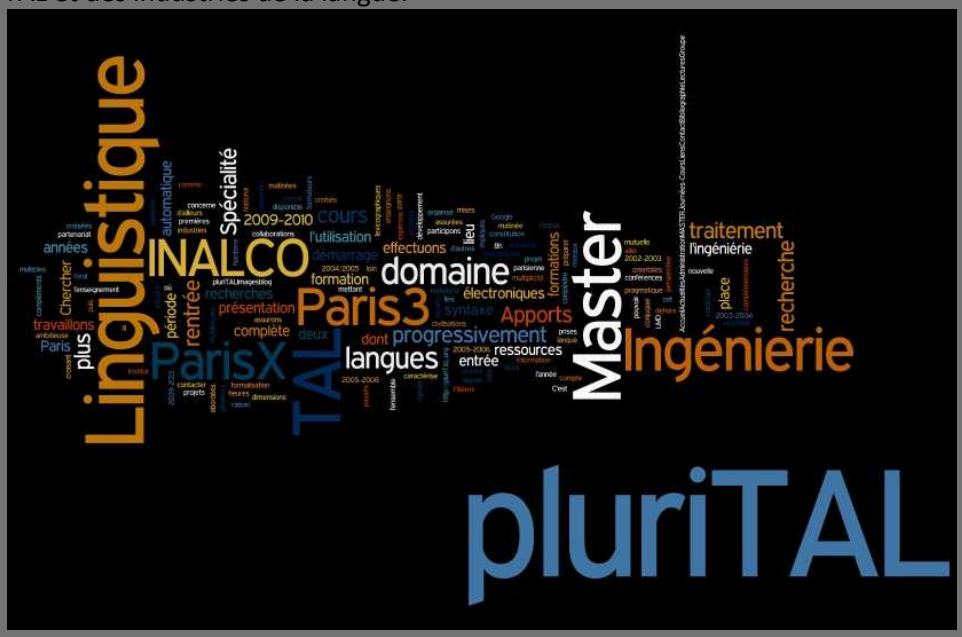
pluriTAL

MASTER Traitement Automatique des Langues

Université Paris Nanterre, Université Sorbonne nouvelle Paris 3, INALCO

La mention TAL concerne la recherche et le développement dans le domaine du TAL et des industries de la langue.



http://plurital.org

Formation proposée par
l'Université Paris Nanterre
l'Université Sorbonne nouvelle Paris 3
l'INALCO

MASTER

Mention : *Traitement Automatique des Langues*

Responsables:

Marie-Anne Moreaux (INALCO)
Sylvain Kahane (Paris Nanterre)
Serge Fleury (Sorbonne nouvelle Paris 3)

Commission pédagogique :

Jean-Michel Daube (INALCO)
Serge Fleury (Sorbonne nouvelle Paris 3)
Sylvain Kahane (Paris Nanterre)

Hypertoile du MASTER : http://plurital.org (désormais « le site pluriTAL»)

Présentation du Master *TAL*

Le diplôme est délivré par les 3 partenaires suivants :

Université PARIS NANTERRE
Université SORBONNE NOUVELLE, PARIS 3
INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES (INALCO)

La formation s'appuie sur les laboratoires : Paris Nanterre - MODYCO, *Modèles, Dynamiques, Corpus*, (UMR 7114), Paris 3 - CLESTHIA *Langage, systèmes, discours* (EA 7345), Paris 3 - LPP *Laboratoire de Phonétique et Phonologie* (UMR 7018), INALCO - ER-TIM *Textes, Informatique, Multilinguisme* (EAD 2540).

La mention *TAL* concerne la recherche et le développement dans le domaine du TAL et des industries de la langue. L'ingénierie linguistique 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 (écrits ou oraux) 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.

L'objectif de la formation est de donner à des étudiants issus de cursus de langues ou de sciences du langage des bases solides qui leur permettent de s'orienter vers les métiers de l'ingénierie linguistique et du TAL, et de les laisser choisir entre diverses perspectives : document électronique, ingénierie multilingue, traductique. Il s'agit aussi de permettre à certains d'entre eux d'opter pour la recherche et le développement en ce domaine.

Sommaire

Contacts

Commission pédagogique

Daube Jean-Michel (jean-michel.daube@inalco.fr)
Fleury Serge (serge.fleury@univ-paris3.fr)
Kahane Sylvain (sylvain@kahane.fr)

Secrétariat administratif

Inalco :
Clémence MILCARECK, 65 rue des grands moulins, 75013 Paris, 01.80.71.11.36, clemence.milcareck@inalco.fr

Sorbonne Nouvelle, Paris 3 (ILPGA) :
Nelly LAPAIX, ILPGA, 2ème étage, 01.44.32.05.81, Nelly.Lapaix@univ-paris3.fr

Paris Nanterre :
Janine BIANCHI, Bureau L-114, 01.40.97.70.75, jbianchi@u-paris10.fr

Réunions de rentrée – Début des cours

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

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

Liste PluriTAL (liste de diffusion)

Inscription **obligatoire** pour tous les étudiants devant suivre des cours du Master T.A.L.

Voir la page « Liste pluriTAL » sur la page web du MASTER (site pluriTAL).

Lien direct : http://plurital.org/groupepluriTAL.html

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

AFTAL

L'AFTAL est dédiée à l'aide à l'insertion professionnelle pour les étudiants issus des formations en SdL/TAL. Nos 3 missions :

- *vous donner un retour d'expérience concret (rencontres avec les anciens, questions que vous posez au groupe)*
- *vous accompagner dans votre insertion professionnelle (conseils, aide au ciblage de candidature)*
- *vous faciliter la prise de contact avec les acteurs de l'industrie (réseau de contacts dans la Recherche et dans l'Entreprise)*

mail : aftal.asso@gmail.com
twitter : @AssoForTAL
LinkedIn : Groupe AFTAL

Localisations

Paris Nanterre

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.

Paris 3 / ILPGA

ILPGA / Université Sorbonne nouvelle Paris 3,
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

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 7 - 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

Les bâtiments de l'**ERTIM** se situent au 2 rue de Lille - 75007 Paris

L'INALCO est aussi dans le 13ème, informations détaillées ici :
http://www.inalco.fr/ina_gabarit_rubrique.php3?id_rubrique=3005

Inscriptions

Inscription en 1ère année

1ère étape de l'inscription en M1

L'étudiant devra être titulaire d'une licence.

Nous accueillons notamment les étudiants issus des 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 »).

Mais aussi des étudiants pouvant faire état d'une appétence pour l'informatique, ou les étudiants en informatique ou mathématiques ayant manifesté de l'intérêt pour les langues ou la linguistique verront notamment leur dossier examiné avec intérêt.

Vous devez contacter par courriel les 3 responsables pédagogiques avant de vous inscrire : sylvain@kahane.fr , serge.fleury@univ-paris3.fr , jean-michel.daube@inalco.fr

2ème étape de l'inscription en M1

En fonction de la réponse de la commission pédagogique, il faudra vous inscrire administrativement dans l'un des 3 établissements (**et un seul**) :

- Inscription Paris 3 - Sorbonne Nouvelle :

http://ecandidat.univ-paris3.fr/

- Inscription Paris Nanterre :

http://ecandidat.u-paris10.fr

- Inscription à l'INALCO

Contact : jean-michel.daube@inalco.fr

Inscription en 2ème année

L'admission en 2ème année se fait sur dossier y compris pour les étudiants reçus en master 1. Les étudiants peuvent ainsi déposer plusieurs dossiers de demande d'admission.

L'inscription en 2ème année nécessite une **formation initiale en T.A.L, linguistique et informatique** équivalente à celle de la première année du master TAL (*cf* liste des cours du M1).

Inscription en M2 parcours « Recherche & Développement » (parcours disponible dans les 3 établissements)

Pour le parcours « *Recherche et Développement* », il est demandé aux étudiants d'obtenir l'accord d'un directeur de recherche qui supervisera leur mémoire. **Vous devez contacter par courriel les 3 responsables pédagogiques avant de vous inscrire** : sylvain@kahane.fr , serge.fleury@univ-paris3.fr , jean-michel.daube@inalco.fr

En fonction de la réponse de la commission pédagogique, il faudra vous inscrire administrativement dans l'un des 3 établissements (et un seul) :

- Inscription Paris 3 - Sorbonne Nouvelle :

http://ecandidat.univ-paris3.fr/

- Inscription Paris Nanterre :

http://ecandidat.u-paris10.fr

- Inscription à l'INALCO

Contact : jean-michel.daube@inalco.fr

Inscription en M2 parcours « Ingénierie Linguistique-Document Electronique et Flux d'information » (IL-DEFI) (parcours disponible uniquement à Paris Nanterre)
Lien : http://ecandidat.u-paris10.fr
Contact : Delphine Battistelli (delphine.battistelli@u-paris10.fr)

Inscription en M2 parcours « Ingénierie Multilingue » (parcours disponible uniquement à l'INALCO)
Contact : jean-michel.daube@inalco.fr

Inscription en M2 parcours « Traductique » (parcours disponible uniquement à l'INALCO)
Contact : jean-michel.daube@inalco.fr

Inscriptions pédagogiques

A l'issue de vos inscriptions administratives, vous devrez effectuer une inscription pédagogique auprès du secrétariat de votre établissement administratif.

Inscriptions pédagogiques :

PARIS 3 (*cf* **secrétariat ILPGA**)	**PARIS X** (*cf* **secrétariat Paris Ouest**)	**INALCO** (*cf* **secrétariat INALCO**)

Formation continue

Le master est aussi ouvert en formation continue pour les traducteurs, documentalistes, bibliothécaires, gestionnaires de sites Web, employés du tertiaire soucieux de se former à des technologies innovantes détenteurs d'une licence ou d'un équivalent par validations d'acquis

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.

Modalités de Contrôle des connaissances (provisoire)

Les enseignements obéissent à la règle du contrôle continu (CC). C'est le régime obligatoire. **Il exige l'assiduité**.

Il s'effectue sous forme d'épreuves évaluées tout au long du semestre : travaux personnels, exposés, partiels en fin de semestre sous la responsabilité de l'enseignant. Leur nature est déterminée par chaque enseignant (oral, dossier, écrit…).

Les enseignements relevant du contrôle continu ne font pas l'objet de dates spécifiques d'examens (inclus dans la durée de l'enseignement) ni de dates spécifiques de rattrapage.

L'organisation du rattrapage de tous les cours est à la charge de l'enseignant qui doit programmer avec le/les étudiant(s) la date de ce rattrapage.

L'enseignant se réserve le droit de ne pas autoriser l'étudiant à rattraper son séminaire si celui-ci ne s'est présenté à aucun de ses cours.

Pour être dispensé(e) d'assiduité et bénéficier de l'inscription en dérogatoire vous devez vous inscrire pédagogiquement auprès du secrétariat, prendre contact avec l'enseignant au début des cours et avoir son accord. **Cette procédure est obligatoire**.

Attention : en Master 1, la moyenne du premier semestre **ne compense pas** celle du second si celle-ci est en-dessous de 10/20.

En d'autres termes, chaque semestre est indépendant l'un de l'autre au regard des moyennes obtenues. En revanche, la moyenne des modules est compensable.

En outre, un étudiant ayant obtenu une note inférieure à 10 à son mémoire ou son rapport de stage n'est pas admis, même si sa moyenne générale est supérieure à 10.

Calendrier 2017-2018

PREMIER SEMESTRE	DEUXIEME SEMESTRE
Journée d'accueil Mercredi 13 septembre **Début des enseignements** Lundi 18 septembre **Vacances de Toussaint** *Cf* calendrier universitaire de chaque établissement **Fin des enseignements** *Cf* calendrier universitaire de chaque établissement **Vacances de Noël** *Cf* calendrier universitaire de chaque établissement	**Début des enseignements** *Cf* calendrier universitaire de chaque établissement **Vacances d'hiver** *Cf* calendrier universitaire de chaque établissement **Vacances de printemps** *Cf* calendrier universitaire de chaque établissement **Fin des enseignements** *Cf* calendrier universitaire de chaque établissement
Jurys de la première session : *Cf* calendrier universitaire de chaque établissement	
Examens de la deuxième session : *Cf* calendrier universitaire de chaque établissement	
Jurys de la seconde session : *Cf* calendrier universitaire de chaque établissement	

La mention T.A.L

La mention T.A.L concerne la recherche et le développement dans le domaine du TAL et des industries de la langue. L'ingénierie linguistique 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 différents domaines des sciences du langage (phonétique et phonologie, morphologie, syntaxe et sémantique) ;
2. Connaître les bases de la recherche et de l'extraction d'information, de la constitution et de la gestion de corpus (écrits ou oraux) 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 de recherche particulier ;
3. Exprimer les règles et les régularités à l'œuvre dans les corpus, par le biais des grammaires formelles et des traitements quantitatifs pour savoir passer d'une description linguistique d'un texte à une représentation plus formelle permettant sa prise en charge par des logiciels.

L'objectif de la formation est de donner à des étudiants issus des cursus de langues ou de sciences du langage des bases solides qui leur permettent de s'orienter vers les métiers de l'ingénierie linguistique, et de leur donner les possibilités de choisir entre diverses perspectives : document électronique, ingénierie multilingue, traductique. Il s'agit aussi de permettre à certains d'entre eux d'opter pour la recherche et le développement en ce domaine.

Un partenariat universitaire pour le TAL (pluriTAL)

Le diplôme est délivré par les 3 partenaires suivants :

- Université Paris Nanterre
- Université Sorbonne nouvelle Paris 3
- Institut National des Langues et Civilisations Orientales (INALCO)

La formation s'appuie sur les laboratoires :

Paris Nanterre - **MODYCO** (Modèles, Dynamiques, Corpus, UMR 7114),

http://www.modyco.fr/

Paris 3 - **CLESTHIA** Langage, systèmes, discours - EA 7345,

http://www.univ-paris3.fr/clesthia-langage-systemes-discours-ea-7345-98241.kjsp

Paris 3 – **LPP** (UMR 7018) Laboratoire de Phonétique et Phonologie

http://lpp.univ-paris3.fr/

INALCO - **ER-TIM** (EAD 2540) : Équipe de Recherche « Textes, Informatique, Multilinguisme »

http://www.crim.fr/

Objectifs d'apprentissage

Objectifs d'apprentissage du master en termes de connaissances (connaissances disciplinaires, connaissances pluridisciplinaires sur l'objet étudié, connaissances méthodologiques, connaissances linguistiques, …)

- Savoirs disciplinaires en linguistique (en complément de bases solides en phonétique/phonologie, morphologie, syntaxe et sémantique) : sémantique formelle, sémantique lexicale, systèmes d'écriture, traductologie, traductique ;
- Savoirs en TAL : grammaires formelles, syntaxe formelle, analyse syntaxique automatique, gestion du multilinguisme, statistique et analyse multidimensionnelle, traitement de l'oral, recherche et extraction d'information, corpus alignés ;
- Savoirs en informatique : programmation et algorithmique spécifique, bases de données, document structuré (XML) ;
- Maîtrise en réception puis en production de l'anglais scientifique.

Objectifs d'apprentissage du master en termes de compétences

Savoir s'intégrer dans un projet collectif multi-disciplinaire :

- comprendre sa contribution spécifique dans le projet ;
- transmettre de manière claire son apport (outils de formalisation) ;
- assurer les coordinations nécessaires.

Objectifs d'apprentissage du master en termes de compétences métier

- Technologies et méthodes de conception et développement : bases de données relationnelles, normes et outils pour documents structurés, conception de produits informationnels ;
- Connaissances des produits et outils industriels en gestion d'information et en traitement des documents.
- Capacité de maîtriser la gestion de projets
- Traitement du document numérique

Débouchés

Métiers auxquels le master permet d'accéder directement

Ingénieur linguiste, terminologue, lexicologue, gestionnaire de site web multilingue, lexicologue, chef de projet multimédia, traducteur, veilleur (économique, stratégique, technologique), chef de projet multimédia, documentaliste spécialisé ou responsable de service de documentation, architecte de système d'information ou responsable d'études informatiques

Code	Intitulé
32213	Webmaster
32214	Documentaliste spécialisé(e) (dans un domaine) ou Responsable du service documentation
32241	Traducteur
32321	Ingénieur de la connaissance
32331	Chef de projet Internet ou multimédia
32341	Architecte système d'information ou Responsable d'études informatiques
35152	Lexicologie, terminologie

Relation avec les milieux professionnels :

EDF, Mondeca, Temis, Arisem, XEROX, PUF, Larousse, Le Robert, Performix, Syllabs, Quensis, Logosapience, Exalead, Thales, France Telecom, Limsi, Bowne Global Solutions, Softissimo, TRADOS, SDL, LIP6, ATILF

Organisation globale des enseignements du master

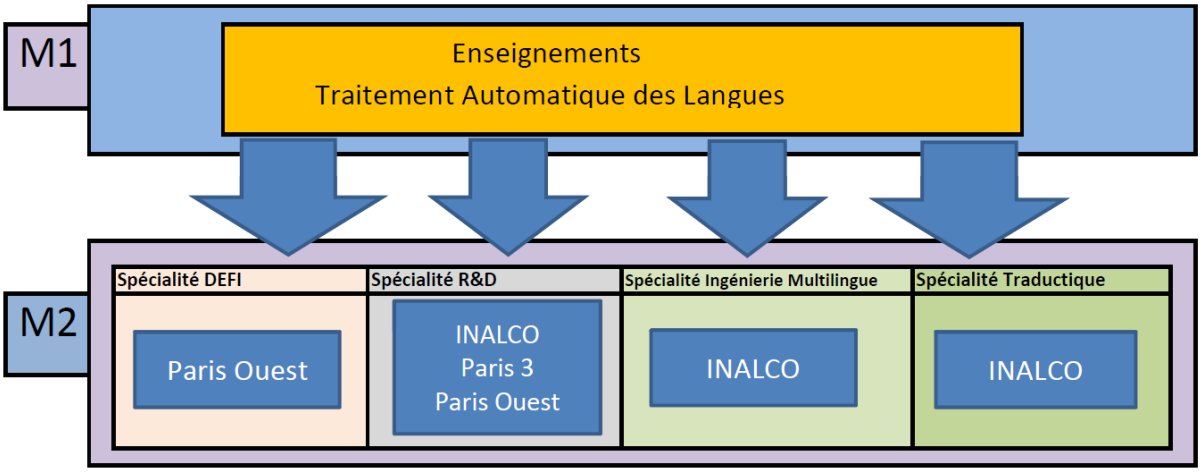
Le master s'organise selon quatre parcours, adossés chacun à un ou plusieurs établissements universitaires :

- le parcours *IL-DEFI* (Ingénierie Linguistqiue - Documents Electroniques et Flux d'Informations), basé à Paris Nanterre ;

- le parcours *Ingénierie multilingue*, basé à l'INALCO ;

- le parcours *Traductique*, également basé à l'INALCO ;

- le parcours *Recherche et Développement*, basé dans les trois établissements.



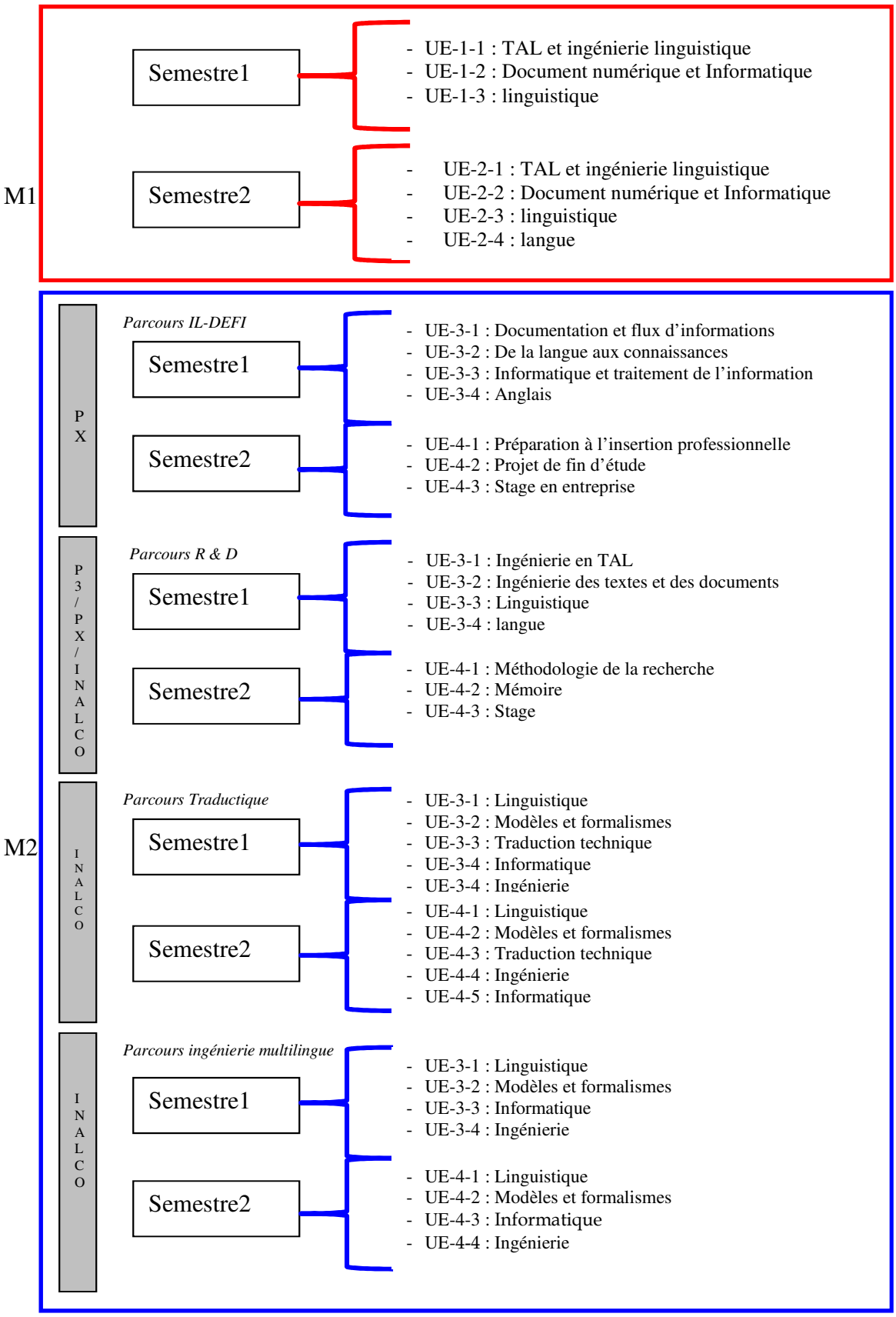
La première année, les UE sont identiques entre les quatre parcours (*cf.* tableau ci-dessous). Les EC qui composent les UE sont soit des EC communes, proposées à l'ensemble des étudiants, soit des EC au choix qui peuvent être pris entre l'un ou l'autre des établissements.

En deuxième année, les parcours se diversifient nettement.

Le parcours *IL-DEFI* et le parcours *Recherche et développement* ont en commun au premier semestre les enseignements d'une UE (celle nommée « Ingénierie des textes et des documents » dans le parcours *Recherche et développement* et celle nommée « De la langue aux connaissances » dans le parcours *IL-DEFI*). Ils se distinguent par les UE : « Documentation et flux d'informations » et « Informatique et traitement de l'information » en *IL-DEFI*, « Ingénierie linguistique et TAL » et « Linguistique » en *Recherche et développement*. Le deuxième semestre est entièrement différencié, puisque notamment dans le parcours *Recherche et développement* il y a obligatoirement un mémoire de recherche.

Enfin, les parcours *Ingénierie multilingue* et *Traductique* se distinguent en deuxième année par des enseignements de « Techniques de traduction » pour ce dernier parcours, ce qui a pour répercussion que les UE « Linguistique », « Modèles et formalismes », « Informatique » et « Ingénierie » ont un moindre poids dans ce parcours.

Tableau synthétique des différents parcours en M1 et M2



MASTER 1ère année

4 parcours :

PRO parcours IL-DEFI (noté **D**)
REC parcours RECHERCHE et DEVELOPPEMENT en TAL (noté **R**)
PRO parcours TRADUCTIQUE (noté **T**)
PRO parcours INGENIERIE MULTILINGUE (noté **I**)

Semestre 1

ECTS	parcours **D**	parcours **R**	parcours **I**	parcours **T**	ECTS
15	TAL ET INGIENIERIE LINGUISTIQUE 1				15
3 3 3	Gestion informatique du multilinguisme Informatique et phonétique Programmation et projet encadré 1				
3 3	Modélisation linguistique pour l'analyse automatique de texte Grammaire formelle		Principes des bases de données 1 Langages réguliers et langages hors contexte1		
6	DOCUMENT NUMERIQUE ET INFORMATIQUE 1				6
3	Bases de données pour linguiste				
3	Programmation et algorithmique 1		Représentation et algorithmes Programmation et abstraction des données 1 Remise à niveau info		
9	LINGUISTIQUE 1				9
3 3 3	Syntaxe formelle Machine creativity and text generation 3 ECTS de linguistique		Compétence avancée en langue Lexique et morphologie Pratiques textuelles et traduction		

4 parcours :

PRO parcours IL-DEFI (noté **D**)
RECH parcours RECHERCHE et DEVELOPPEMENT en TAL (noté **R**)
PRO parcours TRADUCTIQUE (noté **T**)
PRO parcours INGENIERIE MULTILINGUE (noté **I**)

Semestre 2

ECTS	parcours **D**	parcours **R**	parcours **I**	parcours **T**	ECTS
12	TAL ET INGIENIERIE LINGUISTIQUE 2				12
3	Statistique et analyse multidimensionnelle				
3	Programmation et projet encadré 2				
3	Outils de traitements de corpus				
3	Corpus parallèles et comparables				
			Langages réguliers et langages hors contexte2		
9	DOCUMENT NUMERIQUE ET INFORMATIQUE 2				9
3	Recherche et extraction d'information				
3	Document structuré				
3	Programmation et algorithmique 2		Principes des bases de données 2 Remise à niveau info Programmation logique Programmation et abstraction des données 2		
6	LINGUISTIQUE 2				6
3	3 ECTS de linguistique		Sémantique textuelle		
3	Introduction à la fouille de textes		Techniques de traduction		
3	LANGUE				3
3	Langue		Compétence avancée en langue		

Master 1, tous parcours Paris 3, Paris Nanterre

Semestre 1		
TAL et ingénierie linguistique 1	**15**	
Grammaires formelles	3	Paris Nanterre
Modélisation linguistique pour l'analyse automatique de texte	3	Paris Nanterre
Informatique et phonétique	3	Paris 3
Gestion informatique du multilinguisme	3	INALCO
Programmation et Projet encadré 1	3	Paris 3/ INALCO
Linguistique 1	**9**	
Machine creativity and text generation	3	Paris 3
Syntaxe formelle	3	Paris Nanterre
(+ 3 crédits de linguistique)	3	
Document numérique et Informatique 1	**6**	
Bases de données pour linguistes	3	Paris Nanterre
Programmation et algorithmique 1	3	Paris Nanterre
Semestre 2		
TAL et ingénierie linguistique 2	**12**	
Statistique et analyse multidimensionnelle	3	Paris 3
Corpus parallèles et comparables	3	INALCO
Programmation et Projet encadré 2	3	Paris 3/ INALCO
Outils de traitement de corpus	3	INALCO
Linguistique 2	**6**	
(+ 3 crédits de linguistique)	3	Paris 3
Introduction à la fouille de textes	3	Paris 3
Document numérique et Informatique 2	**9**	
Document structuré	3	Paris 3
Programmation et algorithmique 2	3	Paris 3
Recherche et extraction d'information	3	Paris Nanterre
Langue	**3**	Paris 3 / Paris Nanterre / INALCO

Master 1, INALCO

Semestre 1		
TAL et ingénierie linguistique 1	**15**	
Langages réguliers et langages hors contexte 1		INALCO
Informatique et phonétique		Paris 3
Gestion informatique du multilinguisme		INALCO
Programmation et Projet encadré 1		Paris 3/ INALCO
Principes des bases de données		INALCO
Linguistique 1	**9**	
Pratique textuelle		INALCO
Lexique et morphologie		INALCO
Compétence avancée en langue		INALCO
Document numérique et Informatique 1	**6**	
Bases de données pour linguistes		Paris Nanterre
Représentations et algorithmes		INALCO
Programmation et abstraction des données 1		INALCO
Remise à niveau info		INALCO
Semestre 2		
TAL et ingénierie linguistique 2	**12**	
Statistique et analyse multidimensionnelle		Paris 3
Corpus parallèles et comparables		INALCO
Programmation et Projet encadré 2		Paris 3/ INALCO
Outils de traitement de corpus		INALCO
Langages réguliers et langages hors contextes 2		INALCO
Linguistique 2	**6**	
Pratique textuelle		INALCO
Lexique et morphologie		INALCO
Document numérique et Informatique 2	**9**	
Document structuré		Paris 3
Recherche et extraction d'information		INALCO
Programmation logique		INALCO
Programmation et abstraction des données 2		INALCO
Remise à niveau info		INALCO
Langue	**3**	
Compétence avancée en langue		INALCO

Code des enseignements du M1 (semestre 1)

Code EC Paris Nanterre	Code EC Paris 3	Code EC INALCO	Unité D'enseignement
			TAL ET INGIENIERIE LINGUISTIQUE 1
3LTA701T	L7TIL3		Gestion informatique du multilinguisme
3LTA703T	L7TIL4		Informatique et phonétique
3LTAY04T	L7TIL2		Modélisation linguistique pour l'analyse automatique de texte
3LTA705T	L7TIL5		Programmation et projet encadré 1
3LTA702T	L7TIL1		Grammaires formelles
			Principes des bases de données
			Langages réguliers et langages hors contexte 1
			DOCUMENT NUMERIQUE ET INFORMATIQUE 1
3LTA709T	L7DNI2		Bases de données pour linguistes
3LTA710T	L7DNI1		Programmation et algorithmique 1
			Représentations et algorithmes
			Programmation et abstraction des données 1
			Remise à niveau info
			LINGUISTIQUE 1
3LTA706T	LYSL05		Machine creativity and text generation
3LTA708T	L7LIN2		Syntaxe formelle
			Pratique textuelle
			Lexique et morphologie
			Compétence avancée en langue

Code des enseignements du M1 (semestre 2)

Code EC Paris Nanterre	Code EC Paris 3	Code EC INALCO	Unité D'enseignement
			TAL ET INGIENIERIE LINGUISTIQUE 2
3LTA805T	L8TIL1		Statistique et analyse multidimensionnelle
3LTA803U	L8TIL3		Programmation et projet encadré 2
3LTA803T	L8TIL4		Outils de traitements de corpus
3LTA801T	L8TIL2		Corpus parallèles et comparables
			Langages réguliers et langages hors contextes 2
			DOCUMENT NUMERIQUE ET INFORMATIQUE 2
3LTA807T	L8DNI2		Recherche et extraction d'information
3LTA806T	L8DNI1		Document structuré
3LTA808T	L8DNI3		Programmation et algorithmique 2
			Programmation logique
			Programmation et abstraction
			Remise à niveau info
			LINGUISTIQUE 2
3LTA802T	LZST01		Introduction à la fouille de textes
			Sémantique textuelle
			Techniques de traduction
			LANGUE
			Compétence langue
3LTA809T			Langue

MASTER 2ème année

4 blocs distincts

pro	**parcours D**	le M2 parcours *IL-DEFI* Paris Nanterre
recherche	**parcours R**	le M2 parcours *Recherche & Développement* Paris 3/Paris Nanterre /INALCO
pro	**parcours T**	le M2 parcours *Traductique* INALCO
pro	**parcours I**	le M2 parcours *Ingénierie multilingue* INALCO

Les contenus de ces 4 parcours en M2 sont décris *infra*.

Parcours D : M2 IL-DEFI, Paris Nanterre

Semestre 1 (30 ECTS)

Code UE / EC Paris X	*Unités d'enseignement*	*crédits*
	DOCUMENTATION ET FLUX D'INFORMATION	***9***
3LDF904I	Structuration et nature de l'information	3
3LDF905I	Veille et intelligence économique	3
3LDF906I	Management des systèmes d'information documentaire	3
	DE LA LANGUE AUX CONNAISSANCES	***9***
3LDF907I	Corpus annotés et développement de ressources linguistiques	3
3LDF909I	Linguistique outillée et traitements statistiques	3
3LDF908I	Langages du Web sémantique	3
	INFORMATIQUE ET TRAITEMENT DE L'INFORMATION	***9***
3LDF901I	Programmation et programmation orientée objet	3
3LDF902I	Base de données et Web dynamique	3
3LDF903I	Document structuré et écriture numérique	3
	ANGLAIS	***3***
3LIM930I	Anglais	3

Semestre 2 (30 ECTS)

Code UE / EC Paris X	*Unités d'enseignement*	*crédits*
	PREPARATION A L'INSERTION PROFESSIONNELLE	***6***
3LDF003I	Gestion de projets	3
3LDF004I	Conférences professionnelles	3
	PROJET DE FIN D'ETUDE	***9***
3LDF001I	Projet de fin d'étude	9
	STAGE EN ENTREPRISE	***15***
3LDF002I	Stage en entreprise	15

Parcours R : M2 R&D, Paris Nanterre, Paris 3, Inalco

Semestre 1 (30 ECTS)

Code UE/EC Inalco	Code UE/EC P3	Code UE/EC Paris Nanterre	*Unités d'enseignement*	*crédits*
	L9TAL1 L9TAL3		**INGENIERIE EN TAL**	***18***
			6 enseignements à prendre parmi ceux qui suivent (ou d'autres à choisir en accord avec le directeur de recherche) **en plus** du cours *Méthodologie de la Recherche* (validé pour le S2)	
			SITE Paris Nanterre	
	L9ST01	3LFD902S	Modélisation des langues	3
	L9ST02	3LDF910T	TAL et ingénierie des connaissances	3
	L9ST03	3LDF903I	Document structuré et écriture numérique	3
	L9ST04	3LDF902I	Base de données et Web dynamique	3
	L9ST05	3LDF901I	Programmation et programmation orientée objet	3
	L9ST06	3LDF913T	*Méthodologie de la Recherche*. Epistémologie du TAL **(* pour la validation du S2)**	3
				3
	L9ITD1	3LDF907I	Corpus annotés et développement de ressources linguistiques	3
	L9ITD2	3LDF909I	Linguistique outillée et traitements statistiques	3
	L9ITD3	3LDF908I	Langages du Web sémantique	
	L9ST07	3LRD914T	Acquisition, modélisation et représentation des connaissances	
	L9ST08	3LRD9AMR	Ingénierie des connaissances	
			SITE Paris 3	
	LYST02	3LRD906T	Analyse du discours et lexicométrie **(**)**	3
	LYST01	3LRD911T	Fouille de textes	3
	F9SL06		Expérimentation et modalisation dans les humanités numériques	6
			SITE INALCO	
	L9ST09	3LRD915T	Sémantique des textes multilingues	3
	L9ST10	3LRD912T	Genres, textes et usages	3
	L9ST11	----------	Lexicologie, terminologie, dictionnairique	3
			SITE Paris 7	
	L9ST12	3LRD905T	Analyse de données pour le TAL	
	L9ST14	3LRD908T	Analyse sémantique automatique	
	L9ST15	3LRD907T	Analyse et génération de discours	
	L9TAL2		**LINGUISTIQUE**	***9***
			2 ou 3 enseignements de linguistique à prendre en **accord avec le directeur de recherche**	9
	L9TAL4		**LANGUE**	
		3LRD917T	Langue vivante	3

(*) Ce séminaire (qui a lieu au S1) est obligatoire pour la validation du S2 : la note de Méthodologie (S2) sera donnée en fonction du travail effectué à cette occasion en concertation avec le directeur de mémoire.

(**) Ce séminaire doctoral est ouvert aux M1 et M2, il a lieu au S2 (et accessible uniquement pour les M2 n'ayant pas validé le cours « Statistique et analyse multidimensionnelle » du M1)

Semestre 2 (30 ECTS)

Code UE/EC Inalco	Code UE/EC P3	Code UE/EC Paris X	*Unités d'enseignement*	*crédits*
			STAGE	***9***
	L0TAL1	3LRD003T	Stage en laboratoire ou en entreprise	9
			MEMOIRE	***15***
	L0TAL2	3LRD002T	Mémoire de recherche	15
			METHODOLOGIE DE LA RECHERCHE	***6***
	L0TAL3	3LRD001T	*Méthodologie de la recherche*	6

Parcours T : M2 Traductique, Inalco

Semestre 1 (30 ECTS)

Code UE/EC INALCO	*Unités d'enseignement*	*crédits*
	LINGUISTIQUE	*6*
	Sémantique des textes multilingues 1 Lexicologie, terminologie, dictionnairique 1 Genres, textes et usages 1	
	MODELES ET FORMALISMES	*6*
	Acquisition, modélisation et représentation des connaissances Documents structurés	
	TRADUCTION TECHNIQUE	*9*
	Traduction technique 1 Traductologie 1 Conduite de projet de traduction 1	
	INFORMATIQUE	*3*
	Programmation shell Bases de données pour le web	
	INGENIERIE	*6*
	Multimodalité du document numérique (**) Outils de TAO 1 Ecritures et multilinguisme Traitement statistique de corpus	

(**) Ce cours s'intitule désormais « Ingénierie des connaissances »

Semestre 2 (30 ECTS)

Code UE/EC INALCO	*Unités d'enseignement*	*crédits*
	LINGUISTIQUE	*3*
	Sémantique des textes multilingues 2 Lexicologie, terminologie, dictionnairique 2	
	MODELES ET FORMALISMES	*3*
	Indexation et gestion électronique de documents	
	TRADUCTION TECHNIQUE	*6*
	Traduction technique 2 Traductologie 2 Conduite de projet de traduction 2	
	INGENIERIE	*3*
	Techniques web Outils de TAO 2	
	STAGE + MEMOIRE	*15*
	STAGE + MEMOIRE	

Parcours I : M2 Ingénierie Multilingue, Inalco

Semestre 1 (30 ECTS)

Code UE/EC INALCO	*Unités d'enseignement*	*crédits*
	LINGUISTIQUE	*6*
	Sémantique des textes multilingues 1 Lexicologie, terminologie, dictionnairique 1 Genres, textes et usages 1	
	MODELES ET FORMALISMES	*9*
	Calculabilité Acquisition, modélisation et représentation des connaissances Documents structurés	
	INFORMATIQUE	*6*
	Programmation objet 1 Langages de scripts	
	INGENIERIE	*9*
	Multimodalité du document numérique (**) Outils de TAO 1 Analyse du discours et lexicométrie Ecritures et multilinguisme Outils de traitement de corpus Traitement statistique de corpus	

(**) Ce cours s'intitule désormais « Ingénierie des connaissances »

Semestre 2 (30 ECTS)

Code UE/EC INALCO	*Unités d'enseignement*	*crédits*
	LINGUISTIQUE	*3*
	Sémantique des textes multilingues 2 Lexicologie, terminologie, dictionnairique 2	
	MODELES ET FORMALISMES	*3*
	Analyse robuste	
	INFORMATIQUE	*6*
	Programmation objet 2 Programmation itérative et récursive Bases de données sur le web	
	INGENIERIE	*3*
	Techniques web Outils de TAO 2	
	STAGE + MEMOIRE	*15*
	STAGE + MEMOIRE	

Contenu des unités d'enseignement

UE de M1 S1 :
TAL et ingénierie linguistique 1
Cette UE donne un aperçu d'un certain nombre d'outils et de méthodes du traitement automatique des langues. Sont étudiées les Grammaires formelles (ou les langages réguliers), l'analyse syntaxique automatique, Informatique et phonétique, la gestion informatique du multilinguisme. La réalisation d'un premier projet est amorcée.

Linguistique 1
Cette UE a pour objectif d'accroître les connaissances en linguistique générale des étudiants. En dehors d'un enseignement sur la génération automatique de textes (*Machine creativity and text generation*) et d'un enseignement en Syntaxe formelle, ils pourront prendre à leur choix 3 crédits de linguistique.

Document numérique et Informatique 1
Il s'agit, outre la familiarisation avec la notion de bases de données, le début d'un apprentissage systématique et raisonné de la programmation.

UE de M1 S2 :
TAL et ingénierie linguistique 2
On continue d'appréhender les outils et méthodes du traitement automatique des langues, notamment les méthodes statistiques, le traitement de corpus etc. Un second projet est aussi amorcé.

Linguistique 2
Les étudiants poursuivent le perfectionnement de leur culture en linguistique générale, en suivant un enseignement de Fouille de textes à quoi s'ajoutent trois crédits à prendre au choix.

Document numérique et Informatique 2
Outre la poursuite de l'apprentissage de la programmation, les étudiants se familiarisent avec les techniques de Recherche et extraction d'information et de traitement des documents structurés.

UE de M2 IL-DEFI :
Documentation et flux d'informations
On étudie les fondements sur la structuration et la nature de l'information, et il est développé ce que sont la veille et l'intelligence économique

Informatique et traitement de l'information
Les connaissances en programmation sont prolongées par un apprentissage de la programmation orientée objet, par la manipulation des chaînes de traitement sous Unix. Les connaissances en bases de données sont prolongées par la connaissance de la mise en place de sites Web dynamiques. Enfin toute la chaîne de la gestion de l'information électronique est présentée.

Préparation à l'insertion professionnelle

On explique comment sont industrialisés les processus, notamment en ingénierie linguistique. De même est présentée la gestion des projets, et des conférences sont données par des professionnels.

Projet de fin d'études
Par groupes de 4 à 6 étudiants est réalisé un projet qui nécessite une étude de la tâche à accomplir, la définition d'un cahier des charges, la répartition du travail, et qui fait appel à un certain nombre des connaissances techniques acquises dans les autres UE du master.

Stage en entreprise
L'étudiant s'intègre réellement dans une entreprise, et il présente son expérience dans un rapport de stage soutenu oralement.

UE de M2 IL-DEFI et R&D :
De la langue aux connaissances (IL-DEFI) / Ingénierie des textes et des documents (R&D)
On étudie et manipule un certain nombre de logiciels de traitement automatique de la langue et notamment des outils statistiques. On se perfectionne dans le traitement des expressions régulières, et on va aussi loin que possible dans l'appréhension des outils de traitement des documents structurés.

UE de M2 R&D :
Ingénierie en TAL
Il s'agit, pour les étudiants, de se familiariser avec la démarche de recherche en traitement automatique des langues et/ou en linguistique formelle, à travers l'étude systématique de quelques problèmes bien définis et d'approches bien identifiées.

Linguistique
L'objectif est d'approfondir la culture en linguistique des étudiants, en les mettant au fait d'un certain nombre de recherches actuelles dans le domaine.

Méthodologie de la recherche
L'étudiant, à travers notamment une réflexion sur son propre travail d'élaboration d'un mémoire et celui des autres étudiants, progresse dans ses capacités à faire de la recherche dans le domaine.

Stage en laboratoire ou en entreprise
Étude d'un problème dans un contexte de coopération.

Mémoire
Rédaction d'un mémoire de recherche soutenu oralement devant un jury.

UE de M2 Ingénierie multilingue (INALCO):

Linguistique (I)	Sémantique des textes multilingues
	Lexicologie, terminologie, dictionnairique
	Genres, textes et usages
Modèles et formalismes (I)	Documents structurés
	Calculabilité
	Acquisition et modélisation des connaissances

Informatique (I)	Programmation objet 1
	Langages de script
Ingénierie (I)	Multimodalité du document numérique (Ingénierie des connaissances)
	Analyse du discours et lexicométrie
	Outils de TAO 1
	Outils de traitement de corpus
	Traitement statistique de corpus
	Ecritures et multilinguisme
Linguistique (I)	Sémantique des textes multilingues 2
	Lexicologie, terminologie, dictionnairique 2
Modèles et formalismes (I)	Indexation et gestion électronique de documents
Informatique (I)	Analyse robuste
	Programmation itérative et récursive
	Programmation objet 2
	Bases de données sur le web
Ingénierie (I)	techniques web
	Outils de TAO 2

UE de M2 Traductique (INALCO):

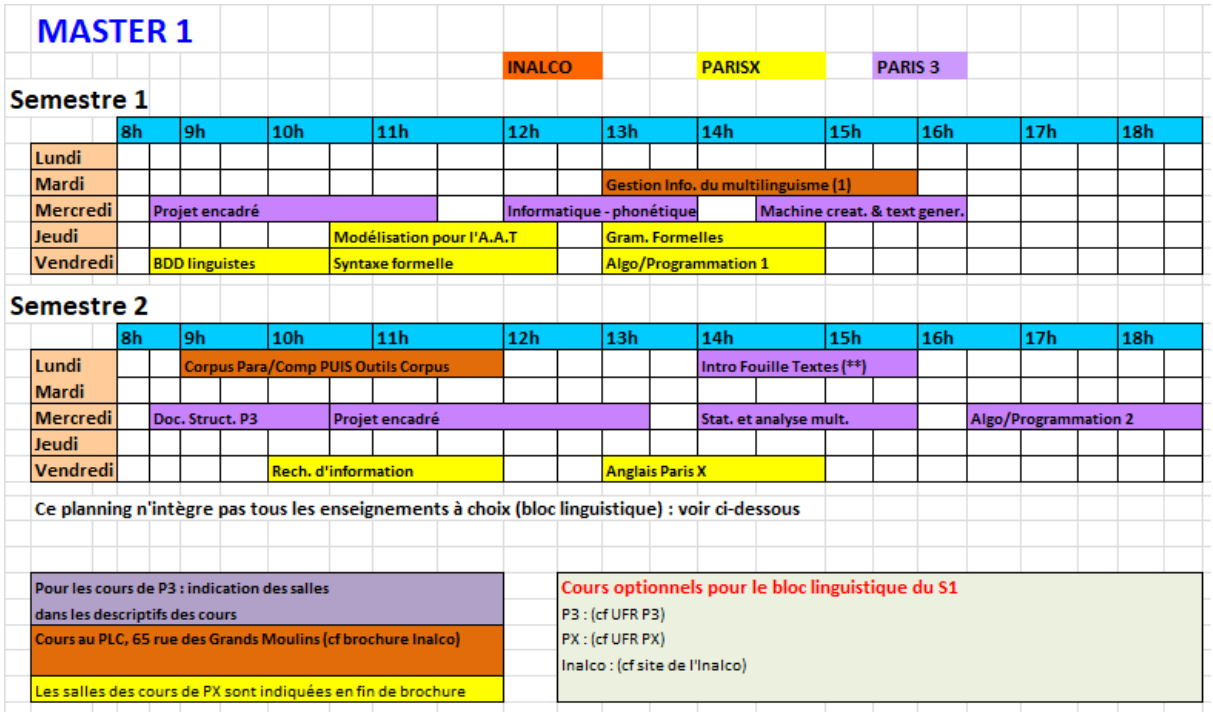
Linguistique (T)	Sémantique des textes multilingues 1
	Lexicologie, terminologie, dictionnairique 1
	Genres, textes et usages
Modèles et formalismes (T)	document structurés
	Acquisition et modélisation des connaissances
Traduction technique (T)	Traductologie 1
	Conduite de projets de traduction 1
Informatique (T)	Bases de données pour le web
	Programmation shell
Ingénierie (T)	Multimodalité du document numérique (Ingénierie des connaissances)
	Outils de TAO 1
	Traitement statistique de corpus
	Ecritures et multilinguisme
Linguistique (T)	Sémantique des textes multilingues 2
	Lexicologie, terminologie, dictionnairique 2
Modèles et formalismes (T)	Indexation et gestion électronique de documents
Traduction technique (T)	Traduction technique 2
	Traductologie 2
	Conduite de projets de traduction 2

Ingénierie (T)	Techniques web
	Outils de TAO 2

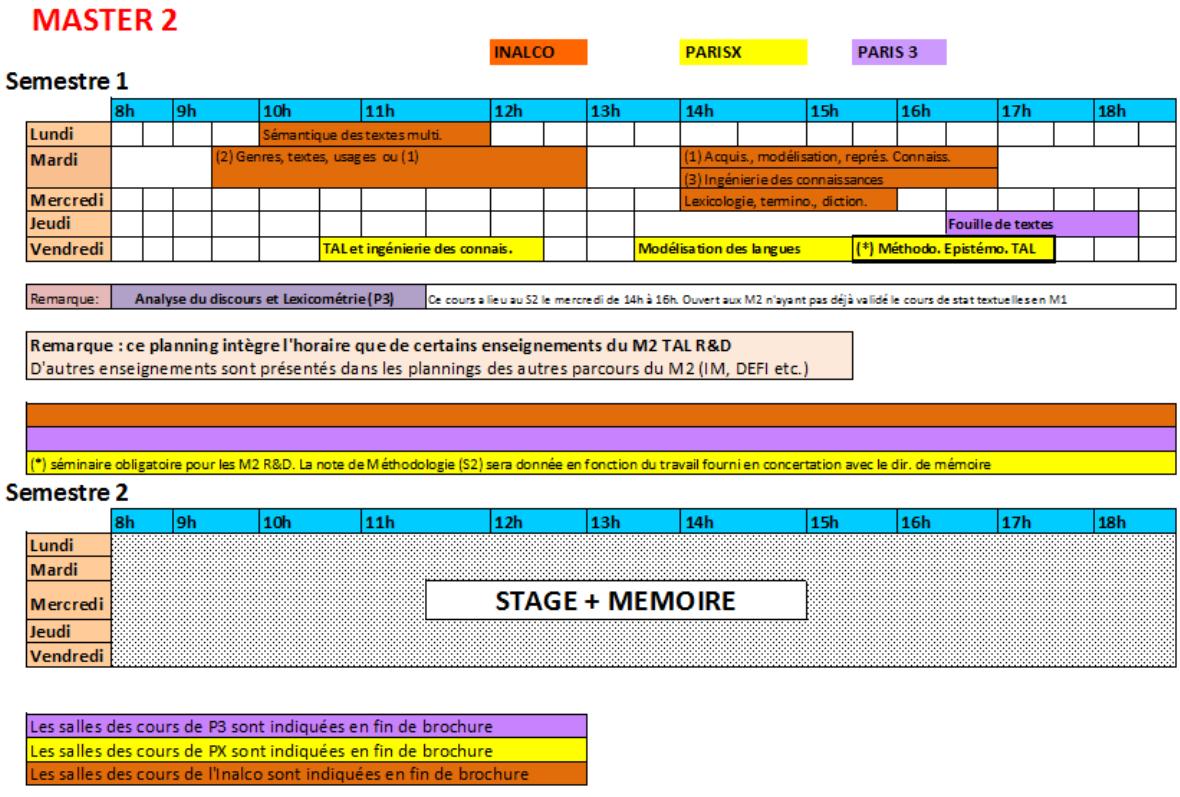
Planning des cours du Tronc Commun du Master T.A.L

LES PLANNINGS QUI SUIVENT SERONT MIS A JOUR DEBUT SEPTEMBRE 2017

Le planning qui suit concerne tous les étudiants M1 de Paris 3 et de Paris Ouest (pour les étudiants de l'Inalco, voir sur le site plurital.org ou sur le site de l'Inalco)



Le planning qui suit concerne les étudiants inscrits dans le parcours R&D en M2 :



Pour les étudiants M2 de l'Inalco, voir sur le site plurital.org
Pour les étudiants M2 de Paris Ouest, voir *infra*

Planning des cours Paris Ouest

- M1 Tal Planning :
  https://planning.u-paris10.fr/direct/index.jsp?login=LLPHIW&projectId=1&showTree=false&displayConfName=Standard%20sans%20fusion&resources=5049

- M2 Tal Planning :
  https://planning.u-paris10.fr/direct/index.jsp?login=LLPHIW&projectId=1&showTree=false&displayConfName=Standard%20sans%20fusion&resources=5051

Equipe pédagogique

Nom : BATTISTELLI Prénom : Delphine
Email : delphine.battistelli@u-paris10.fr
Université / UFR de rattachement : Université Paris Ouest - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114

Nom : CHERFI Prénom : Hacène
Email : hcherfi@u-paris10.fr
Université / UFR de rattachement : Université Paris Ouest - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114 / Mondeca

Nom : CLAVERIE Prénom : Camille
Email : cclaveri@u-paris10.fr
Université / UFR de rattachement : LLPHI
Equipe de recherche : CRIS

Nom : CORI Prénom : Marcel
Email : mcori@u-paris10.fr
Université / UFR de rattachement : Université Paris Ouest - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114

Nom : DESMETS Prénom : Marianne
Email : desmets@u-paris10.fr
Université / UFR de rattachement : Université Paris Ouest - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114

Nom : DAUBE Prénom : Jean-Michel
Email : jean-michel.daube@inalco.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : ER-TIM (EAD 2540)

Nom : FLEURY Prénom : Serge
Email : serge.fleury@univ-paris3.fr
Université / UFR de rattachement : Université Sorbonne nouvelle Paris 3, ILPGA
Equipe de recherche : CLESTHIA (EA7345)

Nom : GENDROT Prénom : Cédric
Email : cgendrot@univ-paris3.fr
Université / UFR de rattachement : Université Sorbonne nouvelle Paris 3
Equipe de recherche : Laboratoire de Phonétique et de Phonologie, UMR7018

Nom : GERDES Prénom : Kim
Email : kim.gerdes@univ-paris3.fr
Université / UFR de rattachement : Université Sorbonne nouvelle Paris 3 , ILPGA
Equipe de recherche : Laboratoire de Phonétique et de Phonologie, UMR7018

Nom : KAHANE Prénom : Sylvain
Email : skahane@u-paris10.fr
Université / UFR de rattachement : Université Paris Ouest - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114

Nom : MINEL Prénom : Jean-Luc
Email : jean-luc.minel@u-paris10.fr
Université / UFR de rattachement : Université Paris 10 – Nanterre
Equipe de recherche : MoDyCo, UMR 7114

Nom : MOREAUX Prénom : Marie-Anne
Email : marie-anne.moreaux@inalco.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : ER-TIM (EAD 2540)

Nom : NOUVEL Prénom : Damien
Email : damien.nouvel@inalco.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : ER-TIM (EAD 2540)

Nom : SALEM Prénom : André
Email : salem@msh-paris.fr
Université / UFR de rattachement : Université Sorbonne nouvelle Paris 3, ILPGA
Equipe de recherche : CLESTHIA (EA7345)

Nom : SEGOND Prénom : Frédérique
Email : frederique.segond@inalco.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : ER-TIM (EAD 2540)

Nom : SLODZIAN Prénom : Monique
Email : mslodz@inalco.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : ER-TIM (EAD 2540)

Nom : TELLIER Prénom : Isabelle
Email : isabelle.tellier@univ-paris3.fr
Université / UFR de rattachement : Université Sorbonne nouvelle Paris 3, ILPGA
Equipe de recherche : UMR 8094 - Langues, Textes, Traitements informatiques, Cognition (LATTICE)

Nom : ZWEIGENBAUM Prénom : Pierre
Email : pz@limsi.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : ER-TIM (EAD 2540)

Descriptif et horaires des cours (1ère et 2ème années)

Les horaires et lieux des cours présentés ci-dessous seront disponibles au moment de la rentrée universitaire (ils seront mis en ligne sur le site pluriTAL et diffusés sur la liste pluriTAL). On obtiendra des renseignements précis et à jour concernant ces cours en s'adressant aux secrétariats des UFRs concernés.

Tous les cours sont accessibles aux étudiants Erasmus.

Descriptif et horaires des cours du master 1ère année

Syntaxe formelle

Enseignant : M. Desmets (Paris Nanterre)
Lieu : Paris Nanterre, salle ?
Horaire : vendredi 10h30-12h30

Il s'agit d'un cours de linguistique qui présente un des modèles théoriques et formels les plus aboutis de la fin du 20ème siècle. Aux côtés de LFG, le modèle HPSG, représentant des modèles basés sur les contraintes, propose une représentation simultanée des différents niveaux de description d'une phrase : morpho-phonologique, syntaxique, sémantique, pragmatique, et s'oppose en cela au Programme Minimaliste. Après une présentation générale du modèle, nous étudierons plus particulièrement certaines des analyses développées par l'école « Franco-californienne » pour le français (I. A. Sag, P. Miller, A. Abeillé, D. Godard, O. Bonami, entre autres), qui concernent le traitement des pronoms clitiques, des prédicats complexes, des dépendances non bornées (relatives, interrogatives). Les dernières séances, nous regarderons le fonctionnement des analyseurs syntagmatiques (chunk parcing) développés dans ce cadre théorique.

Modalités de contrôle
Contrôle continu : La moyenne de l'année est composée de 2 DST de 2h et d'une note d'assiduité.
Contrôle dérogatoire et rattrapage : Examen sur table de 2h.

Espace coursenligne : non.

Grammaires formelles

Enseignant : Sylvain Kahane (Paris Nanterre)
Lieu : Paris Nanterre, salle L200
Horaire : Jeudi 13h20-15h20

Le cours présente la théorie des langages formels et les grammaires formelles de référence en linguistique : les automates à nombre fini d'états et les langages réguliers, 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 modélisation de divers phénomènes linguistiques sera abordée : flexion, sous-catégorisation, actant vs. modifieurs, coordination.

Bibliographie
Abeillé Anne, *Les nouvelles syntaxes : grammaires d'unification et analyse du françai*s, Armand Colin, 1993.
Chomsky Noam, *Syntactic structures*, Mouton & co, 1957 [tr. fr. *Strucrures syntaxiques*, Ed. du Seuil, 1969].
Kahane Sylvain, Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Actes de TALN 2001*, vol. 2, Tours, 2001, 60 pages, www.kahane.fr.
Wehrli Eric, *L'analyse syntaxique des langues naturelles*, Masson, 1997.

Modalités de contrôle
Contrôle continu : La moyenne de l'année est composée d'un DM et d'un DS de 2h à la dernière séance.
Contrôle dérogatoire et rattrapage : Examen sur table de 2h.

Espace coursenligne : non.

Modélisation linguistique pour l'analyse automatique de textes

Enseignant : Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, salle ?
Horaire : jeudi 10h30-12h30

Le cours permet de découvrir, sur la base du recours à divers outils d'annotation automatique, certains aspects de l'analyse linguistique sur corpus avec pour domaine d'application l'interface syntaxe/sémantique/discours. On note aujourd'hui un intérêt marqué pour les unités textuelles/discursives d'une taille possiblement différente de la phrase (cadres de discours, cadres temporels, cadres spatiaux...) ainsi que parfois pour les relations rhétoriques/discursives qui lieraient ces unités (Penn Discourse Tree Bank, …). Dans une perspective d'automatisation de la reconnaissance de ces unités textuelles, on cherchera dans ce cours à exhiber divers types de corrélats linguistiques (morphèmes, lexèmes, constructions syntaxiques) de fonctions discursives spécifiques. L'unité adverbiale sera plus particulièrement étudiée dans ce cadre. La notion de phrase sera en outre discutée.

Bibliographie
GAMON, M. (2004). Sentiment classification on customer feedback data : Noisy data, large feature vectors, and the role of linguistic analysis. In *Proc. of the International Conference on Computational Linguistics* (COLING)

HABERT, B. « Portrait de linguiste(s) à l'instrument ». Texto! [en ligne], décembre 2005, vol. X, n°4 http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

POIBEAU, T., MAUREL, D. « A la fin de : Préposition ou déterminant complexe dans les adverbiaux de temps ? ». Cahiers de Grammaire, 1995, n° 20, pp. 101-111.

VICTORRI, B. « Le modèle en linguistique », Encyclopaedia Universalis, 1997 Version préliminaire disponible sur http://halshs.archives-ouvertes.fr/halshs-00009518.

Modalités de contrôle
Contrôle continu : La moyenne de l'année est composée d'un DM et d'un DS de 2h à la dernière séance.
Contrôle dérogatoire et rattrapage : Examen sur table de 2h.

Espace coursenligne : oui.

Gestion informatique du multilinguisme

Enseignant : Jean François Perrot (INALCO), Marie-Anne Moreaux (INALCO)
Lieu : PLC, rue des Grands Moulins, 7.02 et 5.18
Horaire : mardi 16h00-19h00

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, plate-formes et formats différents.

Informatique et phonétique

Enseignant : Cédric Gendrot (Sorbonne nouvelle Paris 3)
Lieu : Censier, salle D11
Horaire : mercredi 12h00-14h00

Ce cours vise à réaliser un système de synthèse de la parole text-to-speech individualisé pour chaque étudiant. Après une introduction à la phonétique/phonologie et au traitement du signal, un historique de la synthèse de la parole sera présenté. Les différentes étapes de la synthèse text-to-speech sont ensuite présentées avant d'être mises en pratique en cours. L'automatisation de cette synthèse sera réalisée au moyen de langages de programmation utilisés en traitement du signal (Praat / Matlab). La validation consistera en un partiel à mi-semestre, puis un devoir à rendre en fin de semestre.

Programmation et projet encadré (semestre 1)

Enseignant : Jean Michel Daube (INALCO), Serge Fleury (Sorbonne nouvelle Paris 3)
Lieu : Censier, salle D11
Horaire : mercredi 08h30-11h30

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.
URL : http://www.tal.univ-paris3.fr/cours/masterproj.htm

Modalités de contrôle
Contrôle continu : une note de projet.

Espace coursenligne : oui (cf plurital.org)

Bases de données pour linguistes

Enseignant : Jean-Luc Minel (Paris Nanterre)
Lieu : Paris Nanterre, salle L205
Horaire : vendredi 08h20-10h20

La description d'une réalité langagière se fait souvent « à la main » : corpus saisi sous traitement de texte, observations faites également à l'aide d'un traitement de texte, etc. Néanmoins, ces descriptions non structurées sont difficilement analysables lorsque le volume de données devient important. Il est donc nécessaire d'utiliser des langages de description et d'interrogation qui permettent de traiter ces données.
Dans une première partie du cours, nous nous intéresserons à la modélisation sous forme de base de données relationnelle et au langage d'interrogation SQL.
Dans une seconde partie, nous nous intéresserons aux modèles de représentation utilisés dans le web sémantique (ou web de données) et plus spécifiquement aux langages RDF et RDFS. Le langage SPARQL 1.1 qui permet d'exprimer des requêtes dans un Triple store (entrepôt RDF) sera présenté.
Enfin, on présentera les notions qui sous-tendent le traitement des données massives (Big Data)
Des exercices sont systématiquement associés à la présentation des concepts.
Le cours ne suppose pas de connaissances informatiques préalables

Bibliographie :
Jean-Luc Hainaut, *Bases de données et modèles de calcul. Outils et méthodes pour l'utilisateur*, Dunod, 2002, Sciences Sup, Paris, 3ème édition [Une présentation méthodologique des bases de données et des tableurs]

Jacky Akoka & Isabelle Comyn-Wattiau, *Conception des bases de données relationnelles en pratique*, Vuibert, 2001, Informatique, Paris [Pour approfondir la conception et l'utilisation de bases de données]

Dean Allemang & James A. Hendler, *Semantic Web for the Working Ontologist Effective Modeling in Rdfs and Owl.*

Modalités de contrôle
Contrôle continu : La moyenne de l'année est composée de deux DS de 2h.
Contrôle dérogatoire et rattrapage : Examen sur table de 2h.

Espace coursenligne : oui.

Statistique et analyse multidimensionnelle

Enseignant : André Salem (Sorbonne nouvelle Paris 3)
Lieu : ILPGA, LaboC
Horaire : mercredi 14h00-16h00

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.

Corpus parallèles et comparables // Outil de Traitement de Corpus

Enseignant : Pierre Zweigenbaum (INALCO)
Lieu : INALCO, rue de Lille
Horaire : (cf planning INALCO)

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.

Recherche et extraction d'information

Enseignant : Hacène Cherfi <hcherfi@u-paris10.fr> (Paris Nanterre / Mondeca)
Lieu : Paris Nanterre
Horaire : vendredi 10h00-12h00, salle L115

La recherche d'information est à la base des moteurs de recherche sur le Web. Le cours aborde les notions de base de l'implantation d'un système de recherche d'information. Nous y montrons l'intérêt de disposer d'un référentiel, d'un vocabulaire contrôlé et/ou de catégories de recherches pour être efficace en termes de résultats de recherche. Nous montrons également que les moteurs de recherche actuels sont assez performants sans ces structures et uniquement grâce à la cooccurrence de termes d'index. L'extraction d'information produit des réponses plus courtes qu'un document, répondant plus précisément aux demandes d'un utilisateur ; 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 ? »). Les techniques d'analyse syntaxique de contenus sont d'une grande utilité pour l'extraction d'information grâce à des grammaires, des règles et des patrons d'extraction. Enfin, nous abordons succinctement la représentation des connaissances, notamment, afin d'exhiber la possibilité d'aller vers une recherche/extraction d'information dite sémantique pour les contenus. Ce cours est validé par un projet reprenant ces trois volets.

Bibliographie :

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

Cunningham H. et al. *Text Processing with GATE* (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.

Ralph Grishman, Beth Sundheim. Message Understanding Conference - 6: A Brief History. In: *Proceedings of the 16th International Conference on Computational Linguistics* (*COLING*), I, Kopenhagen, 1996, 466–471.

Représentation des connaissances : RDF Primer (W3C) https://www.w3.org/TR/rdf-primer/

Modalités de contrôle
Contrôle continu : Examen sur table de 2h.
Contrôle dérogatoire et rattrapage : Examen sur table de 2h.

Espace coursenligne : oui.

Document structuré

Enseignant : Serge Fleury (Sorbonne nouvelle Paris 3)
Lieu : ILPGA, salle LaboC
Horaire : mercredi 08h30-10h30

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 fournie*s
Polycopié et outils sur pages WEB du cours :

Modalités de contrôle
Contrôle continu : une note de projet.

Espace coursenligne : oui (cf plurital.org)

Programmation et projet encadré (semestre 2)

Enseignant : Jean Michel Daube (INALCO), Serge Fleury (Sorbonne nouvelle Paris 3)
Lieu : ILPGA, salle LaboC
Horaire : mercredi 10h30-13h30
Cf descriptif du premier semestre.
URL : http://www.tal.univ-paris3.fr/cours/masterproj.htm

Modalités de contrôle
Contrôle continu : une note de projet.

Espace coursenligne : oui (cf plurital.org)

Programmation et algorithmique 1 et 2

Enseignant : Iris Eshkol (Paris Nanterre), Kim Gerdes (Sorbonne nouvelle Paris 3)
Lieu : Paris Ouest, salle G213A**,** au semestre 1 ; ILPGA, LaboC au semestre 2
Horaire : semestre 1 : vendredi 13h30-15h30
semestre 2 : Mercredi 15h30-17h30

Programmation et algorithmique 1 (Paris Nanterre, Delphine Battistelli)
Ce cours aborde les notions de structures de données dédiées à la linguistique et au traitement automatique des langues : arbres, automates, graphes, structures de traits. Des algorithmes et des programmes seront écrits sur ces structures, ce qui permettra la réalisation d'application élémentaires et plus élaborées de traitement automatique des langues (analyse morphologique, analyse syntaxique, génération automatique). Les programmes seront encore écrits dans le langage de programmation Python.

Programmation et algorithmique 2 (Paris 3, Kim Gerdes)
Ce cours constitue une introduction aux notions de programmation, théoriques et appliquées, adaptée aux besoins du linguiste informaticien. On abordera les idées sous-jacentes à l'algorithmique : la traduction d'un problème en instructions, la modularisation, les décisions, les boucles, représentées dans des organigrammes. Les exemples seront présentés en Python et étudiés sur machine. On développera des simples outils d'accès à des corpus, nécessaires pour des corrections et des comptages, auxquels tout linguiste est confronté régulièrement. On terminera sur les notions de hiérarchie de classes dans la programmation objet qu'on illustrera avec un travail sur les interfaces graphiques.

Machine creativity and text generation

Enseignant : Kim Gerdes (Sorbonne nouvelle Paris 3)
Lieu : Censier, salle D11
Horaire : mercredi 14h30-16h30

While we are getting used to machines that compute faster than us, play chess better than us, and find information faster than us, Computational Creativity remains an oxymoron, because the world of logic and procedures seems to oppose the free flow of the inspired mind. In our understanding of art, however, unpredictivity and mastership in precision are intertwined as two essential part of the creative process. Using computers, we can untangle this process by simulating, modeling, and analyzing what makes an intelligent system creative. In this way, Artificial Creativity becomes a field where ancient mysteries turn into scientific problems – and where the last bastion is challenged that distinguishes us from machines.
The seminar attempts to present a general overview of the field of Machine Creativity before concentrating on the field of text generation. We will briefly discuss the economic, aesthetic, philosophical, and, above all, scientific implications of machines creating text, where the human authorship consists of not much more than giving instructions to machines. The last sessions will be reserved for the linguistic significance of viewing the language faculty as being centered around the process of text generation as opposed to text analysis.
The students will prepare each class with some readings, the texts will be explained and discussed in detail during the seminar, concerning content as well as terminology.

Bibliography:
http://en.wikipedia.org/wiki/Computational_creativity
Boden, Margaret (1990), *The Creative Mind: Myths and Mechanisms*, London: Weidenfeld and Nicholson
Pearson, Matt (2011), *Generative Art: A Practical Guide Using Processing*, Manning Publications.
Reiter, Ehud, and Robert Dale (2000), *Building Natural Language Generation Systems*. Cambridge University Press.
Summers-Stay, Douglas (2012), *Machinamenta: The thousand year quest to build a creative machine*, CreateSpace Publications.
Veale, Tony (2012), *Exploding The Creativity Myth: The Computational Foundations of Linguistic Creativity*, Bloomsbury Academic

---

Introduction à la fouille de textes

Enseignant : Isabelle Tellier (Sorbonne nouvelle Paris 3)
Lieu : ILPGA
Horaire : lundi, 14h-16h

Ce cours proposera une introduction aux grandes tâches d'ingénierie linguistique qui constituent aujourd'hui ce que l'on résume par le terme de "fouille de textes". Y seront ainsi abordées la segmentation, l'annotation, la classification, la recherche et l'extraction d'information. Ces tâches partagent en effet beaucoup de propriétés :
- représentation des textes sous différentes formes normalisées (sacs de mots, séquence de « tokens")
- utilisation de ressources externes (listes, dictionnaires, thesaurus, ontologies...)
- mesures d'évaluation quantitatives (précision, rappel, F-mesure, exactitude...)
Le cours se concentrera ensuite sur la recherche d'information et ses variantes (booléenne, vectorielle, PageRank...) et sur les différentes techniques actuelles de classification de textes par apprentissage automatique supervisé (Naive Bayes, arbres de décision, SVM...).

Bibliographie

Amini M-R, Gaussier E., *Recherche d'information, Applications, modèles et algorithmes*, Eyrolles 2013.
Cornuejols Antoine, Miclet Laurent, *Apprentissage artificiel, Concepts et Algorithmes*, Eyrolles, 2010 (2ème édition révisée).
Ibekwe-SanJuan F., *Fouille de textes : méthodes, outils et applications*, Hermès, 2007.
Gaussier E., Yvon, F. (cordinateurs), *Modèles statistiques pour l'accès à l'information textuelle*, Hermès, 2011.

Lexique et morphologie

Enseignant : Ch. Jacquet-Pfau (Inalco)
Lieu : PLC
Horaire : (cf planning INALCO)

Il s'agit, dans ce cours, de mettre en place une méthodologie adaptée à l'analyse morphologique, qui, si elle s'inscrit dans un objectif d'implémentation informatique, doit intéresser tout travail sur le lexique. Ainsi les notions de règles et d'exception seront approfondies au cours de l'analyse de corpus que l'on constituera. Un travail sur des sujets représentatifs des principales difficultés auxquelles est confrontée l'analyse morphologique permettra de déterminer des processus d'analyse, pour le français, mais en ayant toujours le souci de la comparer aux systèmes d'autres langues. Une attention toute particulière sera accordée à la notion très large d'emprunt linguistique (abordée dans différentes langues), les emprunts constituants une zone en marge du système de la langue, et posant des problèmes fondamentaux dans de nombreuses applications (analyses orthographiques et sémantiques, traduction, indexation…).
Prérequis : connaissances de base en linguistique.

Descriptif et horaires des cours du master 2ème année

Document structuré et écriture numérique

Enseignant : S. Pouyllau (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
Horaire : (cf. planning en ligne https://goo.gl/QF2IG8)

L'utilisation du langage XML pour décrire des documents semi-structurés nécessite généralement d'effectuer des transformations sur ces documents afin de les utiliser dans des systèmes d'information documentaires. L'utilisation des langage XSLT et XPATH est détaillée afin de montrer le type de transformation qu'il est possible de réaliser. Les langages RDF, RDFS et OWL (utilisés dans le web sémantique) qui appartiennent dorénavant à la galaxie XML sont présentés afin d'en montrer les principales finalités. Le cours est centré sur la réalisation de *mashups* web utilisant des flux XML issus de requête SPARQL dans des triples store RDF et dans des APIs. Il permet de mettre en application XSLT, XPATH et de voir les principaux aspects de la gestion d'un projet de développement de SI documentaire.

Corpus annoté et développement de ressources linguistiques

Enseignant : D. Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
Horaire : (cf. planning en ligne https://goo.gl/QF2IG8)

Ce cours présentera des méthodes, modèles et applications propres à appréhender un niveau d'analyse et d'annotation sémantique des textes. Il exploitera le rapprochement manifeste ces dernières années entre les domaines du TAL et de la Recherche d'Information pour ce qui concerne en particulier la fouille textuelle et/ou l'accès au contenu informationnel des textes. L'enjeu se situe à l'aune d'une masse croissante de documents textuels de types très divers (depuis des fonds d'archives historiques numérisés jusqu'à des ensembles de pages web évolutives en passant par des articles scientifiques du domaine de la biologie) qui peuvent inviter à des traitements sémantiques finalisés différents. Les catégories linguistiques du temps et de la modalité seront ici plus particulièrement abordées.

Bibliographie
A. CONDAMINES (ed), 2005 : *Sémantique et corpus*. Londres : Hermes

Modalités de contrôle
Contrôle continu : Deux dossiers de projets.
Contrôle dérogatoire et rattrapage : Un dossier de projet.

Espace coursenligne : oui.

Langages du Web sémantique

Enseignant : H. Cherfi (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
Horaire : mercredi 9h30-12h30, 8 séances (cf. planning en ligne https://goo.gl/QF2IG8)

Le cours commence par faire un état des lieux des systèmes d'accès au contenu numérique les plus connus ; nous y montrons que la sémantique y est pauvre ; nous illustrons par des cas d'ambiguïtés. La désambiguïsation est un prétexte pour parler de système classificatoire (pour retrouver les documents numériques dans une bibliothèque) ; pour retrouver un article sur un site d'achat. Par la suite, le cours parle de l'initiative de représentation des connaissances pour les humains (notions d'ontologies), puis les rendre opérationnelles pour des machines (Web sémantique) : représentation, requête, échange d'information *sémantisées* entre systèmes qui permettent des raisonnements automatiques. Les langages y afférant sont par la suite présentés : OWL, RDF, SKOS et SPARQL. Une présentation de la plateforme logicielle de représentation d'ontologies Protégé clôture ce cours. Le cours est validé par un projet de modélisation par groupe et par un devoir d'interrogation d'une base

de connaissances (donnée en cours). Le cours se termine avec une présentation des principaux enjeux actuels de l'Open data qui s'appuient largement sur les technologies sémantiques. Ce cours s'articule avec le cours TAL qui utilise les formalismes du Web sémantique afin d'annoter des corpus textuels. Le cours est validé par un projet de modélisation par groupe et par un devoir d'interrogation d'une base de connaissances (donnée en cours).

Analyse du discours et lexicométrie

Enseignant : André Salem (Sorbonne nouvelle Paris 3)
Lieu : ILPGA, salle Benveniste, ILPGA
Horaire : jeudi 14h00-16h00

Le séminaire est consacré à l'étude des corpus de textes à l'aide des méthodes de la textométrie et de celles de l'Analyse de discours. Les cours ne supposent pas de connaissances préalables dans les domaines de la statistique et de l'informatique, de la part des participants. Un premier module est consacré à l'exposé des différentes méthodes de navigation au sein des corpus textuels (index, concordances, segments répétés, etc.) et des méthodes statistiques utilisées en textométrie (spécificités lexicales et segmentales, analyses des cooccurrences, analyses multidimensionnelles, classification automatique, etc.). Le second module est consacré aux applications des méthodes textométriques à différents types de corpus que les chercheurs sont couramment amenés à construire dans le cadre de recherches quantitatives à base de corpus (séries textuelles chronologiques, corpus parallèles, corpus comparables, traitement des corpus multilingues, corpus de traductions et de co-traductions alignées, etc.). Le dernier module permet aux participants d'exposer l'état d'avancement des recherches qu'ils auront entamées, sur un corpus de textes de leur choix, pour appliquer les méthodes qui font l'objet du séminaire. En fin de semestre, un stage pratique d'une journée permet aux participants de bénéficier d'un encadrement plus personnalisé pour la réalisation d'un rapport sur le travail qu'ils ont fourni durant le semestre.

Sémantique des textes multilingues

Enseignant : M. Valette (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

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.

Acquisition, modélisation et représentation des connaissances

Enseignant : F. Segond (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

(descriptif à venir)

Genres, textes, usages

Enseignant : M. Valette (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

(descriptif à venir)

Lexicologie, terminologie, dictionnairique

Enseignant : M. Slodzian (INALCO)

Lieu : INALCO, salle 131
Horaire : (cf planning INALCO)

(descriptif à venir)

Modélisation des langues

Enseignant : Sylvain Kahane (Paris Nanterre)
Lieu : Paris Nanterre, salle L210
Horaire : vendredi 13h30-15h30

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 (morphèmes, unités lexicales, mots, constructions) 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 lexique et grammaire s'articulent. Ce modèle s'inscrit dans le cadre des grammaires de dépendance et plus particulièrement de la Théorie Sens-Texte. Il emprunte aux grammaires lexicalisées le calcul de la structure d'un énoncé par la combinaison de structures élémentaires et aux grammaires d'unification le mode de combinaison de ces structures. Tous les outils mathématiques utilisés seront introduits et motivés par des questions théoriques.

Bibliographie

Bresnan Joan, 2001, *Lexical-Functional Syntax*, Blackwell.
Creissels Denis, 1995, *Eléments de syntaxe générale*, PUF.
Ducrot Oswald, 1995, Unités significatives, in Ducrot & Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, Seuil.
Goldberg Adele, 1995, *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
Kahane Sylvain, 2002, *Grammaire d'Unification Sens-Texte : vers un modèle mathématique articulé de la langue*, Université Paris 7, 82 pages, www.kahane.fr.
Mel'čuk Igor, 1997, *Vers une linguistique Sens-Texte*, Leçon inaugurale au Collège de France, 78 p.
Polguère Alain, 2008, *Lexicologie et sémantique lexicale*, Presses de l'Université de Montréal
Sag Ivan, Thomas Wasow, Emily Bender, 2003, *Syntactic theory: A Formal Introduction*, CSLI Publications, Stanford.
Tesnière Lucien, 1959, *Eléments de syntaxe structurale*, Klincksieck.

MCC : Moyenne sur au moins trois travaux de recherche à la maison.
Espace coursenligne : non.

Expérimentation et modalisation dans les humanités numériques

Enseignant : Ioana Galleron (Sorbonne nouvelle, Paris 3)
Lieu :
Horaire :

Les humanités numériques sont souvent pensées comme l'application d'outils informatiques à des questions de recherche des différentes disciplines du spectre des sciences humaines et sociales. Cependant, pour Willard McCarty leur véritable apport est l'invitation à questionner la construction même du savoir humain, en le confrontant à l'obligation (et aux limites) de la démonstration mécanique. C'est l'ouverture aux méthodes expérimentales, à la modélisation, qui font, dans cette perspective, l'intérêt et la richesse des humanités numériques.
Se concentrant sur les études littéraires, non sans quelques incursions du côté des arts ou de l'histoire des sciences, ce séminaire proposera une exploration de quelques entreprises de modélisation : vision des textes comme arbres (« ordered hierarchies of content objects ») ou comme graphes, création d'une base de données de personnifications, représentations théâtrales virtuelles, reconstitution de bâtiments ou artefacts disparus, e. a.

Dans la mesure où la modélisation a partie liée avec la construction de « jouets » expérimentaux, il proposera également une réflexion sur les possibles applications didactiques de cette approche.

Fouille de textes

Enseignant : Isabelle Tellier (Sorbonne nouvelle, Paris 3)
Lieu : ILPGA
Horaire : Jeudi 16h15-18h15, salle Durand, ILPGA

Ce cours fait suite à celui intitulé « Introduction à la fouille de textes » en M1 mais sera conçu pour être aussi suivi indépendamment, sans pré-requis. Il proposera une introduction aux grandes tâches d'ingénierie linguistique qui constituent aujourd'hui ce que l'on résume par le terme de "fouille de textes". Y seront ainsi abordées la segmentation, l'annotation, la classification, la recherche et l'extraction d'information. Ces tâches partagent en effet beaucoup de propriétés :
- représentation des textes sous différentes formes normalisées (sacs de mots, séquence de « tokens")
- utilisation de ressources externes (listes, dictionnaires, thesaurus, ontologies...)
- mesures d'évaluation quantitatives (précision, rappel, F-mesure, exactitude.
Le cours se concentrera ensuite sur différentes techniques actuelles faisant intervenir de l'apprentissage automatique : méthodes de classification non supervisées (k-means, EM…), modèles d'annotation (HMM, CRF).

Bibliographie :
Amini M.-R., Gaussier E., *Recherche d'information, Applications, modèles et algorithmes*, Eyrolles, 2013.
Cornuejols Antoine, Miclet Laurent, *Apprentissage artificiel, Concepts et Algorithmes*, Eyrolles, 2010 (2ème édition révisée).
Ibekwe-SanJuan F., *Fouille de textes : méthodes, outils et applications*, Hermès, 2007.
Gaussier E., Yvon, F. (coordinateurs), *Modèles statistiques pour l'accès à l'information textuelle*, Hermès, 2011.

Base de Données et Web Dynamique

Enseignant : J.-L. Minel, D. Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
Horaire : (cf. planning en ligne https://goo.gl/QF2IG8)
Ce cours se compose de deux parties. La première partie est consacrée à l'apprentissage des instructions de base du langage PHP en vue de construire des applications qui articulent la gestion de sites Web à l'aide de formulaires, le traitement et le contrôle des données recueillies et la gestion de ces données dans une base données relationnelles à l'aide du lange SQL. Dans la deuxième partie, les fonctionnalités objets de PHP sont présentées ainsi que diverses bibliothèques (XML, SQL). La réalisation est basée sur le trio serveur Apache / MySQL / PHP.

Bibliographie
PHP 5.5 : développez un site web dynamique et interactif, Heurtel, Olivier: Éd. ENI ; 2013

Modalités de contrôle
Contrôle continu : Deux dossiers de projets
Contrôle dérogatoire et rattrapage : Un dossier sur un sujet spécifié par l'enseignant

Espace coursenligne : oui.

TAL et ingénierie des connaissances

Enseignant : Delphine Battistelli, Jean-Luc Minel (Paris Nanterre)
Lieu : Paris Nanterre
Horaire : Vendredi 10h30-12h30

L'Ingénierie des Connaissances (IC) propose des méthodes et des techniques permettant de modéliser, de formaliser et d'acquérir des connaissances dans un but d'opérationnalisation, de structuration ou de gestion au sens large. Les applications concernées sont celles liées à la gestion des connaissances, à la recherche d'information, à l'aide à la navigation ou encore à l'aide à la décision. Dans sa démarche d'ingénierie, l'IC mobilise les techniques de Traitement Automatique des Langues (TAL) en vue notamment de construire des ontologies ou des ressources linguistiques exploitables dans des systèmes de recherche d'information.

Dans une première partie du cours, on approfondira les possibilités d'inférence offertes par les langages RDFS et OWL du Web de données en vue de pouvoir les exploiter pour construire des modèles conceptuels ou des ontologies métiers. Dans une seconde partie, on présentera deux cas d'usage particulièrement illustratifs : l'un accès sur la visualisation de chronologies évènementielles à partir d'un corpus de dépêches AFP ; l'autre accès sur l'analyse de la modalité épistémique dans des textes du domaine de la biologie. Dans les deux cas, il s'agit de montrer que des informations repérées dans les textes sont susceptibles d'être constituées en connaissances par des experts d'un domaine donné et donc de participer à une ingénierie des connaissances textuelles.

Bibliographie :
Dean Allemang & James A. Hendler, *Semantic Web for the Working Ontologist Effective Modeling in Rdfs and Owl*.

Bob DuCharme, *Learning SPARQL, 2nd Edition, Querying and Updating with SPARQL 1.1*, O'Reilly Media.

Modalités de contrôle
Contrôle continu : Une note de travail personnel.
Contrôle dérogatoire et rattrapage : Un dossier de projet.

Espace coursenligne : oui.

Linguistique outillée et traitements statistiques

Enseignant : J.-L. Minel (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
Horaire : (cf. planning en ligne https://goo.gl/QF2IG8)

L'arrivée massive de données textuelles sur support numérique a récemment changé la façon de faire des recherches en linguistique et plus généralement en sciences sociales, notamment en cherchant à exploiter de grands ensembles de données attestées ouvrant ainsi la voie au « Big data ».

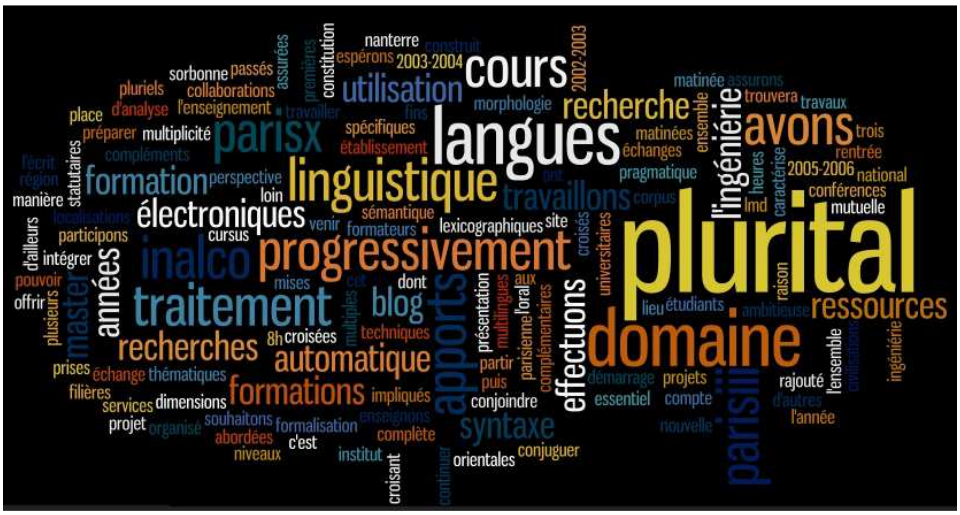
Un des objectif du cours est de permettre aux étudiants de maîtriser les principaux outils statistiques utilisés sciences sociales et plus spécifiquement en linguistique afin d'être capable de les utiliser dans un contexte de applicatif ou de recherche. Au terme de ce cours l'étudiant sera capable de choisir une méthode répondant aux besoins d'une analyse quantitative (analyse univariée et multivariée, test du Chi2, Student et Anova) et de poser un regard critique sur les résultats obtenus. Le cours s'appuie sur des exercices réalisés avec le logiciel libre R avec pour objectif de tester des hypothèses ou d'évaluer les résultats d'applications de TAL.

Une seconde partie du cours est consacrée aux techniques d'annotation automatique symboliques ou par apprentissage automatique supervisé. Des outils d'ingénierie linguistique sont décrits en mettant l'accent sur la nécessité de concevoir ou d'utiliser des modèles de représentation en conformité avec les normes (ISO) ou les standards internationaux (W3C). Différentes applications mettant en œuvre ces outils génériques et ces ressources sont présentées et la question de l'évaluation des outils est discutée. Les étudiants doivent réaliser une ou plusieurs applications en utilisant des outils libres d'accès (Weka).

Bibliographie
Stefan Th. Gries, *Statistics Data for Linguistics With R*, De Gruyter Mouton
Revue TAL : http://www.atala.org/-Revue-TAL-

Modalités de contrôle
Contrôle continu : *Une note de travail personnel*
Contrôle dérogatoire et rattrapage : Un dossier à rendre sur un sujet spécifié par l'enseignant

Espace coursenligne : oui.



http://pluriTAL.org

Contacts

Kahane Sylvain

- Paris Ouest
- MoDyCo
- sylvain@kahane.fr

Daube Jean-Michel

- INALCO
- CRIM
- jean-michel.daube@inalco.fr

Fleury Serge

- Paris 3 / ILPGA
- CLESTHIA
- serge.fleury@univ-paris3.fr