM09	-		<i>Acquisition du langage : narration et interaction adulte-enfant</i>	3
LYM12	-	-	<i>Acquisition du langage et développement pragmatique : le cas de la référence et des conduites dialogiques</i>	
LYM13	-		<i>Génération automatique des textes</i>	3
L7444	-		<i>Phonétique expérimentale</i>	3
LZMAR	-		<i>Phonologie prosodique</i>	3
FYC4L	-		<i>Les connecteurs et l'organisation des textes</i>	3
FYR1L	-		<i>Analyse des discours de transmission de connaissances</i>	3
FYF2L	-		<i>Problèmes d'historiographie des théories linguistiques</i>	3
FYF5L	-		<i>Syntaxe et sémantique : les mots en qu-</i>	3
L7SOC	-		<i>Sociolinguistique (cours de synthèse)</i>	3
L7PHO	-		<i>Phonétique</i>	3
L7FYF	-		<i>Morphosyntaxe-Fondamentaux</i>	3

¹ Les descriptifs des cours de Paris III dans la liste M1-TAL sont disponibles dans la brochure du **MASTER Spécialité "Langage, Langues, Modèles" de la Mention "Sciences du langage ; Didactique des langues"** à cette adresse : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/Brochure-MASTER-LLM.pdf> ou dans la brochure du **MASTER Spécialité Phonétique** de la même mention et disponible à cette adresse : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/lmd-2005-2006/sitewebimdpbonetique.htm>. On pourra aussi se renseigner au secrétariat de l'ILPGA.

L7FYA	-		<i>Morphosyntaxe-Approfondissement</i>	3
L7DID	-		<i>Didactique</i>	3
F7ADI	-		<i>Analyse du discours</i>	3
Code P3	Code PX	Code Inalco	SITE Inalco	crédits

10.2 M2 (1^{ère} année, 2nd semestre)

10.2.1 Tronc commun

Code P3	Code PX	Unités d'enseignement	crédits
L8TIL	LMSUFT22	TAL et ingénierie linguistique 2	9
L8T53	LMSDL281	Statistique et analyse multidimensionnelle	3
L8T01	LMSDL282	Corpus parallèles/comparables	3
L8T02	LMSDL283	Recherche et extraction d'information	3
L8LIN	LMSUFT21	Linguistique 2	9
L8T05	LMSDL217	Sémantique formelle	3
LZM09	-	Syntaxes formelles (Paris III)	3
		Et 1 enseignement à prendre dans la liste M2-TAL <i>(cf infra)</i>	3
L8INF	LMSUCT23	Informatique 2	9
L8T06	LMSDL284	Document structuré	3
L8T07	LMSDL280	Programmation et projet encadré	6
L8LIB	LMSULT24	Libre	3

10.2.2 liste M2-TAL

Code P3	Code PX	Code Inalco	SITE Paris X	Crédits
L8L01	LMSDL221	-	<i>Grammaire et sémantique</i>	6
L8L05	LMSDL235	-	<i>Phonologies multilinéaires et prosodie</i>	6
L8L06	LMSDL241	-	<i>Lexique et didactique</i>	6
L8L07	LMSDL261	-	<i>Ancien français</i>	6
L8L07	LMSDL256	-	<i>Philosophie du langage et épistémologie de la linguistique</i>	
Code P3	Code PX	Code Inalco	SITE Paris III ²	Crédits
LZM01	-	-	<i>Sémantique lexicale : structure argumentale, temps, aspect et aktionsart</i>	
LZM02	-	-	<i>Pédagogie de la phonétique suprasegmentale</i>	
LZM03	-	-	<i>L'acquisition de langue étrangère : perspectives théoriques</i>	
LZM04	-	-	<i>Pédagogie de la phonétique supra-segmentale</i>	
LZM05	-	-	<i>Théories de la construction nationale, planification et aménagement linguistique</i>	
LZM07	-	-	<i>Le français en francophonie : construction politique et sociale</i>	
LZM10	-	-	<i>Approche linguistique de l'apprentissage de l'écrit</i>	
FZB7L	-	-	<i>Sociolinguistique : langues communes, langues d'école</i>	
FZC1L	-	-	<i>Analyse de l'oral premier niveau</i>	
FZL1L	-	-	<i>Énonciation, discours et texte dans l'écrit littéraire</i>	
FZGOL	-	-	<i>Approches linguistiques du texte littéraire médiéval</i>	
L8444	-	-	<i>Phonétique expérimentale</i>	
F8HIL	-	-	<i>Histoire des idées linguistiques (cours de synthèse)</i>	
L8ACQ	-	-	<i>Acquisition (cours de synthèse)</i>	
F8SEM	-	-	<i>Sémantique (cours de synthèse)</i>	
F8DIA	-	-	<i>Diachronie (cours de synthèse)</i>	
L8TAL	-	-	<i>TAL (cours de synthèse)</i>	
Code P3	Code PX	Code Inalco	SITE Inalco	Crédits

² Les descriptifs des cours de Paris III dans la liste M2-TAL sont disponibles dans la brochure du **MASTER Spécialité "Langage, Langues, Modèles" de la Mention "Sciences du langage ; Didactique des langues"** à cette adresse : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/Brochure-MASTER-LLM.pdf> ou dans la brochure du **MASTER Spécialité Phonétique** de la même mention et disponible à cette adresse : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/lmd-2005-2006/siteweb/lmdphonetique.htm>. On pourra aussi se renseigner au secrétariat de l'ILPGA.

10.3 M3 (2^{nde} année, 1^{er} semestre) :

10.3.1 Tronc commun

Code P3	Code PX	Unités d'enseignement	crédits
L9TAL	LMSUFT31	Linguistique et TAL 4 enseignements à prendre parmi ceux qui suivent, ou d'autres enseignements d'universités avec lesquelles existent des conventions	12
		SITE Paris X	3
L9T01	LMSDL311	Sémantique et corpus	3
L9T02	LMSDL316	Modélisation des langues : mots et grammaire	3
L9T03	LMSDL382	Formalismes syntaxiques et traitement automatique	3
L9T04	LMDEF333	Document structuré : XML	3
		SITE Paris III	
L9M02	LMSDL381	Analyse du discours et lexicométrie	3
		SITE INALCO	
L9T05	LMSDL383	Écritures et multilinguisme	3
L9T06	LMSDL384	Multimodalité du document numérique	3
L9T07	LMSDL385	Sémantique des textes multilingues	3
	LMSUFT32	Linguistique 3 3 enseignements à prendre dans la liste M3-TAL (<i>cf infra</i>)	9
L9LVE	LMSUCT34	Langue vivante	3
	LMSUCL33	Synthèse de conférences	3
L9T08	LMSDL390	Synthèse de conférences	
L9LIB	LMSULT35	Libre	3

10.3.2 liste M3-TAL :

Code P3	Code PX	Code Inalco	SITE PARIS X	Crédit
L9L01	LMSDL352	-	<i>Acquisition des langues</i>	6
L9L02	LMSDL367	-	<i>Changement des langues. Diachronie du français</i>	6
L9L03	LMSDL343	-	<i>Géopolitique des langues</i>	6
L9L06	LMSDL363	-	<i>Linguistique discursive et textuelle</i>	6
L9L07	LMSDL374	-	<i>Les bizarreries du français</i>	6
L9L13	LMSDL338	-	<i>Phonologie et sciences cognitives</i>	6
Code P3	Code PX	Code Inalco	SITE Paris III³	Crédit
L9444	-	-	<i>Phonétique expérimentale</i>	
F9B7L	-	-	<i>Analyse du discours : variations et interprétations</i>	
F9M6L	-	-	<i>Morphosyntaxe du français et pratiques langagières</i>	
F9M6K	-	-	<i>Analyse de l'oral. Théorie et interprétation</i>	
L9M01	-	-	<i>Catégorisation : lexicale et cognition</i>	
L9LB1	-	-	<i>Terminologie, théories et applications</i>	
L9M04	-	-	<i>L'analyse du discours : un modèle dialogique</i>	
F9L6L	-	-	<i>La connexion syntaxique et la phrase complexe</i>	
L9M33	-	-	<i>Etude comparée des structures des langues bantu et des langues ouest-atlantiques</i>	
L9MGR			<i>Typologie, variation et invariants en morphosyntaxe</i>	
F9B6L			<i>Question de sémantique</i>	
F0C4L			<i>Anaphore et chaîne de références : approches linguistiques, psycholinguistiques et traitements automatiques</i>	
Code P3	Code PX	Code Inalco	SITE Inalco	Crédit

³ Les descriptifs des cours de Paris III dans la liste M3-TAL sont disponibles dans la brochure du **MASTER Spécialité "Langage, Langues, Modèles" de la Mention "Sciences du langage ; Didactique des langues"** à cette adresse : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/Brochure-MASTER-LLM.pdf> ou dans la brochure du **MASTER Spécialité Phonétique** de la même mention et disponible à cette adresse : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/lmd-2005-2006/siteweb/lmdphonetique.htm>. On pourra aussi se renseigner au secrétariat de l'ILPGA.

10.4 M4 (2^{nde} année, 2nd semestre) :

<i>CODE P3</i>	<i>Code PX</i>	<i>Unités d'enseignement</i>	<i>crédits</i>
L0T01	LMSUFT43	Stage en laboratoire	9
L0T01	LMSDL493	stage	
L0T02	LMSUFT41	Mémoire de recherche	15
L0T02	LMSDL492	mémoire	
L0T03	LMSUFT43	Exposé d'articles	6
L0T03	LMSDL493	exposés	

11 Planning des cours du Tronc commun Master TAL recherche

MASTER 1

INALCO

PARISX

PARIS 3

Semestre 1

	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h
Lundi											
Mardi				Trait. Oral P3							
Mercredi		Programmation/Projet PX/P3/INA									
Jeudi							Gestion Info. Multiling. INA				
Vendredi				Analyse Syntax. Auto. PX		BDD linguistes PX	Gram. formelles PX				

Semestre 2

	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h
Lundi							Corpus parallèles et comparables				
Mardi			Synt. formelles P3								
Mercredi		Doc. Struct. P3	Programmation/projet PX/P3/INA								
Jeudi						Sémantique form. PX	Rech. d'information PX				
Vendredi				Statistiques P3							

Ce planning n'intègre pas les enseignements optionnels (Modules Linguistique 1/2, Langue Vivante, Libre)

Les salles sont indiquées dans les descriptifs de chaque cours.

MASTER 2

INALCO

PARISX

PARIS 3

Semestre 1

	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h	19h
Lundi											Synth. de Conf. PX	
Mardi												
Mercredi		Documents structurés XML PX						Ecritures et multiling. INA				
Jeudi								Anal. du disc. et Lexico. P3			SémTxtMulti INA	
Vendredi							Modélisation lang. PX				Form. Synt. Et Trait. Auto.	

Semestre 2

	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h
Lundi											Exposés d'articles PX
Mardi											
Mercredi		Multi. du doc. Num.									
Jeudi											
Vendredi											

Ce planning n'intègre pas les enseignements optionnels (Modules Linguistique 3, Langue Vivante, Libre)

Les salles sont indiquées dans les descriptifs de chaque cours.

12 Descriptif et horaires des cours (1^{ère} et 2^{ème} années), tronc commun

Les descriptifs, horaires et lieux donnés ci-dessous peuvent avoir été modifiés. On obtiendra des renseignements précis et à jour concernant ces cours en s'adressant aux secrétariat des UFRs concernés.

12.1 Grammaires formelles

Sylvain Kahane (Paris X)

Lieu : Paris X, salle L200

Horaire : Vendredi 15h00-17h00

Le cours présente la théorie des langages formels et les grammaires formelles de référence en linguistique : les incontournables grammaires de réécriture de Chomsky, les grammaires lexicalisées avec les grammaires catégorielles, les grammaires de dépendance et les TAG (Grammaire d'adjonction d'arbres), la synchronisation de grammaires avec LFG (Grammaire Lexicale Fonctionnelle). La nature des représentations des énoncés manipulées par ces grammaires sera également abordée : arbres, arbres ordonnés, graphes acycliques.

Bibliographie

Abeillé A., Les nouvelles syntaxes : grammaires d'unification et analyse du français, Armand Colin, 1993.

Kahane S., Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, Actes de TALN 2001, vol. 2, Tours, 2001.

12.2 Analyse syntaxique automatique

Marcel Cori (Paris X)

Lieu : Paris X, salle R15

Horaire : Vendredi 10h30-12h30

On indique quelles sont les principales composantes qui entrent en jeu dans l'analyse syntaxique automatique : formalisme, grammaire, lexicque, traits. On expose les différents types de méthodes (descendantes ou ascendantes, avec retour en arrière, en parallèle ou déterministe). On oppose l'analyse syntaxique théorique à l'analyse robuste, qui opère sur de « vraies » productions langagières, et qui en général reconnaît des fragments de phrase plutôt que des phrases complètes.

12.3 Gestion informatique du multilinguisme

Jean François Perrot (INALCO), Marie-Anne Moreaux (INALCO)

Lieu : INALCO

Horaire : CM 1 semaine sur 2 jeudi de 14h30 à 17h30.

TD 1 semaine sur 2 Mardi de 14h à 17h.

Ce cours sera centré sur le substrat informatique et webographique en cause dans les problèmes de représentation, codage et transmission de l'information multilingue.

L'objectif est de permettre l'acquisition et la pratique des connaissances nécessaires à l'échange réussi de documents numériques multilingues provenant de machines, plateformes et formats différents.

12.4 Traitement de l'oral / Informatique et phonétique

Cédric Gendrot (Paris 3)

Lieu : ILPGA, LaboC

Horaire : mardi 10h30-12h

3 séances : Introduction à la phonétique. Introduction aux différents types de synthèses de la parole (synthèse totale par règles, synthèse par concaténation) – synthèse par concaténation de diphones avec intégration de règles phonétiques.

3 séances : utilisation d'un phonétiseur écrit en Perl (conversion graphèmes -> phonèmes au moyen de règles phonologiques) - Amélioration de la synthèse Text-To-Speech grâce aux Speech Markup languages (XML, SSML, SAPI, JSML).

3 séances : présentation et utilisation de SMIL (Synchronized Multimedia Integration Language). Ce langage, très proche du XML de part sa structure, permet de créer du contenu multimédia sur Internet, mixant texte, son, vidéo... Ces spécifications ont été travaillées par le W3C.

Ce cours vise à présenter la synthèse de la parole actuellement utilisée par les serveurs vocaux ainsi que l'intégration de l'audio/vidéo sur les pages Internet. Cette optique phonétique permet de passer en revue différents langages inspirés de XML utilisés pour l'amélioration de la synthèse ou pour la mise en parallèle de contenus multimédia.

Bibliographie

Dutoit, T. (1997), High-Quality Text-to-Speech Synthesis : an Overview, Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis, vol. 17 n°1, pp. 25-37.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vrecken, O. (1996) The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396.

Harold E.R., W. S. M. (2001). XML in a nutshell, O'REILLY.

Speech Synthesis Markup Language : SSML (2004) <http://www.w3.org/TR/speech-synthesis/>

JSpeech Markup Language: JSML (2000) <http://www.w3.org/TR/2000/NOTE-jsml-20000605/>

12.5 Programmation et projet encadré (semestre 1)

Jean Michel Daube/Rachid Belmouhoub (INALCO), Serge Fleury (Paris III)

Lieu : ILPGA, LaboC

Horaire : Mercredi 9h00-12h00

Il s'agit d'apprendre à mettre en œuvre une chaîne de traitement textuel semi-automatique,

depuis la récupération des données jusqu'à leur utilisation. Ce cours posera d'abord la question des objectifs linguistiques à atteindre (lexicologie, recherche d'information, traduction...) et fera appel aux méthodes et outils informatiques nécessaires à leur réalisation (récupération de corpus, normalisation des textes, segmentation, étiquetage, extraction, structuration et présentation des résultats...). Ce cours sera aussi l'occasion d'une évaluation critique des résultats obtenus, d'un point de vue quantitatif et qualitatif.

12.6 Bases de données pour linguistes

J.L Milnel (Paris X)

Lieu : Paris X, salle L200

Horaire : Vendredi 13h00-15h00

La description d'une réalité langagière se fait souvent « à la main » : corpus saisi sous traitement de texte, observations faites également sous traitement de texte. Les bases de données relationnelles correspondent à la réalisation informatique de « fiches » sur un sujet quelconque. Par rapport à des fiches papier, on gagne la possibilité de disposer d'un très grand nombre de fiches, comprenant des informations très précises et très riches, triables de multiples façons et entre lesquelles on peut faire des liens.

L'accent est mis sur la démarche de création et d'utilisation de bases de données pour des phénomènes langagiers. La moitié du cours se déroule cependant sur machine, avec utilisation d'un logiciel de gestion de bases de données. Le cours ne suppose néanmoins pas de connaissances informatiques préalables.

12.7 Statistique et analyse multidimensionnelle

André Salem (Paris III)

Lieu : ILPGA, salle LaboC

Horaire : Vendredi 11h-12h30

Les approches quantitatives des corpus textuels sont présentées (historique et tendances récentes). Est abordée la question des unités pour la statistique textuelle (formes, lemmes, segments répétés, cooccurrences). La compréhension d'un certain nombre de propriétés statistiques des textes (Zipf-Pareto, courbes d'accroissement du vocabulaire) fournit le cadre d'analyse des constats effectués. Le cours introduit également aux méthodes de statistiques appliquées aux données textuelles : indices, distances, approches multidimensionnelles.

12.8 Corpus parallèles/comparables

Pierre Zweigenbaum (INALCO)

Lieu : INALCO, 9bis, avenue de la Belle Gabrielle, 75012 PARIS

Horaire : Vendredi de 14h à 17h (les dates précises des cours seront diffusées sur la liste plural)

Ce cours vise à expliciter les objectifs sous-jacents à l'établissement de corpus parallèles (où des textes sont en rapport de traduction) et à exposer les techniques linguistiques et informatiques mises en œuvre pour réaliser un alignement à différents paliers du document (paragraphe, phrase, mot). A partir des limites des corpus parallèles, on expliquera le recours aux corpus comparables (traitant du même domaine et relevant des mêmes

genres), et les outils de traitement associés.

12.9 Recherche et extraction d'information

(Paris X)

Lieu :

Horaire :

La recherche d'information est à la base des moteurs de recherche actuels. Elle recouvre les mécanismes et les modèles permettant de rapatrier les documents correspondant à une requête. L'extraction d'information produit des réponses plus petites qu'un document, répondant de manière plus précise aux demandes d'un utilisateur. Elle est à la base par exemple des systèmes retournant des réponses à des questions factuelles (« En quelle année la France a-t-elle gagné la Coupe du Monde de Football ? »)

12.10 Sémantique formelle

Sarah de Vogué (Paris X)

Lieu : Paris X, salle ...

Horaire : Jeudi 13h-15h

12.11 Syntaxes formelles

Pollet SAMVELIAN (Paris III)

Ce séminaire se donne un double objectif :

1. Familiariser les étudiants avec un formalisme grammatical particulier, en l'occurrence le modèle HPSG (Head-Driven Phrase Structure Grammar, tr. fr.: grammaires syntagmatiques guidées par les têtes).
2. Mener une réflexion générale, par le biais de la comparaison entre le modèle présenté et l'une des variantes du modèle génératif-transformationnel (P&P), sur les différentes possibilités d'analyses pour un même phénomène syntaxique, selon les caractéristiques du modèle que l'on adopte.

HPSG appartient au paradigme des grammaires dites d'unification. Ces dernières, nées dans les années quatre-vingts, constituent une alternative au modèle génératif-transformationnel dans ses diverses variantes. Malgré leurs différences plus ou moins importantes, elles partagent généralement les caractéristiques suivantes, qui en font des outils adaptés pour l'implémentation en TAL :

- Absence de tout mécanisme transformationnel (donc de mouvement).
- Réhabilitation des descriptions linguistiques proches de la structure dite de surface.
- Recours au mécanisme d'unification de traits.

Les premières séances seront consacrées à la mise en place des notions communes à la plupart de ces modèles. On passera ensuite au modèle HPSG en abordant le(s) traitement(s) proposé(s) dans ce cadre pour les phénomènes suivants : montée et contrôle, dépendances à distance et extraction, interface morphologie-syntaxe (traitement des clitiques).

Pour le plan détaillé, l'organisation des séances et la bibliographie commentée se référer à :

http://www.cavi.univ-paris3.fr/ilpga/pages_personnelles/samvelian/index.htm

12.12 Document structuré

Serge Fleury (Paris III)

Lieu : ILPGA, salle LaboC

Horaire : Mercredi 9h00-10h30

Les textes sont des documents structurés. Un article comporte un titre, un ou des auteur(s), des sections, une bibliographie. La présentation permet d'appréhender cette structure (taille des caractères, jeu sur le gras, etc.). Lorsqu'on rend explicite cette structure (par le moyen de balisages en XML), on peut manipuler le texte comme unité structurée (extraire automatiquement les titres pour une table des matières, chercher les paragraphes introductifs, etc.). Le cours présente la manière de rendre explicite et fiable la structure des documents (en les assortissant d'une « grammaire textuelle » dite DTD). Il aborde les transformations réglées de textes qui deviennent possibles.

Bibliographie

P. Bonhomme, « Codage et normalisation de ressources textuelles », *in* Ingénierie des langues, J.-M. Pierrel (ed), p. 173-192, Hermès Science, 2000, Paris.

Ressources fournies

Polycopié et outils sur pages WEB du cours :

<http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/doc-structure.htm>

<http://e-ilpga.univ-paris3.fr/campus/>

12.13 Programmation et projet encadré (semestre 2)

Jean Michel Daube/Jean-Michel Daube (INALCO), Serge Fleury (Paris III)

Lieu : ILPGA, salle LaboC

Horaire : Mercredi 11h30-14h30

Cf descriptif du premier semestre.

12.14 Modélisation des langues : mots et grammaire

Sylvain Kahane (Paris X)

Lieu : Paris X, salle L...

Horaire : Vendredi 13h00-15h00

L'objectif est de présenter un modèle d'une langue naturelle, c'est-à-dire un dispositif permettant de simuler un sujet parlant, du sens qu'il souhaite communiquer au son qu'il produit (et notamment la prosodie). Nous aborderons la question des unités linguistiques élémentaires (mots, morphèmes, unités lexicales, signes linguistiques) et la question des différents types d'organisation de ces unités (organisation discursive et structure communicative, structure prédicat-argument, dépendance syntaxique, constituants topologiques, constituants prosodiques). Nous construirons ensemble un fragment de modèle pour le français et nous verrons comment lexicque et grammaire s'articulent. Ce modèle sera contrasté avec d'autres approches comme les grammaires d'unification et les grammaires lexicalisées. Tous les outils mathématiques utilisés seront introduits et motivés par des questions théoriques.

Bibliographie

- BRESNAN Joan, 2001, Lexical-Functional Syntax, Blackwell.
 CREISSELS Denis, 1995, Elements de syntaxe générale, PUF.
 KAHANE S., 2002, Grammaire d'Unification Sens-Texte : vers un modèle mathématique articulé de la langue, Université Paris 7, 82 p., <http://panini.u-paris10.fr/kahane>
 MEL'CUK I., 1997, Vers une linguistique Sens-Texte, Leçon inaugurale au Collège de France, 78 p.
 POLGUERE Alain, 2003, Lexicologie et sémantique lexicale, Presses de l'Université de Montréal
 SAG I. & WASOW T., 1999, Syntactic theory; A Formal Introduction, CSLI Publications, Stanford.
 TESNIÈRE L. 1959, Elements de syntaxe structurale, Klincksieck.

12.15 Formalismes syntaxiques et traitement automatique

M. Cori, M. Desmets (Paris X)

Lieu : Paris X, salle L...

Horaire : Vendredi 17h30-19h30

12.16 Document structuré : XML

Jean Paul Jorda (Paris X)

Lieu : Paris X, salle ...

Horaire : Mercredi 9h-12h

L'utilisation informatique aisée de documents (documentations techniques, mise en ligne sur le Web, archivage) implique désormais qu'ils soient structurés explicitement. On présente les techniques permettant de passer de textes faiblement ou pas structurés à des textes structurés, permettant de vérifier que des textes structurés obéissent bien à la structure souhaitée. Les transformations réglées d'un texte relevant d'une structure pour qu'il obéisse à une autre (passage d'un livre à une version sur la Toile, qui suppose des découpages fins et des liens) sont présentées et mises en œuvre.

12.17 Analyse du discours et lexicométrie

André Salem (Paris 3)

Lieu : ILPGA, salle Benveniste

Horaire : jeudi 15h-17h30

12.18 Écritures et multilinguisme

M. Bottin (INALCO)

Lieu : INALCO

Horaire : Mercredi 14h-16h

Ce cours a pour objectif de sensibiliser les étudiants à la diversité des langues et de leurs écritures, en vue de ne pas plaquer les méthodes et outils informatiques d'une aire linguistique (monde anglo-saxon en particulier) sur une autre. On abordera en particulier les types d'écriture, les directions d'écriture, les problèmes de ligatures, majuscules, diacritiques, ponctuation, segmentation ... pour bien prendre la mesure des problèmes liés au multilinguisme.

12.19 Multimodalité du document numérique

P. Stockinder, D. Piotrowski (INALCO)

Lieu : INALCO

Horaire : 10h-12h

Le document numérique n'est plus que faiblement textuel. Il associe également des images, du son. Il repose sur des interfaces d'accès qui influent sur sa compréhension et son utilisation. Le cours introduit à la dimension sémiotique de ces fonctionnements ainsi qu'à leur support technique.

12.20 Sémantique des textes multilingues

M. Slodzian, E. Bourion, M. Valette, C. Podat (INALCO)

Lieu : INALCO

Horaire : 18h00-20h00

Le passage d'une langue à l'autre est filtré par les usages et les cultures associées à chacune des langues. De la structuration globale des textes aux séquences renvoyant à des entités du monde (institutions, événements), la sémantique doit rendre compte de cette dimension.

12.21 Synthèse de conférences

Lieu : Paris X, salle ...

12.22 Exposés d'articles

Lieu : Paris X, salle ...