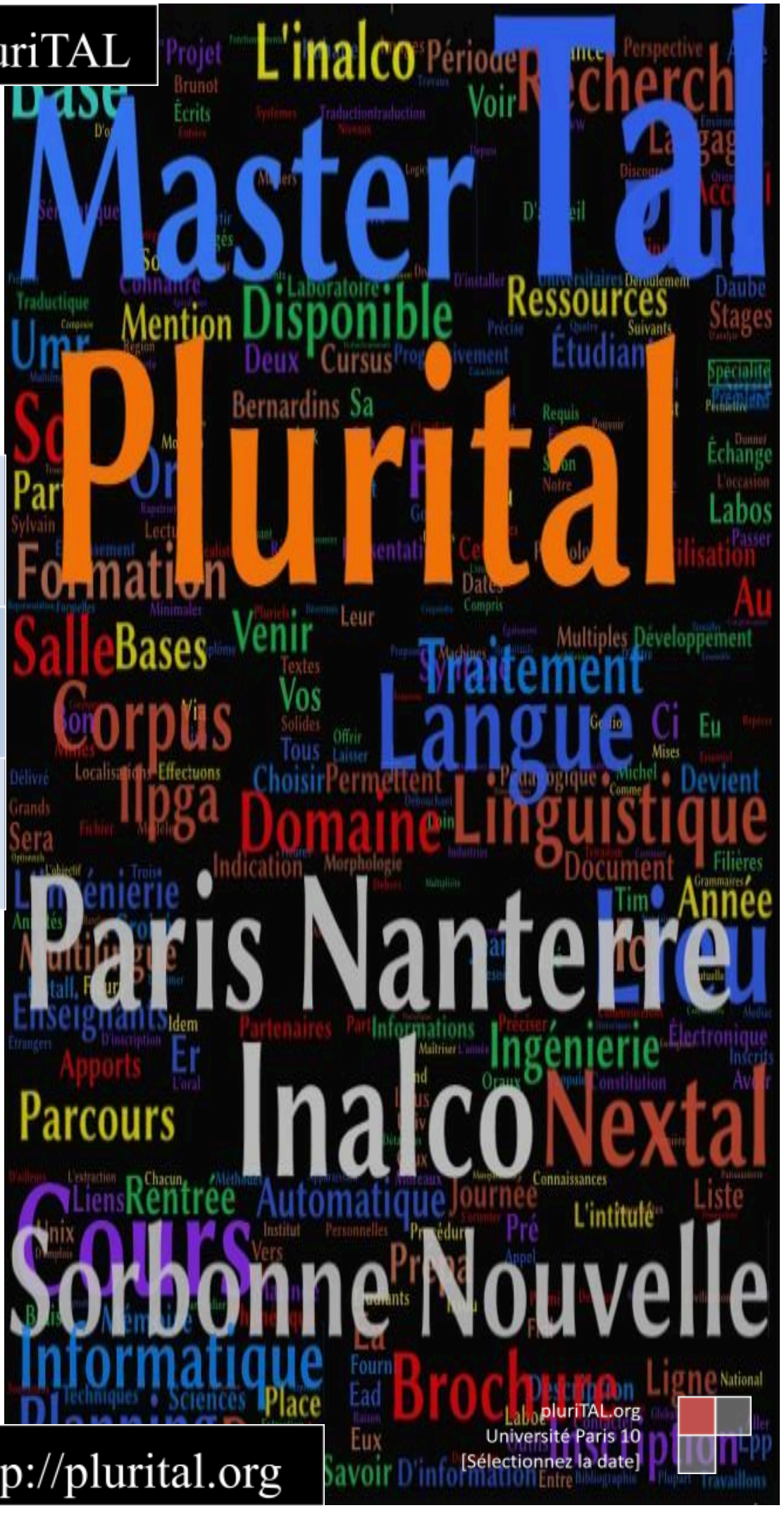
pluriTAL



http://plurital.org

pluriTAL.org
Université Paris 10
[Sélectionnez la date]

Formation proposée par
l'Université Paris Nanterre
l'Université Sorbonne Nouvelle
l'INALCO

MASTER

Mention : *Traitement Automatique des Langues*

Responsables:

Mathieu Valette (INALCO)
Iris Taravella / Delphine Battistelli (Paris Nanterre)
Cédric Gendrot (Sorbonne Nouvelle)

Commission pédagogique :

Mathieu Valette (INALCO)
Cédric Gendrot (Sorbonne Nouvelle)
Iris Taravella / Delphine Battistelli (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 : ER-TIM *Textes, Informatique, Multilinguisme* (EAD 2540, Inalco), Lattice, *Langues, Textes, Traitements informatiques, Cognition* (UMR 8094, ENS/PSL, CNRS & Sorbonne Nouvelle), LPP, *Laboratoire de Phonétique et Phonologie* (UMR 7018, CNRS & Sorbonne Nouvelle), MoDyCo, *Modèles, Dynamiques, Corpus* (UMR 7114, CNRS & Paris Nanterre).

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 notamment 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. 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.

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, compréhension et utilisation des réseaux de neurones profonds et des transformers ;
- Savoirs en informatique : programmation et algorithmique spécifique, bases de données, document structuré (XML) ;
- Maîtrise en réception puis en production de l'anglais scientifique.
- Maîtrise des concepts et utilisation des algorithmes d'intelligence artificielle (CNN, RNN, LLM, etc.)

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).

Sorbonne Nouvelle — inalco Institut national des langues et civilisations orientales — Université Paris Nanterre

Sommaire

Contacts

Direction pédagogique

Mathieu Valette (mathieu.valette@inalco.fr)
Cédric Gendrot (cedric.gendrot@sorbonne-nouvelle.fr)
Iris Taravella (ieshkolt@parisnanterre.fr)
Delphine Battistelli (dbattist@parisnanterre.fr)

Secrétariats administratifs

Inalco :
Bassir HAMID
Responsable de formation
Filière Traitement automatique des langues (TAL)
Masters TI & SDL
Bureau N° 3.22A
65 rue des grands moulins, 75013 Paris
01.80.71.11.36
bassir.hamid@inalco.fr

Sorbonne Nouvelle (ILPGA) :
Marie-Claudette Baremon
Bureau A513
01.87.86.13.18
marie-claudette.baremon@sorbonne-nouvelle.fr

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

Réunions de rentrée – Début des cours

Journée d'accueil du MASTER : (dates disponibles en ligne sur le site pluriTAL)

Sorbonne Nouvelle - campus Nation
- 10 septembre 2024 de 10h à 11h : réunion M2 en C117
- 10 septembre 2024 de 11h15 à 12h30 : réunion M1 en C117

pluriTAL : http://plurital.org

- 10 septembre 2024 de 11h15 à 13h30 : Atelier nexTAL pour les M2 R&D, de 11h15 à 13h30 en B112
- 10 septembre 2024 à partir de 14h : intégration M1 en C117

- 11 septembre 2024 : install party **obligatoire** pour les M1 ou les nouveaux M2, de 10h à 16h en C117

Début des cours : (dates disponibles en ligne sur le site pluriTAL)

INALCO : https://www.inalco.fr/calendrier-universitaire : 16 septembre 2024
Nanterre : https://www.parisnanterre.fr/calendrier-universitaire-2024-2025 : 16 septembre 2024 (M2), 23 septembre 2024 (M1)
Sorbonne Nouvelle : http://www.univ-paris3.fr/le-calendrier-universitaire-116398.kjsp : 23 septembre 2024

Liste PluriTAL (liste de diffusion)

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

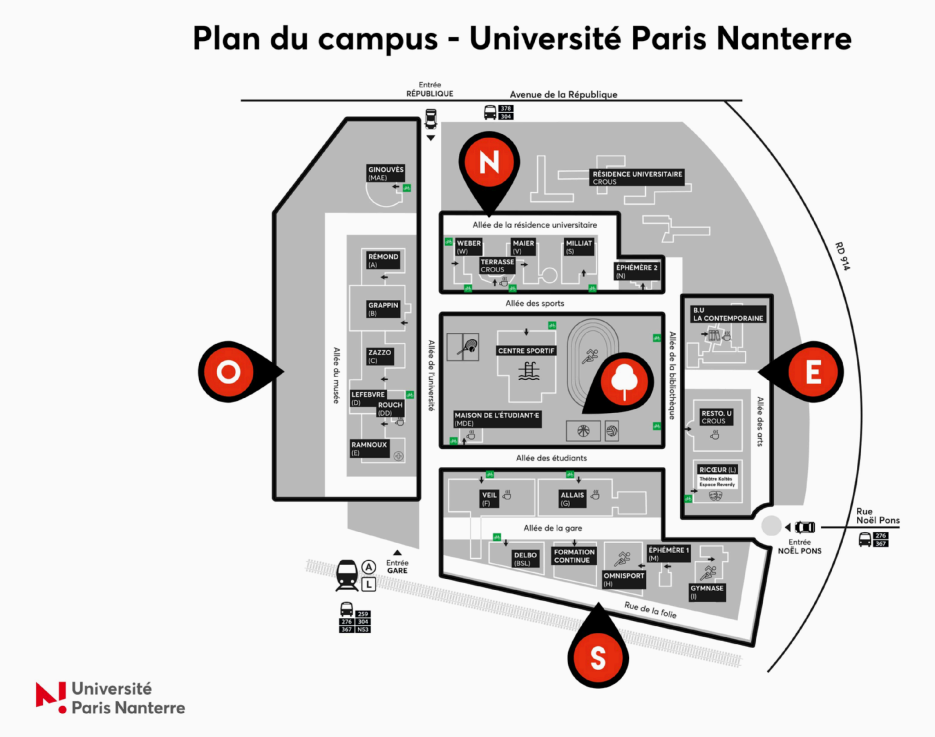
Voir la page « Liste pluriTAL » sur la page web du MASTER (site pluriTAL).

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

Localisations

Paris Nanterre

Pour se rendre à 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.
https://gestion.parisnanterre.fr/informations-pratiques-et-acces.html



Sorbonne Nouvelle / ILPGA

Campus Nation:
8 avenue de Saint-Mandé - 75012 Paris. Ouvert du lundi au vendredi de 7h30 à 21h, le samedi 7h30 à 18h, fermé le dimanche et les jours fériés

https://sites.google.com/sorbonne-nouvelle.fr/campus-nation/accueil

INALCO

Les cours de l'INALCO sont le plus souvent sur le campus des Grands Moulins, 65 rue des Grands Moulins, 75013 Paris.
Les bâtiments de l'**ERTIM** se situent au 2 rue de Lille - 75007 Paris
https://www.inalco.fr/sorienter

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 : dbattist@parisnanterre.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, celui qui vous a accepté**) :

- Inscription Sorbonne Nouvelle :
http://www.univ-paris3.fr/inscription-administrative-170406.kjsp

- Inscription Paris Nanterre :
https://candidatures-inscriptions.parisnanterre.fr/

- Inscription à l'INALCO
https://www.inalco.fr/inscriptions-administratives

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 3 responsables pédagogiques avant de vous inscrire :cedric.gendrot@sorbonne-nouvelle.fr, mathieu.valette@inalco.fr, ieshkolt@parisnanterre.fr

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

- Inscription Sorbonne Nouvelle :
  http://ecandidat.univ-paris3.fr/
- Inscription à l'INALCO
  mathieu.valette@inalco.fr
- Inscription à Nanterre
  https://ecandidat.parisnanterre.fr/#accueilView

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

Contact : 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 » (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 Nanterre	INALCO
(*cf* secrétariat ILPGA)	(*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 :Iris Taravella (ieshkolt@parisnanterre.fr), Nanka Stoyanov (nanka.stoyanova@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 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. De même, il n'y a pas de compensation entre les blocs. Il n'y a compensation qu'au sein de chaque bloc.

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 2024-2025

Rentrée

Voir section : Réunions de rentrée – Début des cours

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

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

Les emplois du temps sont disponibles ici. Pour l'instant, l'emploi du temps est identique à celui de l'année 2023-2024, mais reste sujet aux changements. Ce lien sera mis à jour régulièrement.

https://docs.google.com/spreadsheets/d/1NPydiKWIS9-UA_f_jgbY22SFSniofWMI/edit?gid=930560273#gid=930560273

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

Fin d'Année universitaire

Le jeudi 26 juin 2025 auront lieu une journée de présentation des sujets des M2 R&D (soutenances, pré-soutenances, travaux en cours en fonction de l'avancement des étudiants)

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
32000	Ingénieur NLP, métiers de l'IA, préparation données, évaluation, etc

Relation avec les milieux professionnels :

EDF, Mondeca, Arisem, XEROX, PUF, Larousse, Le Robert, Performix, Syllabs, Quensis, Logosapience, Exalead, Thales, France Telecom, Bowne Global Solutions, Softissimo, TRADOS, SDL, LIP6, ATILF, Synapse, Qwant, Lunii, ATDI, Deezer, Ubisoft; etc.

Organisation globale des enseignements du master

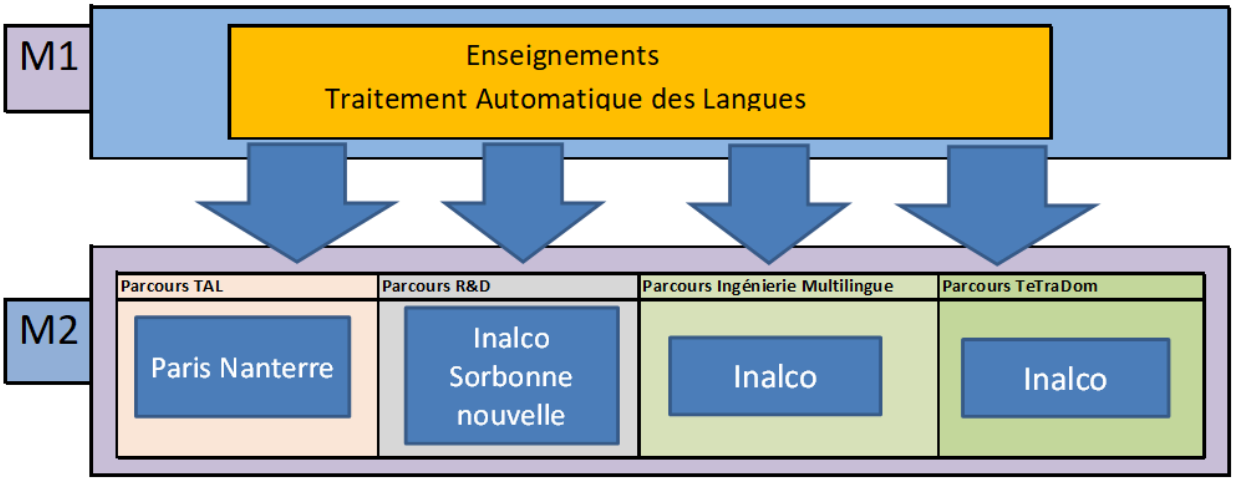
Le master s'organise selon une première année composée d'enseignements en Traitement Automatique des langues puis quatre parcours en M2, adossés chacun à un ou plusieurs établissements universitaires :

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

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

- le parcours *TeTraDom*, 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

Master 1, 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 / Nanterre**
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

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

Parcours T : M2 Technologies de la traduction et traitement des données multilingues (TeTraDom), Inalco

Parcours I : M2 Ingénierie Multilingue, Inalco

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@ens.fr
Université / UFR de rattachement : Université Sorbonne nouvelle, ILPGA
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 : BRIGLIA Prénom : Andrea
Email : andrea.briglia@grenoble-inp .fr
Université / UFR de rattachement : Université Sorbonne nouvelle, ILPGA
Equipe de recherche : Lattice, UMR 8094

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

Nom : DUPONT Prénom : Yoann
Email : yoann.dupont@sorbonne-nouvelle.fr
Université / UFR de rattachement : Université Sorbonne nouvelle, ILPGA
Equipe de recherche : Lattice, UMR 8094

Nom : TARAVELLA Prénom : Iris
Email : ieshkolt@parisnanterre.fr
Université / UFR de rattachement : Université Paris - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114

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 : lgrobol@parisnanterre.fr
Université / UFR de rattachement : Université Paris - Nanterre, UFR Phillia
Equipe de recherche : MoDyCo, UMR 7114

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

Nom : GABOR Prénom : Kata
Email : kata.gabor@inalco.fr
Université / UFR de rattachement : INaLCO

Equipe de recherche : ER-TIM (EAD 2540)

Nom : MAGISTRY Prénom : Pierre
Email : pierre.magistry@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 : VALETTE Prénom : Mathieu
Email : mvalette@inalco.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

L7DN002 Corpus arboré et parsing

Enseignant : Aleksandra Miletić (Paris Nanterre) & Santiago Herrera (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, son utilisation pour le TAL et la linguistique ainsi qu'une introduction à l'analyse syntaxique automatique. Les principales notions de syntaxe (unité syntaxique, tête, dépendance, constituant, relation syntaxique) sont introduites. Le guide d'annotation UD (Universal Dependencies) et SUD (Surface-Syntactic UD) est présenté et chaque étudiant procède à l'annotation d'un fragment de corpus de français. L'exploration du corpus et l'analyse syntaxique seront effectuées à l'aide de grammaires de réécriture de graphes et de bibliothèques Python.

Bibliographie

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*, Language Science Press, 2022, open edition.
Kahane, Sylvain, Nicolas Mazziotta (2022). Les corpus arborés avant et après le numérique. *Revue TAL*, *63*(3), 63-88.
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.
De Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, Daniel Zeman (2021). Universal Dependencies. *Computational linguistics*, *47*(2), 255-308.
Tesnière Lucien, *Eléments de syntaxe structurale*, Klincksieck, 1959.

Ressources

ArboratorGrew, arborator.ilpga.fr.
Universal Dependencies Treebanks, universaldependencies.org.
Grew-match, match.grew.fr.

Modalités de contrôle

Contrôle continu : La moyenne de l'année est composée d'un projet de développement de treebank, 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.

Grammaires formelles

Enseignant : Loïc Grobol (Paris Nanterre)

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

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 L 308, bâtiment Ricoeur
Horaire : jeudi 10h30-12h30

Le cours propose une initiation aux questions méthodologiques posées par la démarche de modélisation en linguistique. Il vise à montrer également que le TAL a besoin de modèles linguistiques rigoureux à implémenter pour optimiser les applications. Dans la construction d'un modèle sont ainsi mis à contribution des mathématiques de manière directe ou indirecte (via le recours à des logiciels comme Unitex en particulier pour ce cours). Deux types d'unités linguistiques sont analysées à partir de leurs modes d'expression dans des textes : 1) les unités adverbiales dites temporelles ; 2) les unités lexicales dites émotionnelles (une année sur deux, ce sont les unités lexicales et grammaticales relevant de formes de distanciation qui sont étudiées).

Bibliographie indicative

Desclés, J.-P. (1969), "Linguistique et Mathématiques", l'Homme, 9-3, pp. 93-99

Tanguy, L. (2012), "Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes", HDR Linguistique. Université Toulouse le Mirail - Toulouse II.

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 DST de 2h à mi-parcours.
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 : Nation salle B326
Horaire : mercredi 13h00-15h00

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 très rapidement 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 / Praat). Les approches neuronales les plus récentes seront également abordées en fin de semestre. La validation consistera en un partiel à mi-semestre, puis un devoir à rendre en fin de semestre. Il est nécessaire de venir avec son ordinateur portable afin de réaliser les exercices en cours chaque semaine.

- Text to Speech Synthesis : https://www.youtube.com/watch?v=0GmFA5t_IKo
- Speech and Language Processing (Jurafsky & Martin) : https://web.stanford.edu/~jurafsky/slp3/
- An introduction to text-to-speech synthesis (Dutoit)
https://archive.org/details/introductiontote0000duto/page/n9/mode/2up
- Neural Text-to-Speech Synthesis (Xu Tan)

Espace cours en ligne : https://nextcloud.laboratoirephonetiquephonologie.fr/index.php/s/tGgTKeTdbWDdWkt

---

Programmation et projet encadré (semestre 1)

Enseignant : Yoann Dupont (Sorbonne nouvelle), Pierre Magistry (INALCO)
Lieu : Sorbonne Nouvelle, salle à préciser
Horaire : mercredi 9h-12h

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

Ressources
GitHub (créer un compte avant le premier cours) : https://github.com

Modalités de contrôle
Contrôle continu : une note individuelle et une note de projet.

Espace cours en ligne : espace icampus https://icampus.univ-paris3.fr

---

Programmation orienté objet C++

Enseignant : Florent Jacquemard (Inria)
Lieu : INALCO salle 3.16
Horaire : semestre 1: jeudi 9h-12h, semestre 2: à préciser

Le cours du premier semestre est une initiation à la programmation impérative, avec l'introduction d'éléments de base: variables, types, instructions de contrôle, opérateurs et expressions, fonctions... Au second semestre sont abordés des concepts fondamentaux de la programmation orienté objet, tels qu'abstraction, encapsulation, héritage et polymorphisme.

Aucune connaissance préalable en programmation n'est requise. L'illustration des concepts présentés en cours, ainsi que les exercices, sont en langage C++, mais ce cours a vocation a introduire des principes généraux de programmation orientée objet, valables dans d'autres langages tels que Python ou Java.

Espace cours en ligne : dépot git https://gitlab.inria.fr/jacquema/ooptal et MOODLE INALCO

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

Le cours vise à introduire les bases de données relationnelles. Les étudiants apprendront à construire les bases de données avec les outils disponibles et à formuler les requpetes en utilisant le langage SQL. Dans une seconde partie, nous nous intéresserons aux bases de données basées sur les graphes comme Neo4jDes 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 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.

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

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

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 : 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, étiquetage 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.

Documents structurés

Enseignant : Michel Jacobson (Sorbonne nouvelle)
Lieu : Nation
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 : Yoann Dupont (Sorbonne nouvelle), Pierre Magistry (INALCO)
Lieu : Sorbonne Nouvelle, salle à préciser
Horaire : mercredi 10h30-13h30

Cf descriptif du premier semestre.
URL : http://www.tal.univ-paris3.fr/cours/masterproj.htm

Ressources
GitLab (*pas* GitHub du S1. Créer un compte avant le premier cours) : https://gitlab.com

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), Mathieu Dehouck (Sorbonne nouvelle)
Lieu : Semestre 1 : Paris Nanterre, salle à préciser PX
Semestre 2 : Sorbonne Nouvelle, salle à préciser
Horaire : Semestre 1 : vendredi 13h30-15h30
Semestre 2 : Mercredi 14h00-16h00

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 acquérons 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 scripts 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'une moyenne des notes pour les exrcices rendus et d'un examen sur table de 2h.

Espace cours en ligne : oui.

Programmation et algorithmique 2 (Mathieu Dehouck)
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, algorithmes récursifs) 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.

Modalités de contrôle : contrôle continu (6 x 10%), contrôle final (40%)

Espace cours en ligne : espace icampus https://icampus.univ-paris3.fr

---

Machine creativity and text generation

Enseignant : Andrea Briglia (Sorbonne Nouvelle)
Lieu : salle B307 à l'ILPGA (Nation)
Horaire : mercredi de 15h00 à 17h00 , semestre 1

Modalité de contrôle : trois devoirs à rendre (à partir de la 4ème séance), analyse critique d'article scientifique (anglais ou français) en binôme, projet à rendre avant la fin du S1, en binôme.

Espace cours en ligne : espace icampus https://icampus.univ-paris3.fr

Descriptif : Le cours commence par un CM où l'on définit le concept de créativité et on propose des exemples de génération de texte. Ensuite, des rappels de programmation Python seront proposés en fonction du niveau de la classe (sur Jupyter notebook, tutoriel disponible). Analyse des extraits générés automatiquement selon les critères de cohérence, respect de la syntaxe et idiomaticité. Importance du corpus d'entraînement (taille, biais, *overfitting*). Chaînes de Markov (librairie *markovify*) et HMM, méthodes probabilistes et méthodes neuronales. Des RNN aux transformeurs (encodage et décodage, mécanisme d'attention, différence entre entraînement et *fine-tuning*) Cas d'étude : évaluation automatique et humaine de l'expressivité en lecture, qui prend en compte des éléments à la fois textuels et oraux. Notions de vérité de terrain, *crowdsourcing*, accord inter-évaluateurs (calcul du k de Cohen, f-score).

Bibliographie:

Oliveira, H.G. (2009). Automatic generation of poetry: an overview.
Automatic Grammatical Error Correction for Sequence-to-sequence Text Generation: An Empirical Study (Ge et al., ACL 2019)
Yeo, C., & Chen, A. (2020). Defining and Evaluating Fair Natural Language Generation. *ArXiv, abs/2008.01548*.
Malmi, E., Takala, P., Toivonen, H., Raiko, T., & Gionis, A. (2016). Dopelearning: A computational approach to rap lyrics generation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 195-204).
François Portet, Albert Gatt, Jim Hunter, Ehud Reiter, and Somayajulu Sripada. 2009. Le projet BabyTalk : génération de texte à partir de données hétérogènes pour la prise de décision en unité néonatale. In *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 121–130, Senlis, France. ATALA.
Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B.C. (2005). Freudbot: An Investigation of Chatbot Technology in Distance Education.
Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799

Fouille de textes

Enseignant : Yoann Dupont (Sorbonne Nouvelle)
Lieu : Sorbonne Nouvelle, B307
Horaire : jeudi 13h-15h (S2)

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...).

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

Espace cours en ligne : espace icampus https://icampus.univ-paris3.fr

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%)

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

---

Sémantique lexicale et sémantique textuelle

Enseignant : Mathieu Valette (INALCO)
Lieu : INALCO
Horaire : au S2 (Jeudi 9h30-11h00, cf planning INALCO)

Ce cours propose des bases théoriques et critiques 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 et d'une approche référentielle du lexique de ce qui relève de la sémantique et d'une approche pratique (ou praxéologique) du lexique. 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, néosémie, 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.
Rastier, François, Mathieu Valette (2009) « De la polysémie à la néosémie », *Le français moderne*, 77, pp. 97-116.
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.

Langages de script

Enseignant : Alexandre Roulois (IRIF, université Paris-Cité, CNRS)
Lieu : INALCO, Grands Moulins, salle 5.10
Horaire : au S1 (Lundi 12h00-14h00, cf planning INALCO)

Entièrement dédié à l'apprentissage du langage Python, notre parcours connaîtra deux temps forts. Dans le premier, nous aborderons les fondamentaux du langage, de la découverte des structures de données à leur manipulation dans un programme grâce aux structures de contrôle (conditions, boucles…) ; puis, nous serons amené·es à mobiliser certaines des bibliothèques logicielles les plus connues pour résoudre des tâches classiques en traitement automatique du langage : tokenisation, lemmatisation, vectorisation de mots, évaluation de l'importance d'un terme… De nombreux exercices seront proposés tout au long des séances afin de consolider les apprentissages.

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

Espace de cours en ligne : https://github.com/Alex-bzh/python-M1TAL

Descriptif et horaires des cours du master 2ème année

Sémantique computationnelle

Enseignant : Pascal Amsili
Lieu : Sorbonne Nouvelle, ILPGA B324 (Sorbonne Nouvelle)
Horaire : vendredi, 10h-12h

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 sémantique computationnelle, comme la résolution de coréférences, ou la détection des inférences naturelles.
Voir la page du cours 2023/24 pour se faire une idée du contenu:
https://lattice.cnrs.fr/amsili/Ens24/LYST001.php
Plan du cours :

- Ch1: Sémantique lexicale
- Ch2: L'hypothèse distributionnelle
- Ch3: Applications
- Ch4: Réduction de dimensionnalité
- Ch5: Plongements lexicaux

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 : ILPGA
Horaire : mercredi 17h30-19h30 salle B307

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%)

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

---

Outils de TAL
Enseignant : Damien Nouvel (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

Ce cours propose un cas d'utilisation des outils TAL standard sur une problématique dédiée, qui change chaque année. L'objectif principal est d'aborder une problématique inconnue, d'identifier les différents traitements à mettre en œuvre, de répartir le travail par groupes, d'implémenter les algorithmes en coordination. Le travail est réalisé en séance et en dehors des cours est l'objet d'une évaluation. Une autre évaluation est réalisée en cours, qui porte sur la programmation de quelques traitements TAL simples en temps limité.

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

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

---

Méthodes en apprentissage automatique
Enseignant : Pierre Magistry (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

---

Documents structurés
Enseignant : Rim Abrougui (INALCO)
Lieu : INALCO
Horaire : (cf planning INALCO)

Ce cours se concentre sur l'évaluation et le ré-entraînement des outils de traitement automatique des langues (TAL), en abordant les bonnes pratiques d'évaluation et la réplicabilité des résultats. Il alterne entre des séances théoriques sur les méthodes d'évaluation et la lecture d'articles scientifiques, et des travaux pratiques où les étudiants ré-évaluent et ré-entraînent des modèles sur des données spécifiques.

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

Arbres, graphes

Enseignant : Loic Grobol (Paris Nanterre)
Lieu : Paris Nanterre, salle à préciser
Horaire : (cf. planning en ligne https://goo.gl/QF2IG8)

Attention, contenu du cours différent du titre

Partie 2 du cours apprentissage automatique. Réseaux de neurones

---

Acquisition, modélisation et représentation des connaissances

Enseignant : Ons Aouina (Inserm)
Lieu : Inalco, rue de Lille, rue de lille
Horaire : Mardi 14h-16h

Ce cours explore les techniques avancées de représentation des connaissances sous forme de thésaurus, d'ontologies, de terminologies et de réseaux sémantiques. Les étudiants apprendront à modéliser des connaissances avec ces structures, à publier des données liées sur le Web via RDF, à interroger ces données avec SPARQL, et à raisonner sur les ontologies en utilisant RDFS, OWL et SKOS. L'objectif est de former les étudiants aux technologies du Web sémantique.

Le cours inclut également des travaux pratiques qui permettront aux étudiants de mettre en œuvre les concepts étudiés. Ils commenceront par la création d'ontologies, où ils apprendront à modéliser des connaissances à l'aide d'outils comme Protégé, en définissant les concepts, relations et instances pour un domaine spécifique. Ensuite, un TP sur le requêtage d'ontologies avec SPARQL leur permettra d'interroger ces données de manière précise. Enfin, un TP sur le raisonnement sur les ontologies (RDFS, OWL) initiera les étudiants aux techniques de raisonnement automatique pour inférer de nouvelles connaissances et enrichir les ontologies créées.

Bibliographie :
Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199-220.
Un texte fondamental pour comprendre les bases de la création d'ontologies.

Antoniou, G., & Van Harmelen, F. (2004). A Semantic Web Primer. MIT Press.
Un ouvrage de référence sur les technologies du Web sémantique, y compris RDF, OWL et SPARQL.
Staab, S., & Studer, R. (Eds.). (2009). Handbook on Ontologies. Springer Science & Business Media.
Un recueil complet couvrant les aspects théoriques et pratiques des ontologies.
Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1. W3C Recommendation.
Un document officiel qui définit le cadre RDF(S) pour la modélisation des connaissances.
Hitzler, P., Krötzsch, M., Rudolph, S., & Sure, Y. (2008). Semantic Web: Concepts, Technologies, and Applications. Springer.
Une introduction complète aux concepts et aux technologies du Web sémantique, avec des exemples concrets d'applications.
W3C (2013). SPARQL 1.1 Query Language. W3C Recommendation.
Le document officiel décrivant le langage de requête SPARQL, essentiel pour interroger des ontologies.

Annotations sémantiques et applications en recherche d'information

Enseignant : Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, Bâtiment BFC, salle 408
Horaire : lundi, 9h30-12h30

Ce cours présente des méthodes, modèles et applications propres à appréhender un niveau d'analyse et d'annotation sémantique des textes. Il exploite le rapprochement manifeste ces dernières années entre les domaines du TAL et de la Recherche d'Information (RI) pour ce qui concerne en particulier la fouille textuelle et/ou l'accès au contenu informationnel des textes. Tout au long des séances, nous allons explorer des dimensions sémantiques d'annotation des textes (tout particulièrement celles en lien avec les catégories de temps, aspect, modalité et énonciation - sous l'acronyme TAME). Ces dimensions répondent toutes à des besoins en RI formulés par des communautés d'utilisateurs données (ex. : historiens, biologistes, journalistes, …). Ces besoins sont formulés à l'aide d'une terminologie assez fluctuante, oscillant entre termes de désignation des catégories linguistiques elles-mêmes (ex. TAME) et formulations intuitives de besoins en accès au contenu informationnel des textes (ex. : ordonner temporellement et/ou dater des évènements, repérer des opinions, typer des émotions, dire si une information est certaine ou incertaine, …).

Bibliographie
BATTISTELLI, D., *Linguistique et recherche d'information : la problématique du temps*, Hermes
CONDAMINES A. (ed), 2005 : *Sémantique et corpus*. Londres : Hermes

Modalités de contrôle
Contrôle continu : 1 DM + 1 DST + 1 exposé en groupe .
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. 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, Lundi, 15h00-16h30, salle LO.01, rue de Lille (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. La tâche est décidée au cours du 1er semestre en concertation avec les étudiants. L'année 2024-2025 sera consacrée à la question des cultures humaines en TAL.

Bibliographie

Pincemin B. (2011) - «Sémantique interprétative et textométrie », *Corpus*, 10, 259-269.
Rastier, F. (2011) *La mesure et le grain. Sémantique de corpus*, Paris, Champion.
Valette, M. (2018) "Elements of a Corpus Semantics for Humanities. Application to the Classification of Subjective Texts", D. Compagno, eds., Quantitative Semiotic Analysis, Springer International Publishing.

Acquisition, modélisation et représentation des connaissances

Enseignant : Ons Aouina (Inserm)
Lieu : Inalco, rue de Lille, rue de lille
Horaire : Mardi 14h-16h (cf planning INALCO

Ce cours explore les techniques avancées de représentation des connaissances sous forme de thésaurus, d'ontologies, de terminologies et de réseaux sémantiques. Les étudiants apprendront à modéliser des connaissances avec ces structures, à publier des données liées sur le Web via RDF, à interroger ces données avec SPARQL, et à raisonner sur les ontologies en utilisant RDFS, OWL et SKOS. L'objectif est de former les étudiants aux technologies du Web sémantique.

Le cours inclut également des travaux pratiques qui permettront aux étudiants de mettre en œuvre les concepts étudiés. Ils commenceront par la création d'ontologies, où ils apprendront à modéliser des connaissances à l'aide d'outils comme Protégé, en définissant les concepts, relations et instances pour un domaine spécifique. Ensuite, un TP sur le requêtage d'ontologies avec SPARQL leur permettra d'interroger ces données de manière précise. Enfin, un TP sur le raisonnement sur les ontologies (RDFS, OWL) initiera les étudiants aux techniques de raisonnement automatique pour inférer de nouvelles connaissances et enrichir les ontologies créées.

Bibliographie :

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199-220.
Un texte fondamental pour comprendre les bases de la création d'ontologies.
Antoniou, G., & Van Harmelen, F. (2004). A Semantic Web Primer. MIT Press.
Un ouvrage de référence sur les technologies du Web sémantique, y compris RDF, OWL et SPARQL.
Staab, S., & Studer, R. (Eds.). (2009). Handbook on Ontologies. Springer Science & Business Media.
Un recueil complet couvrant les aspects théoriques et pratiques des ontologies.
Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1. W3C Recommendation.
Un document officiel qui définit le cadre RDF(S) pour la modélisation des connaissances.
Hitzler, P., Krötzsch, M., Rudolph, S., & Sure, Y. (2008). Semantic Web: Concepts, Technologies, and Applications. Springer.
Une introduction complète aux concepts et aux technologies du Web sémantique, avec des exemples concrets d'applications.

W3C (2013). SPARQL 1.1 Query Language. W3C Recommendation.
Le document officiel décrivant le langage de requête SPARQL, essentiel pour interroger des ontologies.

Genres, textes, usages

Enseignant : Mathieu Valette (INALCO)
Lieu : INALCO
Horaire : au S1, Lundi, 13h30-15h00, salle LO.01, rue de Lille (cf planning INALCO)

Cultures et Large Language Models
Le cours s'organise comme un séminaire de recherche. Le projet poursuivi depuis quelques années est de travailler aux points de tension épistémologiques entre la linguistique en tant que Science Humaine et Sociale (et science de la culture) et le TAL comme technologie. L'année 2024-2025 sera consacrée à la problématique des cultures humaines dans le TAL et dans l'IA génératives (grands modèles de langues) : quels sont les enjeux de la prise en compte du *fait culturel* dans les recherches en TAL ? quelles sont les propositions actuelles en termes d'implémentation de la culture en TAL, quelles en sont les soubassements politiques et culturels ? les qualités et les limites ? Le séminaire s'organisera autour de 3 axes : (i) le passé : débuts de la traduction automatique, "inculte" par méthode – minimalisme culturel de la modélisation sémantique – lexiques et les ontologies – corpus vs data) ; (ii) le présent : les nouveaux réductionnismes culturels des LLMs (cultureLLMs (Li et al. 20024), Constitutional AI ((Bai et al. 2022), Perspectivism (Cabitza et al., 2023), etc.), (iii) le futur : objets pseudo-culturels et fragmentation de l'environnement symbolique

Bibliographie
Bai et al. (2022) Constitutional AI: Harmlessness from AI Feedback, arXiv:2212.08073v1 [cs.CL] 15 Dec 2022.
Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. Proceedings of the AAAI Conference on Artificial Intelligence, 37(6), 6860-6868.
Léon, J. (2015). *Histoire de l'automatisation des sciences du langage*. ENS Éditions.
Li et al. (2024) CultureLLM: Incorporating Cultural Differences into Large Language Models, https://doi.org/10.48550/arXiv.2402.10946.
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.
Valette, M. (2024a) « Guerre cognitive, culture et récit national », *Ingénierie cognitique*, ISTE OpenScience, Londres, Vol. 24-7, 6-12.
Valette, M. (2024b) "What Does Perspectivism Mean? An Ethical and Methodological Counter-criticism", *NLPerspectives 2024@LREC-COLING 2024*, 21 May, 2024, pages 111-115.

Modélisation des langues

Enseignant : Sylvain Kahane (Paris Nanterre)
Lieu : Paris Nanterre, salle BFC 408
Horaire : mercredi 13h30-16h30

Nous nous intéresserons à différents modèles linguistiques depuis les grammaires formelles du 20e siècle aux LLM d'aujourd'hui. L'emphase est mise sur les modèles interprétables par des humains et qui permettent aux linguistes de comprendre la structure des langues, de comparer les langues entre elles et de faire de la typologie des langues. L'approche est résolument orientée vers les corpus et les méthodes, automatisées ou non, qui permettent de faire émerger de la connaissance grammaticale à partir de données linguistiques. Ce cours est aussi l'occasion de redéfinir les concepts utiles à l'élaboration d'un modèle linguistique et les liens logiques entre eux.

Bibliographie
Bresnan Joan, 2001, *Lexical-Functional Syntax*, Blackwell.
Goldberg Adele, 1995, *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Herrera S., Corro C., Kahane S. (2024) Sparse Logistic Regression with High-order Features for Automatic Grammar Rule Extraction from Treebanks, *Proceedings of LREC-Coling*.
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
Haspelmath Martin (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3) 663-687.
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.

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

Espace cours en ligne : oui.

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, bâtiment BFC, salle 408
Horaire : (cf. planning en ligne sur plurital.org)

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 / Karin Heidelmayr (Paris Nanterre)
Lieu : Paris Nanterre, bâtiment BFC, salle 408
Horaire : (cf. planning en ligne sur plurital.org)

Un des objectifs du cours est de permettre aux étudiants de maîtriser les principaux outils statistiques utilisés en sciences humaines et sociales et plus spécifiquement en linguistique afin d'être capable de les utiliser dans un contexte 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, etc.) 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. Il permet également un regard critique et concret sur des questionnements à la croisée du TAL et de la psycholinguistique (ex. analyse de documents textuels produits par/pour des enfants, analyse de documents textuels produits par des patients atteints d'une maladie neuro-dégénérative, etc.).

Bibliographie indicative

Fuchs, C., 2014. "Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs". *L'information grammaticale*, Peeters Publishers, 2014, pp.8-13.

Gries S. Th., *Statistics Data for Linguistics With R*, De Gruyter Mouton

Sébillot, P., 2014. "Le traitement automatique des langues face aux données textuelles volumineuses et potentiellement dégradées : qu'est-ce que cela change ?". Lisette Calderan; Pascale Laurent; Hélène Lowinger; Jacques Millet. Big data : nouvelles partitions de l'information. Actes du séminaire IST INRIA,, octobre 2014, De Boeck, pp.43-60, 2015, Information et stratégie.

Tanguy, C., Fabre, C. 2014. "Évolutions de la linguistique outillée : méfaits et bienfaits du TAL". *L'information grammaticale*, Peeters Publishers, 2014, pp.15-23.

Modalités de contrôle

Contrôle continu : 4 exercices en CC + 1 DM
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. À 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 : Loic Grobol (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*.

Interfaces pour le Web

Enseignant : Loic Grobol (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 et Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, Bâtiment BFC, salle 408
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.

—----------

Conférences professionnelles

Enseignant : Iris Taravella et Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, bâtiment BFC, salle 408
Horaire : (cf. planning en ligne sur plurital.org)

Ce cours permet aux étudiants de découvrir les recherches actuelles du TAL dans le monde industriel. Il est composé d'une série d'interventions de représentants des entreprises travaillant dans le domaine du TAL.

Langages de script

Parcours : TeTraDom
Enseignant : Alexandre Roulois (IRIF, université Paris-Cité, CNRS)
Lieu : INALCO, rue de Lille, salle LO.01
Horaire : au S1 (Mercredi 9h00-11h00, cf planning INALCO)

Après quelques révisions sur l'utilisation de la ligne de commande et sur les notions-clés du langage Python dans une perspective de traitement de la langue, notre parcours visitera quelques bibliothèques logicielles importantes du domaine en vue de structurer et de traiter des données. De nombreux exercices seront proposés tout au long des séances afin de consolider les apprentissages.

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

Espace de cours en ligne : https://github.com/Alex-bzh/python-M2TTD

Conduite de projets de traduction

Parcours : TeTraDom
Enseignant : Sophie Gaumet (S1), Carine Kerne (S2)
Lieu : INALCO, rue de Lille
Horaire : Mardi 9h00-11h00, cf planning INALCO

Ce cours permet de découvrir la conduite de projets de traduction.

- Définition de l'activité de conduite de projet de traduction.

- Etude du cycle de vie d'un projet de traduction, de son déclenchement à sa clôture : préparation/vente de la prestation, organisation de la production, contrôle qualité, gestion des risques, livraison et facturation.

- Sujets d'ouverture : projets d'interprétariat, métiers exercés dans les agences de traduction, survol du marché de la traduction en agence (acteurs, outils, tendances).

Modalités de contrôle : contrôle continu

Langage de scripts : niveau 2

Parcours : Ingénierie Linguistique
Enseignant : Louis Jourdain (ChapsVision)
Lieu : INALCO, rue de Lille
Horaire : Mardi 14h00-16h00, cf planning INALCO

Après la révision de points vus en première année (structures de données, traitement de fichiers...) on renforcera les acquis des étudiants en travaillant sur deux points :

* La préparation des entretiens techniques de recrutement par le traitement chaque semaine d'un problème d'algorithmique

* L'élargissement de la boîte à outils du programmeur en introduisant des librairies et points de syntaxe utiles (programmation asynchrone, décorateurs, pandas, matplotlib et seaborn...). qui seront illustrés par des travaux pratiques.

Les sujets traités peuvent être adaptés aux besoins des étudiants.

Modalités de contrôle : petits devoirs d'algorithmique, certains TPs notés et un projet final

Ingénierie des connaissances (INALCO)

Parcours : Ingénierie Linguistique
Enseignant : Louis Jourdain (ChapsVision)
Lieu : INALCO, rue de Lille
Horaire : Mardi 14h30-16h30, cf planning INALCO

Ce cours est l'équivalent du cours proposé à Nanterre / Sorbonne nouvelle et présente la notion d'ontologie et les langages RDF et SparQL. On présentera également les modes de constitution et différentes utilisations des graphes de connaissance (*Knowledge Graphs*)
L'accent sera particulièrement mis sur les emplois actuels des ontologies et des graphes de connaissances dans l'industrie pour renforcer d'autres systèmes de TAL (systèmes de NER, Retrieval Augmented Generation...) sur des cas d'applications spécifiques ainsi que sur l'impact du déploiement des LLMs sur la constitution des ontologies (challenges et opportunités).

Tout cela sera mis en pratique en langage python.

Si besoin, des rappels d'algorithmique sur les graphes seront faits.

Modalités de contrôle : Devoir sur table, certains TPs notés et projet final

http://pluriTAL.org

Sorbonne Nouvelle

inalco
Institut national des langues et civilisations orientales

Université Paris Nanterre

Contacts

Taravella Iris

- Paris Nanterre
- MoDyCo
- ieshkolt@parisnanterre.fr

Battistelli Delphine

- Paris Nanterre
- MoDyCo
- dbattist@parisnanterre.fr

Valette Mathieu

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

Cédric Gendrot

- Sorbonne Nouvelle / ILPGA
- LPP
- cedric.gendrot@sorbonne-nouvelle.fr