pluriTAL

MASTER Traitement Automatique des Langues

Université Paris Nanterre, Université Sorbonne nouvelle, 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
l'INALCO

MASTER
Mention : *Traitement Automatique des Langues*

Responsables:

Mathieu Valette (INALCO)
Sylvain Kahane (Paris Nanterre)
Cédric Gendrot (Sorbonne Nouvelle)

Commission pédagogique :

Mathieu Valette (INALCO)
Cédric Gendrot (Sorbonne Nouvelle)
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
INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES (INALCO)

La formation s'appuie sur les laboratoires : Paris Nanterre - MODYCO, *Modèles, Dynamiques, Corpus*, (UMR 7114), Sorbonne Nouvelle - CLESTHIA *Langage, systèmes, discours* (EA 7345), Sorbonne Nouvelle - LPP *Laboratoire de Phonétique et Phonologie* (UMR 7018), Sorbonne Nouvelle – Lattice (Langues, Textes, Traitements informatiques, Cognition, UMR CNRS/P3/ENS 8094), 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, discours) ;
2. Acquérir de solides compétences en informatique en général et en programmation en particulier, dans le domaine du Traitement Automatique des Langues (ou NLP pour Natural Language Processing en anglais), dans une perspective autant théorique qu'appliquée, pour l'analyse et la compréhension automatique des langues naturelles, ou leur génération automatique.
3. 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 ;
4. 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

Valette Mathieu (mathieu.valette@inalco.fr)
Cédric Gendrot (cedric.gendrot@sorbonne-nouvelle.fr)
Kahane Sylvain (sylvain@kahane.fr)

Secrétariat administratif

Inalco :
Jessica MAUVIERES
65 rue des grands moulins, 75013 Paris
01.80.71.11.36
jessica.mauvieres@inalco.fr

Sorbonne Nouvelle :
Marie-Claudette Baremon
marie-claudette.baremon@sorbonne-nouvelle.fr
Campus Nation
8 avenue de Saint-Mandé, 75012 Paris
www.sorbonne-nouvelle.fr/ufr-lld
Tél. : 01 45 87 78 08 / 79 07…

Paris Nanterre :
Silva semedo costa Joyce
Bureau L-114
01.40.97.70.75
ssc.joyce@parisnanterre.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.

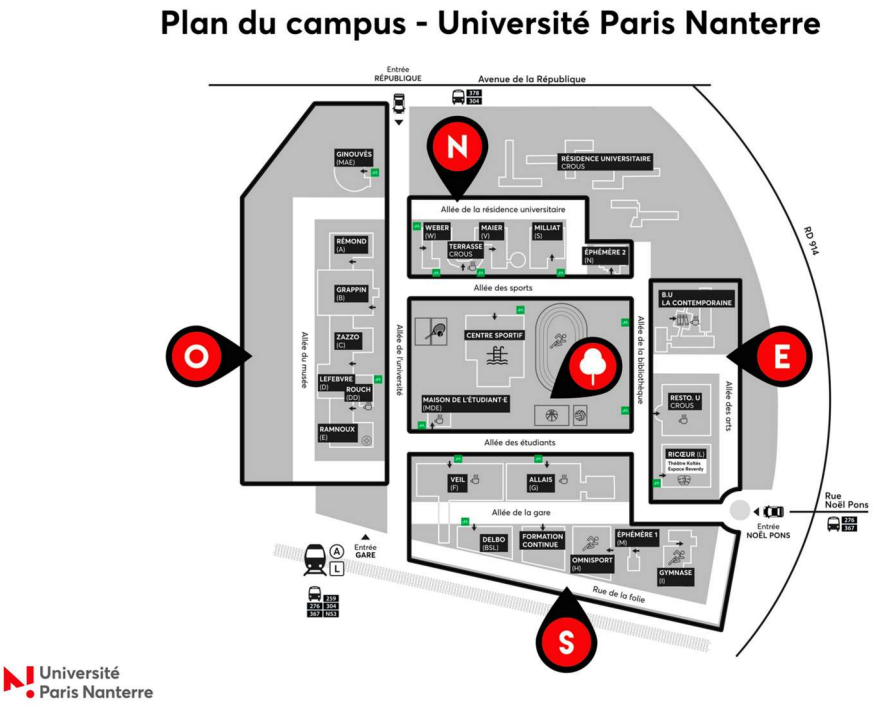
Voir la page « Liste pluriTAL » sur la page web du MASTER (site pluriTAL).

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

Localisations

Paris Nanterre

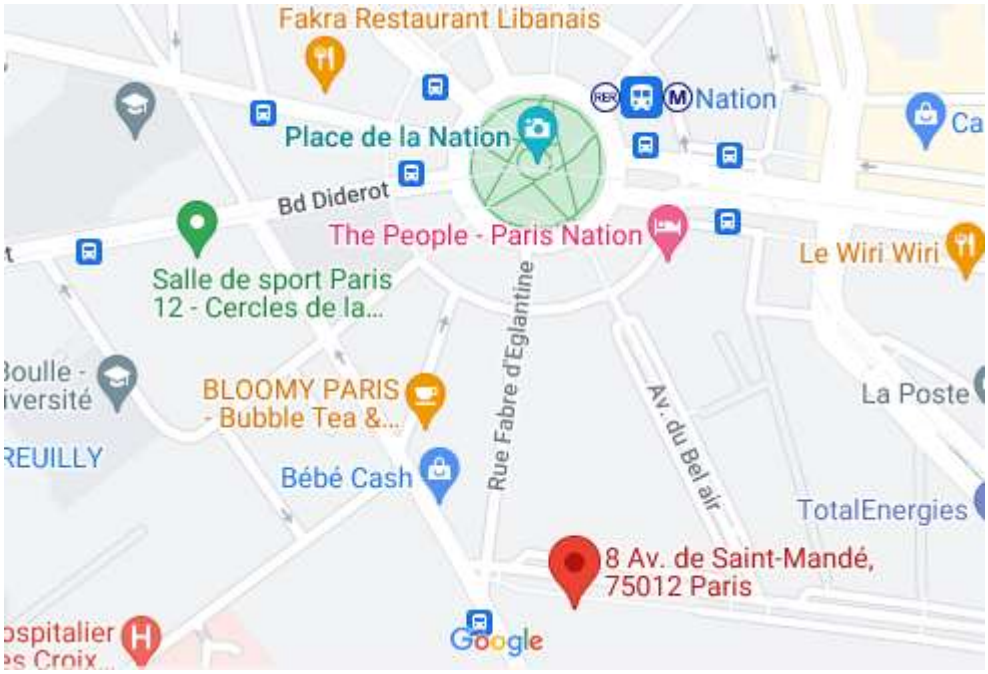
Pour se rendre l'Université de Paris Nanterre : RER A, Direction Saint Germain-en-Laye, station Nanterre Université. Les cours ont lieu dans le bâtiment Paul Ricœur (L), qui se trouve à droite du restaurant universitaire, lui même à droite du bâtiment de la bibliothèque universitaire



Sorbonne Nouvelle / Paris 3

Campus Nation
8 avenue de Saint-Mandé, 75012 Paris
www.sorbonne-nouvelle.fr/ufr-lld
Tél. : 01 45 87 78 08 / 79 07…

Site de Nation ***8 avenue de Saint-Mandé - 75012 Paris (Visualiser)***



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 essentiellement des étudiants issus de filières orientées vers la linguistique ou l'informatique.

Par exemple :

- pour la linguistique, les spécialités : « 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 »).
- et pour l'informatique, les spécialités « Informatique » ; « Mathématique » (etc.).

Il est important que ces étudiants, quelle que soit leur provenance, aient une appétence pour l'ensemble des matières :

- pour les linguistes un intérêt et une base de compétences en informatique,
- pour ceux provenant de filières informatiques un intérêt marqué pour l'étude des langues naturelles autant par des moyens linguistiques qu'informatiques.

Vous devez dans un premier temps contacter par courriel les 3 responsables pédagogiques avant de vous inscrire : sylvain@kahane.fr, cedric.gendrot@sorbonne-nouvelle.fr, mathieu.valette@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 Sorbonne Nouvelle :

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

- Inscription Paris Nanterre :

http://ecandidat.parisnanterre.fr

- Inscription à l'INALCO

Contact : mathieu.valette@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 »

Vous devez contacter par courriel les 2 responsables pédagogiques avant de vous inscrire : cedric.gendrot@sorbonne-nouvelle.fr, mathieu.valette@inalco.fr

En fonction de la réponse de la commission pédagogique, il faudra vous inscrire administrativement dans l'un des 2 établissements (et un seul) :

- Inscription Sorbonne Nouvelle :
http://ecandidat.univ-paris3.fr/

- Inscription à l'INALCO
Contact : mathieu.valette@inalco.fr

Inscription en M2 parcours « Traitement Automatique des langues » (parcours disponible uniquement à Paris Nanterre)

Contact : Delphine Battistelli (delphine.battistelli@parisnanterre.fr) et Iris Eshkol-Taravella (ieshkolt@parisnanterre.fr)
Ce parcours offre un itinéraire avec des cours rassemblés sur 2 jours et demi compatible avec un contrat en alternance et un itinéraire libre équivalent à « Recherche et Développement ».

Inscription en M2 parcours « Ingénierie Multilingue » (parcours disponible uniquement à l'INALCO)

Contact : mathieu.valette@inalco.fr

Inscription en M2 parcours « TeTraDom (Traductique) » (parcours disponible uniquement à l'INALCO)

Contact : mathieu.valette@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 :

SORBONNE NOUVELLE	PARIS X	INALCO
(*cf* secrétariat master TAL)	(*cf* secrétariat Paris Nanterre)	(*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

Formation en alternance

Le M2 parcours « Traitement Automatique des Langues » (parcours disponible uniquement à Paris Nanterre) est ouvert à la formation en alternance.
Contact : Delphine Battistelli (delphine.battistelli@parisnanterre.fr) et Iris Eshkol-Taravella (ieshkolt@parisnanterre.fr)

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

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 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 2020-2021

<table>
<tr><th>PREMIER SEMESTRE</th><th>DEUXIEME SEMESTRE</th></tr>
<tr><td>Journée d’accueil
(cf site plurital)

Début des enseignements
(cf site plurital)

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</td><td>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</td></tr>
<tr><td colspan="2">Jurys de la première session : Cf calendrier universitaire de chaque établissement</td></tr>
<tr><td colspan="2">Examens de la deuxième session : Cf calendrier universitaire de chaque établissement</td></tr>
<tr><td colspan="2">Jurys de la seconde session : Cf calendrier universitaire de chaque établissement</td></tr>
</table>

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.
4. Savoir utiliser des logiciels utiles au traitement de données linguistiques, manipuler des données textuelles pour leur analyse et leur stockage, implémenter des algorithmes réalisant le traitement automatique des langues et les évaluer dans un contexte applicatif donné.

L'objectif de la formation est de donner à des étudiants issus des cursus de langues, de sciences du langage ou d'informatique 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
- 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/

Sorbonne Nouvelle - **CLESTHIA** Langage, systèmes, discours - EA 7345,
http://www.univ-paris3.fr/clesthia-langage-systemes-discours-ea-7345-98241.kjsp

Sorbonne Nouvelle – **LPP** (UMR 7018) Laboratoire de Phonétique et Phonologie
http://lpp.univ-paris3.fr/

Sorbonne Nouvelle – **Lattice** (UMR 8094, CNRS/ENS-PSL)
https://www.lattice.cnrs.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, sémantique, discours) : 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
- Implémentation d'algorithmes de manipulation de données textuelles, autant par méthodes statistiques que par méthode symboliques à base de règles, pour des tâches variées (extraction d'information, représentation des connaissances, traduction automatique ou assistée).

Débouchés

Métiers auxquels le master permet d'accéder directement

Ingénieur linguiste, *data scientist*, 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, programmeur / développeur TAL

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, Synapse, Qwant, Lunii

Organisation globale des enseignements du master

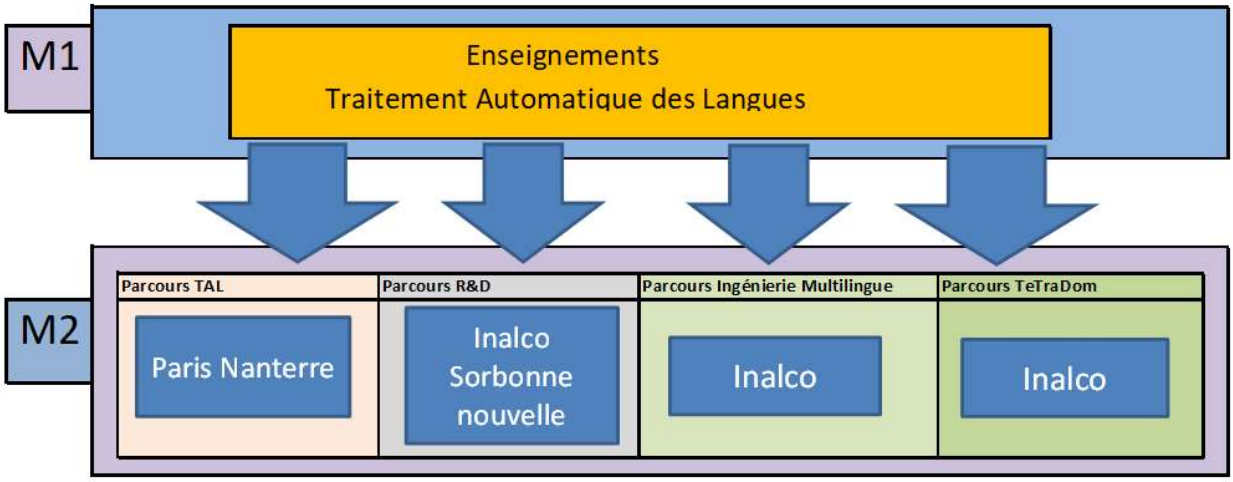
Le master s'organise selon une première année composée d'enseignements en Traitement Automatique des langues puis quatre parcours en M2, adossés chacun à un ou plusieurs établissements universitaires :

- le parcours *TAL*, basé à Paris Nanterre ;

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

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

- le parcours *Recherche et Développement*, basé à l'Inalco et à la Sorbonne Nouvelle.



Master 1, tous parcours Sorbonne Nouvelle, Paris Nanterre

SEMESTRE 1			
TAL ET INGÉNIERIE LINGUISTIQUE 1		Etablissement	ECTS
	Langages réguliers	Inalco	3
	Modélisation linguistique pour l'analyse automatique de textes	Paris Nanterre	3
	Phonétique et synthèse de la parole	Sorbonne nouvelle	3
	Gestion informatique du multilinguisme	Inalco	3
	Programmation et projet encadré 1	Sorbonne nouvelle / Inalco	3
LINGUISTIQUE 1		Etablissement	ECTS
	Machine creativity and text generation	Sorbonne nouvelle	3
	Corpus arborés et parsing	Paris Nanterre	3
	Un enseignement de linguistique au choix	Sorbonne nouvelle/Inalco/Paris Nanterre	3
DOCUMENT NUMÉRIQUE ET INFORMATIQUE 1		Etablissement	ECTS
	Bases de données pour linguistes	Paris Nanterre	3
	Programmation et algorithmique 1	Paris Nanterre	3
SEMESTRE 2			
TAL ET INGÉNIERIE LINGUISTIQUE 2		Etablissement	ECTS
	Statistiques textuelles	Sorbonne nouvelle/Inalco	3
	Corpus parallèles et comparables	Inalco	3
	Programmation et projet encadré 2	Sorbonne nouvelle/Inalco	3
	Outils de traitement de corpus	Inalco	3
LINGUISTIQUE 2		Etablissement	ECTS
	Grammaires formelles	Paris Nanterre	3
	Fouille de textes	Paris Nanterre	3
DOCUMENT NUMÉRIQUE ET INFORMATIQUE 2		Etablissement	ECTS
	Document structuré	Sorbonne nouvelle	3
	Programmation et algorithmique 2	Sorbonne nouvelle	3
	Enrichissement de corpus	Paris Nanterre	3
LANGUE		Etablissement	
	Un enseignement de langue au choix	Sorbonne nouvelle/Inalco/Paris Nanterre	3

Le cours « Machine creativity and text generation» (Sorbonne nouvelle) n'est plus assuré. **Un enseignement de substitution est mis en place (cf Descriptifs).**

Master 1, INALCO

M1 Traitement Automatique des Langues - TAL	60	
Miroir de VET		
Semestre 1	30	
LANGUES VIVANTES ÉTRANGÈRES - 1 EC	4	
Anglais		Inalco
langue de l'inalco (présentiel ou mooc)		
LINGUISTIQUE	7	
Pratiques textuelles et traduction		Inalco
Lexique et morphologie		Inalco
Phonétique et synthèse de la parole		Paris3
MODÈLES ET FORMALISMES	6	
Langages réguliers		Inalco
corpus arborés et parsing		Paris-Nanterre
Principes des bases de données		Inalco
INFORMATIQUE	7	
Bases de données pour linguistes		Paris Nanterre
Programmation objet 1		Inalco
Langages de script		Inalco
INGÉNIERIE	6	
Gestion informatique du multilinguisme		Inalco
Programmation et Projet encadré 1		Paris3 / Inalco
Semestre 2	30	
LANGUES VIVANTES ÉTRANGÈRES - 1EC	4	
Anglais		Inalco
langue de l'inalco (présentiel ou mooc)		
LINGUISTIQUE	5	
Traduction automatique et assistée		Inalco
Sémantique lexicale et textuelle		Inalco
MODÈLES ET FORMALISMES	5	
Documents structurés		Paris3
Grammaires formelles*		Paris Nanterre
INFORMATIQUE	5	
Programmation objet 2		Inalco
Outils de traitement de corpus		Inalco
INGÉNIERIE	5	
Statistiques textuelles		Paris3
Corpus parallèles et comparables		Inalco
Extraction d'informations		Inalco
PROJET	6	
Programmation et Projet encadré 2		Paris3 / Inalco
*optionnel pour M1 InaLCO		

MASTER 2ème année

4 blocs distincts

parcours D	le M2 parcours *TAL* **Paris Nanterre**
parcours R	le M2 parcours *Recherche & Développement* **Sorbonne Nouvelle / INALCO**
parcours T	le M2 parcours *TeTraDom (Traductique)* **INALCO**
parcours I	le M2 parcours *Ingénierie multilingue* **INALCO**

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

Parcours D : M2 TAL, Paris Nanterre

SEMESTRE 1			
DOCUMENTATION ET FLUX D'INFORMATION		**nb H**	ECTS
	Structuration et nature de l'information	24	3
	Veille et intelligence économique	24	3
	Management des systèmes d'information documentaire	16	3
DE LA LANGUE AUX CONNAISSANCES		**nb H**	ECTS
	TAL et Ingénierie des Connaissances	24	3
	Linguistique outillée et traitements statistiques	24	3
	Langages du Web sémantique	24	3
INFORMATION ET TRAITEMENT DE L'INFORMATION		**nb H**	ECTS
	Programmation et programmation orientée objet	24	3
	Base de données et Web dynamique	24	3
	Document structuré et écriture numérique	24	3
ANGLAIS		**nb H**	ECTS
	anglais	20	3
SEMESTRE 2			
PRÉPARATION À L'INSERTION PROFESSIONNELLE		**nb H**	ECTS
	Gestion de projets	24	3
	Conférences professionnelles	10	**3**
PROJET DE FIN D'ÉTUDE			
	Projet de fin d'étude	24	9
STAGE EN ENTREPRISE			
	Stage en entreprise		15

Parcours R : M2 R&D, Paris Nanterre, Sorbonne Nouvelle, Inalco

INGÉNIERIE EN TAL		nb H	ECTS
	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		
	Ingénierie des connaissances	24	3
	Document structuré et écriture numérique	24	3
	Base de données et Web dynamique	24	3
	Programmation et programmation orientée objet	24	3
	Méthodologie de la Recherche. Epistémologie du TAL (pour la validation du S2)	24	3
	Corpus annotés et développement de ressources linguistiques	24	3
	Linguistique outillée et traitements statistiques	24	3
	Langages du Web sémantique	24	3
	Annotations sémantiques et applications en recherche d'information	24	3
	De la modélisation au traitement automatique des données linguistiques	24	3
	Apprentissage automatique	24	3
	TAL et linguistique de corpus	24	3
	SITE Sorbonne nouvelle		
	Sémantique computationnelle	24	3
	Expérimentation et modalisation dans les humanités numériques	24	3
	Réseau de neurones pour la reconnaissance de l'oral et applications linguistiques	26	3
	SITE INALCO		
	Sémantique des textes multilingues	26	3
	Genres, textes et usages	26	3
	Lexicologie, terminologie, dictionnairique	26	3
	Fouille de textes	26	3
	Traitements statistique de corpus	26	3
	Méthodes en apprentissage automatique	26	3
	SITE Paris 7		
	Analyse sémantique automatique		3
LINGUISTIQUE		**nb H**	**ECTS**
	Modélisation des langues (Paris Nanterre)	24	3
	2 enseignements de linguistique à prendre en **accord avec le directeur de recherche**	48	6
LANGUE		**nb H**	**ECTS**
	Langue vivante	24 ou 26	3
	SEMESTRE 2		
STAGE			**ECTS**
	Stage en laboratoire ou en entreprise		9
MÉMOIRE			**ECTS**
	Mémoire de recherche		15
MÉTHODOLOGIE DE LA RECHERCHE			**ECTS**
	Méthodologie de la recherche		6

Parcours T : M2 TeTraDom (Traductique), Inalco

M2 TAL - parcours Technologies de la Traduction et Traitement des Données Multilingues (TeTraDom)	60
Copie de Miroir VET de M1	
Semestre 9	30
LINGUISTIQUE	6 ECTS
Sémantique des textes multilingues 1	
Lexicologie, terminologie et dictionnairique 1	
Genres, textes et usages	
INGÉNIERIE	9 ECTS
Ingénierie des connaissances	
Traitement statistique de corpus	
Fouille de textes 1	
Outils de TAO 1	
Écriture et multilinguisme	
TRADUCTION TECHNIQUE	9 ECTS
Traduction technique 1	
Traductologie 1	
Conduite de projet de traduction 1	
INFORMATIQUE	3 ECTS
Base de données pour le web	
Langages de script	
MODÈLES ET FORMALISMES	3 ECTS
Documents structurés	
Acquisition, modélisation et représentation des connaissances	
Semestre 10	30
LINGUISTIQUE	3 ECTS
Sémantique des textes multilingues 2	
Lexicologie, terminologie et dictionnairique 2	
INGÉNIERIE	3 ECTS
Fouille de textes 2	
Outils de TAO 2	
Web, BD, sites multilingues et localisation	
TRADUCTION TECHNIQUE	6 ECTS
Traduction technique 2	
Traductologie 2	
Conduite de projet de traduction 2	
STAGE ET MÉMOIRE	18 ECTS

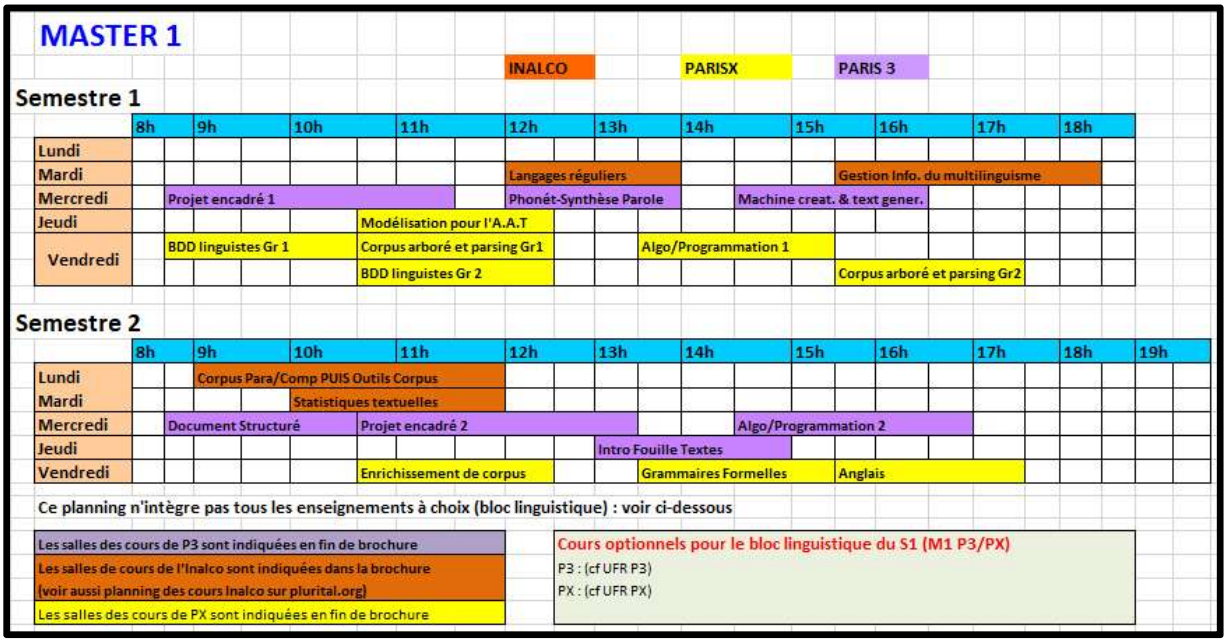
Parcours I : M2 Ingénierie Multilingue, Inalco

M2 TAL - parcours Ingénierie Multilingue		60
Copie de Miroir VET de M1		
Semestre 1		30
LINGUISTIQUE		6 ECTS
	Sémantique des textes multilingues 1	
	Lexicologie, terminologie et dictionnairique 1	
	Genres, textes et usages	
INGÉNIERIE		9 ECTS
	Ingénierie des connaissances	
	Traitement statistique de corpus	
	Fouille de textes 1	
	Outils de traitement de corpus	
	Réseaux de neurones pour la reconnaissance de l'oral et applications linguistiques	
INFORMATQUE		6 ECTS
	Programmation objet 1	
	Langages de script	
MODÈLES ET FORMALISMES		9 ECTS
	Documents structurés	
	Calculabilité	
	Acquisition, modélisation et représentation des connaissances	
Semestre 2		30
LINGUISTIQUE		3 ECTS
	Sémantique des textes multilingues 2	
	Lexicologie, terminologie et dictionnairique 2	
INGÉNIERIE		6 ECTS
	Fouille de textes 2	
	Techniques web	
	Web, BD, sites multilingues et localisation	
	Méthodes en apprentissage automatique	
INFORMATQUE		3 ECTS
	Programmation objet 2	
	Programmation itérative et récursive	
STAGE ET MÉMOIRE		18 ECTS

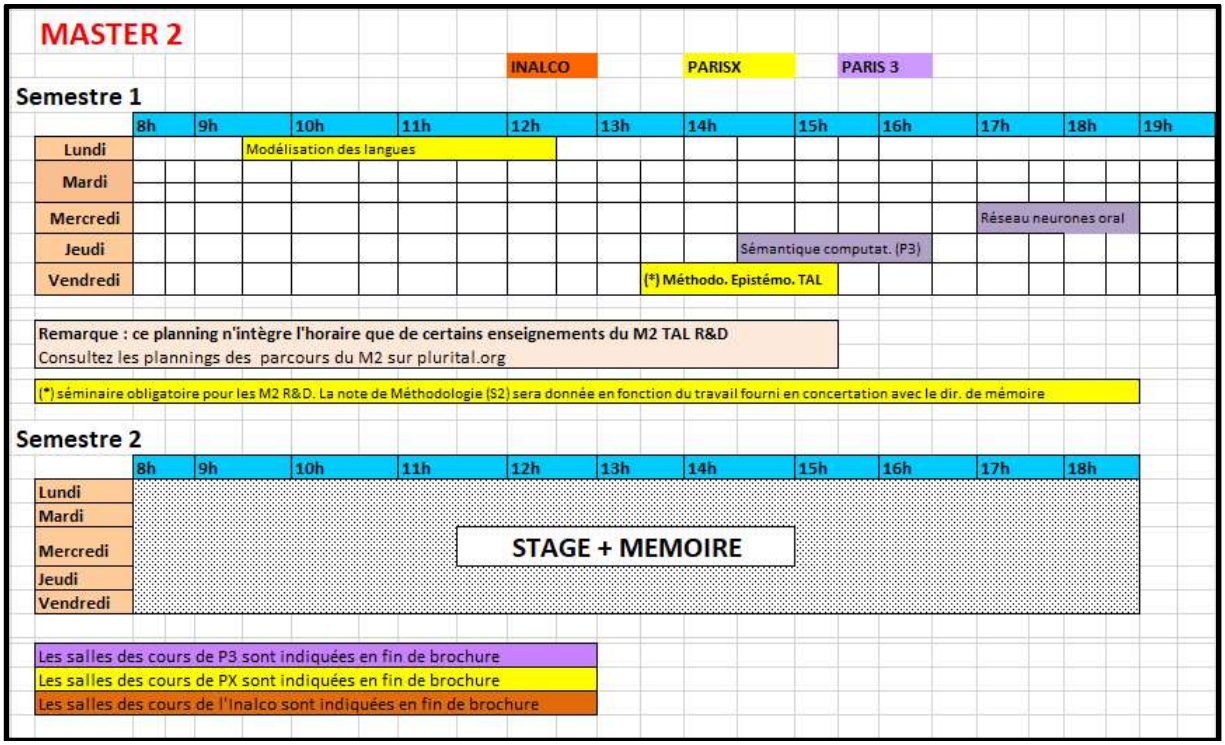
Planning des cours du Tronc Commun du Master T.A.L

LES PLANNINGS QUI SUIVENT SERONT MIS A JOUR AU DEBUT DE L'ANNEE UNIVERSITAIRE

Le planning qui suit regroupe essentiellement **les cours communs à tous les étudiants M1** (pour les M1 de l'Inalco, voir sur le site plurital.org ou sur le site de l'Inalco)



Le planning « partiel » qui suit concerne les **étudiants inscrits dans le parcours R&D en M2** : plannings complémentaires sur plurital.org



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

Planning des cours Paris Nanterre

- 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 : AMSILI Prénom : Pascal
Email : pascal.amsili@gmx.fr
Université / UFR de rattachement : Sorbonne Université
Equipe de recherche : Lattice, UMR 8094

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

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

Nom : ESHKOL-TARAVELLA Prénom : Iris
Email : ieshkolt@parisnanterre.fr
Université / UFR de rattachement : Université Paris 10 – Nanterre
Equipe de recherche : MODYCO

Nom : FLEURY Prénom : Serge
Email : serge.fleury@sorbonne-nouvelle.fr
Université / UFR de rattachement : Université Sorbonne Nouvelle
Equipe de recherche : CLESTHIA (EA7345)

Nom : GENDROT Prénom : Cédric
Email : cedric.gendrot@sorbonne-nouvelle.fr
Université / UFR de rattachement : Université Sorbonne nouvelle
Equipe de recherche : Laboratoire de Phonétique et de Phonologie, UMR7018

Nom : GROBOL Prénom : Loïc
Email :
Université / UFR de rattachement : Université Paris - Nanterre, UFR Phillia
Equipe de recherche :

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

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

Nom : PLANCQ Prénom : Clément
Email : clement.plancq@ens.fr
Université / UFR de rattachement : INaLCO
Equipe de recherche : LaTTiCe, UMR 8094

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

Corpus arboré et parsing

Enseignant : Sylvain Kahane (Paris Nanterre)
Lieu : Paris Nanterre (à préciser à la rentrée)
Horaire : Gr1 : Vendredi 10h30-12h30 ; Gr2 : Vendredi 15h30-17h30

Le cours présente la constitution d'un corpus annoté en syntaxe de dépendance et son utilisation pour le parsing. Les principales notions de syntaxe (unité syntaxique, tête, dépendance, constituant, fonction, macrosyntaxe) sont introduites. Le guide d'annotation SUD (Surface-Syntactic Universal Dependancies) est présenté et chaque étudiant procède à l'annotation d'un fragment de corpus de français parlé.

Bibliographie

Creissels Denis, *Eléments de syntaxe générale*, PUF, 1995.
Fort Karën, *Les ressources annotées, un enjeu pour l'analyse de contenu: vers une méthodologie de l'annotation manuelle de corpus*. Thèse de doctorat. Université Paris-Nord-Paris XIII, 2012, en ligne.
Gerdes Kim, Bruno Guillaume, Sylvain Kahane, Guy Perrier. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of the Universal Dependencies Workshop (UDW)*. 2018.
Kahane Sylvain, Kim Gerdes, *Syntaxe théorique et formelle*, à paraître.
Kübler, Sandra, Ryan McDonald, Joakim Nivre, *Dependency parsing*, Synthesis Lectures on Human Language Technologies, 2009.
Mel'čuk Igor, Jasmina Milićević. *Introduction à la linguistique, vol. 2 : Syntaxe*, Hermann, 2011.
Nivre J., M. C. De Marneffe, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, 1659-1666.
Tesnière Lucien, *Eléments de syntaxe structurale*, Klincksieck, 1959.

Ressources

Arborator, arborator.ilpga.fr.
Universal Dependencies Treebanks, universaldependencies.org.
Grew-match, match.grew.fr.
CEFC (Corpus d'Etude pour le Français Contemporain)/Orféo, ortolang.fr

Modalités de contrôle

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

Espace cours en ligne : non.

Grammaires formelles

Enseignant : Sylvain Kahane (Paris Nanterre)

Lieu : Paris Nanterre, (à préciser à la rentrée)
Horaire : Vendredi 13h30-15h30

Le cours présente la théorie des langages formels et les grammaires formelles de référence en linguistique : les les langages réguliers et leurs limites, les incontournables grammaires de réécriture de Chomsky, les grammaires lexicalisées avec les grammaires catégorielles les TAG (Grammaire d'adjonction d'arbres), les grammaires de dépendance et la Théorie Sens-Texte, et les grammaires basées sur l'unification de structures de traits (HPSG, Head-Driven Phrase Structure Grammar). La modélisation de divers phénomènes linguistiques sera abordée : sous-catégorisation, actant vs. modifieurs, locutions (expressions multi-mots), interface syntaxe-sémantique.

Bibliographie

Abeillé Anne, *Les nouvelles syntaxes : grammaires d'unification et analyse du françai*s, Armand Colin, 2000.
Bresnan Joan, 2001, *Lexical-Functional Syntax*, Blackwell.
Chomsky Noam, *Syntactic structures*, Mouton & co, 1957 [tr. fr. *Structures syntaxiques*, Ed. du Seuil, 1969].
Kahane Sylvain, Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Actes de TALN*, vol. 2, Tours, 2001, 60 p., www.kahane.fr.
Kahane Sylvain, *Grammaire d'Unification Sens-Texte : vers un modèle mathématique articulé de la langue*, 2002, 82 pages, www.kahane.fr.
Mel'čuk Igor, *Vers une linguistique Sens-Texte*, Leçon inaugurale au Collège de France, 1997, 78 p, en ligne.
Polguère Alain, *Lexicologie et sémantique lexicale*, Presses de l'Université de Montréal, 2008.
Rétoré Christian, *Systèmes déductifs et traitement des langues: un panorama des grammaires catégorielles*, rapport INRIA, 2000, 37 p., en ligne.
Sag Ivan, Thomas Wasow, Emily Bender, *Syntactic theory: A Formal Introduction*, CSLI Publications, Stanford, 2003.

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 cours en ligne : non.

Modélisation linguistique pour l'analyse automatique de textes

Enseignant : Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, salle M 202
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

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.

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 cours en ligne : oui.

Gestion informatique du multilinguisme

Enseignant : Ilaine Wang (INALCO)
Lieu : PLC, rue des Grands Moulins
Horaire : (cf planning INALCO)

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.

Phonétique et synthèse de la parole

Enseignant : Cédric Gendrot (Sorbonne nouvelle)
Lieu :
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 (Python / 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 : Serge Fleury (Sorbonne nouvelle)
Lieu :
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 cours en ligne : oui (cf plurital.org)

Bases de données pour linguistes

Enseignant : Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre, salle L308
Horaire : Gr 1 : vendredi 08h30-10h30 ; Gr 2 : vendredi 10h30-12h30

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 bases de données basées sur les graphes comme Neo4j et aux modèles de représentation utilisés dans le web sémantique (ou web de données) et plus spécifiquement aux langages RDF et RDFS. Le langage SPARQL 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 d'un projet et d'un examen sur table de 2h.

Espace cours en ligne : oui.

Statistiques textuelles

Enseignant : Damien Nouvel
Lieu : INALCO
Horaire : (cf planning INALCO)

Ce cours est orienté sur la présentation des bases mathématiques générales pour les probabilités et statistiques (évènements, lois, distributions, séries). Il aborde aussi la loi hypergéométrique (spécificités) et la présentation de quelques outils de statistiques textuelles.

Corpus parallèles et comparables

Enseignant : Pierre Zweigenbaum (INALCO)
Lieu : INALCO
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.

Outils de Traitement de Corpus

Enseignant : Clément Plancq (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

Les outils de traitement de corpus, qu'ils soient issus de la linguistique outillée ou du TAL sont nombreux et évoluent rapidement. Ce cours se veut être une présentation raisonnée des outils disponibles. Il met l'accent sur les méthodes mises en œuvre par ces outils, les formats de données et les langages de requête. La première partie du cours présente les étapes de traitement de données, leurs enjeux et les formats de données (tokenization, étiqueteage en POS, lemmatisation, NER, analyse syntaxique, …). La deuxième partie traite des outils d'interrogation sur les corpus annotés avec des exercices sur le langage CQL (corpus query langage) puis sur l'outil Grew (interrogation de corpus annotés en dépendance). La troisième partie traite des étiqueteurs en POS et de leurs différentes méthodes. La dernière partie du cours présente une bibliothèque logicielle intégrant une chaîne de traitements de TAL (NLTK ou Spacy) à partir d'exemples et de travaux pratiques.

Bibliographie

Tony McEnery and Andrew Wilson. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001 (second edition).
Céline Poudat et Frédéric Landragin. Explorer un corpus textuel : Méthodes – pratiques – outils. De Boeck Supérieur, 2017.
Daniel Jurafsky and James H. Martin. Speech and Language Processing. Pearson, 2008 (second edition).

Enrichissement de corpus

Enseignant : Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre
Horaire : vendredi 10h30-12h30

Le séminaire sera consacré au processus de l'annotation qui consiste dans l'ajout de l'information linguistique et extralinguistique aux corpus bruts. Le processus de l'annotation, les normes, la méthodologie et les techniques utilisées avec des exemples concrets seront présentés. Suite à cette introduction théorique, les étudiants travailleront ensemble par petits groupes de 2-4 personnes sur l'annotation et l'analyse des différents corpus (oraux et écrits) en utilisant et/ou en développant les différents outils informatiques. A la fin, ils présenteront à l'oral les résultats de l'analyse et rendront un petit mémoire écrit.

Document structuré

Enseignant : Serge Fleury (Sorbonne nouvelle)
Lieu : (à préciser)
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 cours en ligne : oui (cf plurital.org)

Programmation et projet encadré (semestre 2)

Enseignant : Serge Fleury (Sorbonne nouvelle)
Lieu : (à préciser)
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 cours en ligne : oui (cf plurital.org)

Programmation et algorithmique 1 et 2

Enseignant : Iris Taravella (Paris Nanterre), Pascal Amsili (Sorbonne nouvelle)
Lieu : Semestre 1 : Paris Ouest, salle Info PX
Semestre 2 :
Horaire : Semestre 1 : vendredi 13h30-15h30
Semestre 2 : Mercredi 17h30-19h30

Programmation et algorithmique 1 (Paris Nanterre, Iris Taravella)
Ce cours aborde les notions de base du langage Python 3 : types de données (données numériques, chaînes de caractères, listes, dictionnaires, tuples), fonctions, etc. Les étudiants acquiérons des compétences dans la manipulation de fichiers, dans la définition de fonctions, dans l'importation de modules et dans l'utilisation d'outils TAL sur les corpus à l'aide de scrpts Python.

Des exercices sont systématiquement associés à la présentation des concepts.
Le cours ne suppose pas de connaissances informatiques préalables

Modalités de contrôle
Contrôle continu : La moyenne de l'année est composée d'un DS de 2h et d'un examen sur table de 2h.

Espace cours en ligne : oui.

Programmation et algorithmique 2 (Pascal Amsili)
Dans la continuité du cours du premier semestre, ce cours poursuit la présentation du langage python en abordant en parallèle les premiers concepts d'algorithmique : notions de qualité de programme et de complexité, étude de quelques algorithmes (tris, parcours d'arbre) et de quelques structures de données (piles, files, listes). On insistera sur le savoir-faire en programmation en proposant de nombreux TPs, et en commençant chaque séance par un exercice de programmation.

Machine creativity and text generation (indisponible à partir 2020-2021)

Un enseignement de substitution est assuré par Marine Delaborde le mercredi de 14h30 à 16h30 en salle Brunot à l'ILGA

Fouille de textes

Enseignant :
Lieu :
Horaire :

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.

Langages réguliers

Enseignant : Damien Nouvel (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

Analyser ou générer des langages naturels à l'aide d'un ordinateur requiert une formalisation algébrique, afin que ces processus soient automatiques et déterministes. D'un point de vue théorique, il s'agit de déterminer les symboles (d'un alphabet) utilisés par un langage et l'opération qui permet de générer un texte (concaténation ou produit). Par suite, il devient possible de modéliser à l'aide d'automates deux classes de langages : les langages "réguliers" et les langages hors-contexte (grammaires). Ce cours présentera la théorie formelle, un temps important est consacré aux exercices et aux travaux pratiques.

Modalités de contrôle : contrôle (50%) et examen (50%)

Bibliographie
Chomsky, N. (1956). Three models for the description of language. IRE Transactions on information theory, 2(3), 113-124.
https://pages.lip6.fr/Jean-Francois.Perrot/inalco/Automates/
http://www.math-info.univ-paris5.fr/~bonzon/Cours/Langages
http://www.labri.fr/perso/anca/Langages/cours/cfl.pdf

Espace cours en ligne : http://damien.nouvels.net/fr/enseignement

---

Sémantique lexicale et sémantique textuelle

Enseignant : Mathieu Valette (INALCO)
Lieu : INALCO
Horaire : au S2 (cf planning INALCO)

Ce cours propose des bases théoriques et pratiques pour l'analyse du lexique du point de vue de la textualité. Dans un premier temps, nous définissons le signe linguistique par opposition au concept et au terme, traditionnellement promus dans le TAL. Il sera notamment question de distinguer ce qui relève de l'ontologie de ce qui relève de la sémantique proprement dite. Dans un deuxième temps, nous aborderons la question de la structuration du lexique en classes sémantiques conçues comme un type de formalisation alternatif ou complémentaire aux ontologies, qui prend en compte l'ancrage social et culturel du lexique. Enfin, nous aborderons, à partir d'exemples concrets, l'analyse comparée des unités lexicales (lexèmes) et des unités sémantiques non lexicalisées (isotopies, formes sémantiques, etc.).

Bibliographie
Gianola, L., M. Valette (2018) « Spécialisation générique et discursive d'une unité lexicale. L'exemple de *joggeuse* dans la presse quotidienne régionale », *Proceedings of the 14th International Conference on Statistical Analysisi of textual Data (JADT'18)*, UniversItalia, vol. 1, 312-318.
Valette, Mathieu (2010b), « Propositions pour une lexicologie textuelle », *Les configurations du sens*, *Zeitschrift für Französische Sprache und Literatur*, 37, Franz Steiner Verlag, éd., pp. 171-188.

Descriptif et horaires des cours du master 2ème année

Sémantique computationnelle

Enseignant : Pascal Amsili
Lieu : Sorbonne Nouvelle, (salle à préciser)
Horaire : jeudi, 14h30-16h30

Ce cours présente dans un premier temps les approches dites distributionnelles qui visent à représenter les mots par des vecteurs encodant les aspects pertinents de leur voisinage, approchent qui ont contribué, sous la forme des plongements lexicaux, aux succès récents des méthodes dites neuronales en IA et en TAL. Dans un deuxième temps, on abordera, de façon plus applicative, des tâches relevant de la sémangtique computationnelle, comme la résolution de coréférences, ou la détection des inférences naturelles.

Fouille de textes

Enseignant : Cyril Grouin
Lieu : INALCO
Horaire : au S1 et au S2 (cf planning INALCO)

L'objectif du cours vise, d'une part à découvrir plusieurs méthodes pour effectuer de la fouille de texte, dans un scénario précis (le repérage d'entités nommées, REN), et d'autre part à évaluer les résultats produits. Pour le REN, nous utiliserons l'annotation manuelle, les méthodes symboliques (règles et lexiques), et les méthodes par apprentissage statistique (CRF de chaîne linéaire). En recherche d'informations (RI), le cours sera l'occasion de découvrir la représentation vectorielle des documents (avec une représentation pondérée des éléments du vecteur au moyen du tf*idf) et le calcul de similarité entre vecteurs (produit scalaire, cosinus).
Nous traiterons l'évaluation des résultats produits, à la fois par les humains (accords inter-annotateurs de la famille des Kappa) mais également par les machines au moyen des mesures utilisées en recherche d'information (rappel, précision, exactitude, spécificité, F-mesure, coefficient de Dice ; macro et micro-mesures ; Slot Error Rate). Au-delà des mesures utilisées et des résultats obtenus, le cours sera l'occasion de s'interroger sur la lecture et l'interprétation de ces résultats.
De manière plus générale, le cours vise également à sensibiliser les étudiants à la méthodologie de travail à adopter, tant dans le cadre du cours (production de guides d'annotation, comparaison qualitative et quantitative des pré-annotations, évaluation en validation croisée, études ablatives, analyse des résultats produits, etc.) qu'au-delà (cahiers d'expériences, structure des devoirs et du mémoire, organisation du travail de groupe).
Le cours repose sur des outils connus du TAL (tel que le TreeTagger pour l'étiquetage-lemmatisation) et des langages de script (Perl, Python) à produire. Nous utilisons également BRAT pour l'annotation de corpus, DARK pour l'annotation à base de règles, et Wapiti pour l'apprentissage statistique. De manière complémentaire, des outils de clustering (implémentation du clustering de Brown), de calcul du code Soundex, etc., pourront venir compléter la chaîne de traitements utilisée. Il est nécessaire d'avoir une machine de type Unix (Linux, Mac OS X) ou d'utiliser les machines en salle de cours. Les émulateurs de type CygWin ne permettent pas d'utiliser efficacement les outils et sont donc à proscrire.

Modalités de contrôle : note de présence (coeff. 3), note de participation (coeff. 3), contrôle continu (coeff. 6), projet/examen final (coeff. 8)

Espace cours en ligne : https://perso.limsi.fr/grouin/inalco/

Réseau de neurones pour la reconnaissance de l'oral et applications linguistiques

Enseignant : Cédric Gendrot, Nicolas Audibert
Lieu :
Horaire :

Dans ce cours, nous proposerons une introduction pratique aux réseaux de neurones pour l'application à des données orales (reconnaissance de la parole, du locuteur, des émotions, etc.) ainsi que d'autres applications linguistiques possibles. Des connaissances solides en python sont exigées pour ce cours.

Traitement statistique de corpus

Enseignant : Damien Nouvel (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

Ce cours enseigne les notions mathématiques utiles en TAL et en textométrie. Dans un premier temps, les notions de bases en statistiques sont présentées (variables, échantillons, moyennes, écarts-types), avant d'aborder les probabilités (variables, dépendance, lois de probabilités, méthodes bayésiennes, entropie, théorie de l'information). Le cours alterne entre théorie et de nombreux exercices pratiques pour une progression s'adaptant aux étudiants ayant peu de notions en mathématiques.

Modalités de contrôle : contrôle (50%) et examen (50%)

Bibliographie

Sites web

- http://archives.limsi.fr/Individu/jardino/coursTCAN2005.pdf
- http://courses.ischool.berkeley.edu/i256/f06/sched.html
- http://web.stanford.edu/class/cs276b/
- http://norvig.com/chomsky.html
- Initiation aux méthodes de la statistique linguistique (Muller, 1993)
- Modèles statistique pour l'accès à l'information textuelle (Gaussier, Yvon, 2011)
- Statistique textuelle (Lebart, Salem, 1994)
- Foundation of Statistical NLP (Manning, Schütze, 1999)

Espace cours en ligne : http://damien.nouvels.net/fr/enseignement

Méthodes en apprentissage automatique

Enseignant : Damien Nouvel (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

L'apprentissage automatique (méthodes issues de l'intelligence artificielles) sont maintenant devenues incontournables en TAL, autant dans les milieux académiques que dans les entreprises. En s'appuyant sur le cours de statistiques, cet enseignement présente la méthodologie et différentes classes d'algorithmes pouvant être utilisés pour réaliser l'apprentissage automatique d'un modèle. De nombreux exercices sont proposés (jeux de données à l'appui, comparaison et paramétrage d'algorithmes, évaluation des résultats, etc.) sur des jeux de données réels.

Modalités de contrôle : projet (50%) et examen (50%)

Bibliographie :
- http://scikit-learn.org/stable/
- http://www.cs.waikato.ac.nz/ml/weka/

Espace cours en ligne : http://damien.nouvels.net/fr/enseignement

Document structuré et écriture numérique

Enseignant : S. Pouyllau (Paris Nanterre)
Lieu : Paris Nanterre, salle L115

Horaire : (cf. planning en ligne https://goo.gl/QF2IG8)

Le cours portera sur l'écriture numérique dans le contexte de l'open data et de la science ouverte (open science) car de nos jours les publications (articles scientifiques, thèses, mémoires, rapports, littérature grise) embarquent des données issues de bases de connaissances, de bases de données, d'API, du Web sémantique. Dans ce contexte, les publications embarquent de nos jours non seulement du texte et des références bibliographiques, mais aussi des données (sérielles, documents, etc.) et des programmes informatiques qui traitent ces dernières. Quels sont les enjeux de ces nouvelles formes de publication ? Comment « écrire » des programmes dans un document ? Quels rôles jouent les vocabulaires documentaires, mais aussi les API et les SPARQL endpoint ? Quels sont les standards qui s'en dégagent ? Est-ce une nouvelle forme de publication ? Comment pérenniser ces documents ?

Dans ce contexte l'utilisation des langages du Web, de XML à JSON, pour décrire des documents semi-structurés et des métadonnées nécessite généralement d'effectuer des transformations sur ces documents afin de les utiliser dans des systèmes d'information documentaires. L'utilisation des langages XSLT et XPATH, mais de plus en plus Python est utilisé afin de montrer le type de transformation qu'il est possible de réaliser. Les langages RDF, RDFS et OWL (utilisés dans le web sémantique) sont présentés et mis en perspective du monde XML afin d'en montrer les principales finalités dans une perspective de construction de SI documentaire. Le cours est centré sur la réalisation de "mashups" web utilisant flux XML issus de requêtes en SPARQL du moteur de recherche ISIDORE (https://isidore.science). Il permet de voir les principaux aspects de la gestion d'un projet de développement de SI documentaire.

Bibliographie
Bibliographie et ressources disponibles sur http://blog.stephanepouyllau.org/1009

Annotations sémantiques et applications en recherche d'information

Enseignant : Delphine 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 cours en ligne : oui.

Langages du Web sémantique

Enseignant Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
Horaire :

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). Les langages de modélisation et de représentation des connaissances seront présentés : OWL, RDF, SPARQL. La pratique de ces langages se fera à

l'aide de la plateforme logicielle de représentation d'ontologies Protégé. 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-IC 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 sur table.

Sémantique des textes multilingues

Enseignant : Mathieu Valette (INALCO)
Lieu : INALCO
Horaire : au S1 et au S2 (cf planning INALCO)

Ce cours s'organise en TP et peut être considéré comme le prolongement pratique du cours *Genres, textes et usages*. Il a pour objectif la réalisation collective d'une étude de sémantique comparée multilingue articulant linguistique de corpus et TAL. L'objectif est de répondre à une problématique sémantique à partir de la combinaison des outils conceptuels et théoriques de la sémantique textuelle (F. Rastier, 2011), des outils techniques de la linguistique de corpus (textométrie) et des méthodes de validation du TAL. L'application est à déterminer en concertation (fouille de texte, analyse des données subjectives, etc.).

Bibliographie

Eensoo, E., Valette, M. (2015) « Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité », *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*, Caen (France).
Pincemin B. (2011) - «Sémantique interprétative et textométrie », *Corpus*, 10, 259-269.
Rastier, F. (2011) *La mesure et le grain. Sémantique de corpus*, Paris, Champion.

Acquisition, modélisation et représentation des connaissances

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

Genres, textes, usages

Enseignant : Mathieu Valette (INALCO)
Lieu : INALCO
Horaire : au S1 (cf planning INALCO)

Ce cours s'organise comme un séminaire de recherche et a pour objet les relations entre la linguistique de corpus et le traitement automatique des langues. Nous questionnons les points de contacts et les divergences des deux sous-disciplines de manière à évaluer en quoi elles peuvent se féconder mutuellement. Puis, considérant les corpus de textes organisés de manière réflexive en discours et en genres comme des représentations des usages culturels et linguistiques, nous élaborons les conditions d'une sémantique de corpus pour le TAL à partir d'exemples issus d'applications (analyse des sentiments, fouille d'opinion, chatbot). Ce cours peut être suivi comme une introduction épistémologique au cours sémantique des textes multilingue.

Bibliographie

Cori, M., David, S. et Léon, J. (2002). « Pour un travail épistémologique sur le TAL », *TAL*, 43/3 : 7-20.
Tanguy L. et Fabre C., (2014). « Évolutions de la linguistique outillée : méfaits et bienfaits du TAL », *L'information grammaticale*, 142, p. 15-23.

Valette, M. (2016). « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée », *Proceedings of 13th International Conference on Statistical Analysis of Textual Data*, Vol. II, 697-706.

Modélisation des langues

Enseignant : Sylvain Kahane (Paris Nanterre)
Lieu : Paris Nanterre, salle L418
Horaire : lundi 9h30-12h30

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 (prosodie comprise). 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, Lareau François, 2005, Grammaire d'Unification Sens-Texte : modularité et polarisation, *Actes de TALN*, 23-32.
Kahane Sylvain, 2015, Les trois dimensions d'une modélisation formelle de la langue : syntagmatique, paradigmatique et sémiotique, *TAL*, 56.1, 39-63.
Kahane Sylvain, Kim Gerdes, 2020, *Syntaxe théorique et formelle, Volume 1 : Modélisation, unités, structures*, Language Science Press, https://langsci-press.org/catalog/book/241
Mel'čuk Igor, 1997, *Vers une linguistique Sens-Texte*, Leçon inaugurale au Collège de France, 78 p.
Mel'čuk Igor, Milićević Jasmina, 2014, *Introduction à la linguistique*, 3 volumes, Hermann.
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 cours en ligne : non.

De la modélisation au traitement automatique des données linguistiques

Enseignant : Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre, salle L418
Horaire : lundi 13h30-15h30

Une des problématiques principales du TAL est d'arriver à ce que les machines comprennent le langage humain en s'appuyant sur des indices présents directement dans les données linguistiques. Depuis quelques années, les travaux en TAL portent de plus en plus vers le repérage de l'information « implicite », déduite en quelque sorte

du corpus. Il s'agit tout d'abord de la fouille d'opinions visant la détection des avis, des émotions ou des sentiments. D'autres travaux se concentrent sur le repérage des intentions émises dans les discussions orales ou sur Internet, dans les avis sur les lieux visités, etc. Toutes ces informations qu'on cherche à détecter doivent être d'abord modélisées. Cette modélisation guidera ensuite les technologies utilisées : méthodes symboliques, apprentissage supervisé de surface ou apprentissage profond.

Bibliographie

Barbedette A., Eshkol-Taravella I. (2020), « Prédire automatiquement les intentions du locuteur dans des questions issues du discours oral spontané », TALN2020
Bunt, H., Petukhova, V., Traum, D., and Alexandersson,J. (2017). Dialogue Act Annotation with the ISO 24617-2 Standard, pages 109–135.
Eshkol-Taravella I., Kang H. J. (2019). « Observation de l'expérience client dans les restaurants », TALN2019, 1-5 juillet 2019, Toulouse, France.
Flamein H., Eshkol-Taravella I. (2020), « De la parole à la carte : repérage, analyse et visualisation automatique de la perception d'une ville », CMLF2020
Flamein H., Eshkol-Taravella I. (2020). « Noms de lieux dans le corpus de français parlé : Une approche symbolique pour un traitement automatisé », Le français moderne 2020, n.1
Grabar N., Eshkol-Taravella I. (2016). Prédiction automatique de fonctions pragmatiques dans les reformulations. *TALN2016,* Paris, France.
Grice, H. (1975). Logic and conversation. Syntax and Semantics 3 : Speech Acts. New York : Academic Press, 41-58. Reprinted in Grice, pages 22–40.
Jakobson, R. (1963). Linguistique et poétique. Essais de Linguistique Générale, pages 209–248.
Kang H. J., Eshkol-Taravella I. (2020), « Les avis sur les restaurants à l'épreuve de l'apprentissage automatique », TALN2020
Karoui, J., Benamara Zitoune, F., Moriceau, V., Aussenac-Gilles, N., and Hadrich Belguith, L. (2015). Détection automatique de l'ironie dans les tweets en français. In 22eme Conference sur le Traitement Automatique des Langues Naturelles (TALN 2015), pages 1–6, Caen, France.

Espace cours en ligne : oui.

Expérimentation et modalisation dans les humanités numériques

Enseignant : Ioana Galleron (Sorbonne nouvelle)
Lieu : Sorbonne Nouvelle
Horaire : (cf planning Sorbonne Nouvelle)

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.

Ingénierie des connaissances

Enseignant : Delphine Battistelli, Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre, salle L115
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 présentera différents modèles de représentation de connaissances (réseaux sémantiques, logiques de description, ontologies). 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
Un devoir sur table de 3h

Espace cours en ligne : oui.

Linguistique outillée et traitements statistiques

Enseignant : Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre
Horaire : (cf. planning en ligne sur plurital.org)

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 cours en ligne : oui.

Méthodologie de la recherche et épistémologie du TAL

Enseignant : Marcel Cori (Paris Nanterre)
Lieu : Paris Nanterre, salle L312
Horaire : 15h30-17h30

L'objet de ce séminaire est de fournir aux étudiants des outils conceptuels et pratiques leur permettant de mener à bien leurs travaux personnels de recherche. Il s'agit tout d'abord de dessiner un cadre général dans lequel pourront s'inscrire les approches de chacun. A cet effet, on se penchera sur l'histoire du Traitement automatique des langues, en montrant comment s'est construit ce domaine de recherche. Les grands types de méthodes qui ont cours ou ont eu cours dans le domaine seront décrits, en opposant notamment les méthodes qui s'appuient sur une analyse linguistique des données à celles fondées sur les statistiques et l'apprentissage. On essayera d'en déduire une caractérisation du domaine, d'étudier ses rapports avec des domaines liés comme l'informatique et la linguistique. Plus concrètement, les étudiants seront accompagnés dans tous les aspects que requiert un travail de recherche : détermination d'un sujet, établissement d'un état de l'art, construction d'une problématique, rédaction d'un mémoire (ou d'un article), réalisation d'un exposé.

Bibliographie

Cori, M. et Léon, J. (2002). La constitution du TAL. Étude historique des dénominations et des concepts. *TAL*, 43(3):21_55.
Léon, J. (2015*). Histoire de l'automatisation des sciences du langage*. coll. « Langages ». ENS Éditions, Lyon.
Pierrel, J.-M. (2000). *Ingénierie des langues*. Hermes.

Lexicologie, terminologie, dictionnairique

Enseignant : Kata GABOR (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

Le cours est consacré à la représentation du sens des mots et des documents pour le TAL, avec un accent particulier sur les word embeddings (représentations vectorielles). Après une introduction sur les opérations vectorielles et matricielles, nous apprendrons à construire des matrices de co-occurrences terme-document et mot-mot et d'en extraire de l'information sémantique telle que la similarité sémantique entre mots et documents, la pertinence d'un document par rapport à une requête utilisateur, les différentes relations lexicales entre mots, et des topic models sur une collection de documents. Nous aborderons aussi les différentes méthodes d'extraction terminologique et d'extraction des expressions composées. Ce cours est également une introduction au package de calcul matriciel numpy.

Apprentissage automatique

Enseignant : Jade MEKKI (Paris Nanterre)
Lieu : Paris Nanterre
Horaire : (cf. planning en ligne sur plurital.org)

Donner les outils techniques et théoriques afin de mener à bien un projet d'apprentissage automatique dans une démarche critique quant à ces derniers
Ce cours présentera toutes les étapes nécessaires à un projet d'apprentissage automatique : depuis la récole du corpus jusqu'à l'évaluation des résultats. Chaque étape sera l'occasion d'exercices de programmation à partir de données réelles. Différents modèles d'apprentissage automatique seront introduits puis développés : supervisés, semi-supervisés et non supervisés. L'objectif de cet enseignement est de vous donner les outils techniques et théoriques afin de mener à bien un projet d'apprentissage automatique dans une démarche critique quant à ces derniers.

Bibliographie

GOODFELLOW, Ian, BENGIO, Yoshua, et COURVILLE, Aaron. *Deep learning*. MIT press, 2016.
TELLIER, I. Introduction à la fouille de textes. *Université de Paris*.

Base de données et Web dynamique

Enseignant : Jade MEKKI (Paris Nanterre)
Lieu : Paris Nanterre
Horaire : (cf. planning en ligne sur plurital.org)

Savoir développer un site web dynamique depuis son interface jusqu'à sa base de données
Il s'agira de développer un site web dynamique depuis son interface jusqu'à sa base de données. Vous verrez toutes les étapes nécessaires à l'élaboration d'une application web et concevrez un cahier des charges du projet. Vous apprendrez notamment à analyser l'existant ainsi que des besoins clients, à décrire des fonctionnalités attendues, des utilisateurs, des exigences techniques et enfin à établir des modalités de validation des prestations. Vous développerez cette application web dans une démarche comparative afin de comprendre les avantages et désavantages des différents outils utilisés.

Bibliographie

JAZAYERI, Mehdi. Some trends in web application development. In : *Future of Software Engineering (FOSE'07)*. IEEE, 2007. p. 199-213. PROKOFYEVA, Natalya et BOLTUNOVA, Victoria. Analysis and Practical Application of PHP Frameworks in Development of Web Information Systems. *Procedia Computer Science*, 2017, vol. 104, p. 51-56.
RIGAUX, Philippe. *Pratique de MySQL et PHP: Conception et réalisation de sites web dynamiques*. Dunod, 2009.

TAL et linguistique de corpus

Enseignant : Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre
Horaire : (cf. planning en ligne sur plurital.org)

Découvrir les recherches actuelles dans les domaines du TAL et de la linguistique de corpus et la (les) manière(s) dont elles s'interpénètrent.
Les relations entre la linguistique de corpus (LC) et le TAL sont de nature complexe du fait de méthodologies et d'objectifs finaux bien distincts alors même que tous deux sont concernés aujourd'hui par l'analyse de (grands) corpus. Dans ce séminaire, les enseignants et chercheurs montreront dans quelle mesure ces deux domaines peuvent et doivent s'interpénétrer pour une meilleure prise en compte de conceptualisations strictement linguistiques, et pour démontrer ainsi qu'il est non seulement possible mais en fait indispensable pour des résultats robustes en TAL de (re)mettre au centre des préoccupations la langue, vue à travers des corpus de types variés.

http://pluriTAL.org

Sorbonne Nouvelle

inalco
Institut national des langues et civilisations orientales

Université Paris Nanterre

Contacts

Kahane Sylvain
- Paris Ouest
- MoDyCo
- sylvain@kahane.fr

Valette Mathieu
- INALCO
- CRIM
- mathieu.valette@inalco.fr

Cédric Gendrot
- Sorbonne Nouvelle
- CLESTHIA
- cedric.gendrot@sorbonne-nouvelle.fr