Master
Traitement Automatique des Langues

PluriTAL

Version provisoire - Sept. 2025, peu donner lieu à des modifications

Table des matières

Responsables:

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

Site web de la formation : http://plurital.org

1- Le MASTER TAL : L'informatique et l'IA au service des langues

Le Traitement Automatique des Langues (ou NLP pour Natural Language Processing) est un domaine à l'interface de la linguistique, de l'informatique et de l'Intelligence Artificielle axé sur les méthodes, algorithmes et technologies dédiées aux langues et aux textes. Les débouchés de ce domaine sont soit applicatifs (réalisation d'applications spécialisées en traduction automatique, agents conversationnels – chatbots, veille, filtrage et recherche d'information, extraction de connaissances, etc.) soit scientifiques (outils pour l'analyse linguistique automatique ou assistée, humanités numériques).

Le Master TAL repose sur partenariat entre l'Université Sorbonne Nouvelle (USN), l'Université Paris Nanterre (UPN) et l'institut National des Langues et Civilisations Orientales (Inalco). Il offre une formation fondamentale (modélisation linguistique, programmation et algorithmique, apprentissage automatique…) et appliquée (génie logiciel, chaînes de traitement, constitution et exploitation de corpus et jeux de données…).
Le master TAL répond aux attentes économiques et sociales créées par la nécessité de penser le numérique en termes de diversité des langues et des cultures. Il forme aux métiers du numérique : ingénieur linguiste, ingénieur NLP, data scientist, ingénieur LLM, ingénieur IA, etc.

La formation s'appuie sur les laboratoires : ERTIM (Textes, Informatique, Multilinguisme (Inalco), Lattice (Langues, Textes, Traitements informatiques, Cognition (ENS/PSL, CNRS & Sorbonne Nouvelle), LPP, Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), MoDyCo, (Modèles, Dynamiques, Corpus (CNRS & Paris Nanterre).

L'organisation du Master TAL

La 1re année de Master (M1) consiste en un tronc commun aux 3 établissement où sont enseignés les fondamentaux en linguistique pour le TAL, mathématiques, algorithmiques et programmation. L'ensemble des cours est obligatoire et correspond à un volume d'environ 540 heures réparties sur 24 semaines.

La 2e année (M2) permet de choisir entre 4 parcours :

- **Recherche et développement (R&D) :** parcours commun aux 3 établissements, il permet de constituer un programme pédagogique « à la carte » destiné à préparer un projet de recherche doctorale. La formation s'étend sur un semestre (12 semaines, environ 240 heures) suivi d'un stage donnant lieu à un mémoire de recherche.

- **Ingénierie multilingue (IM) :** formation aux métiers de l'ingénierie NLP et IA, écrit et oral. Le parcours se déroule sur un semestre et demi (340 heures) et est suivi d'un stage en entreprise ou en laboratoire donnant lieu à un mémoire de recherche. Sa spécificité est de se focaliser sur les problématiques du multilinguisme, de la variation linguistique et du traitement automatique des langues du domaine Inalco qui disposent parfois de peu de données disponibles et d'applications dédiées. Cette 2e année met donc l'accent sur les

systèmes d'écriture, d'encodage, les caractéristiques linguistiques de ces langues dites « peu dotées ».

- **Technologies de la Traduction et Traitement des Données Multilingues (TETRADOM) :** formation équivalente au parcours « Ingénierie Multilingue » mais spécialisée en traduction automatique pour les langues peu dotées. Attention, ce parcours ne sera pas ouvert à l'inscription en 2025.

- **Alternance :** [...]

Le master TAL est multi-site : en M1 et en M2 R&D les cours ont lieu au Pôle des langues et civilisations (Inalco, 13e), sur le Campus Nation (Sorbonne Nouvelle, 12e) et sur le Campus Nanterre (Paris Nanterre, 92).
Les cours du parcours M2 Alternance ont lieu sur le campus Nanterre.
Les cours des parcours M2 IM et TETRADOM à la Maison de la recherche de l'Inalco (7e).

Public attendu

Le Master s'adresse :

- à des personnes ayant une formation initiale en **sciences humaines et sociales** (langues, linguistique ou autres) n'ayant pas obligatoirement de compétences informatiques mais désireux d'acquérir une formation technique approfondie favorisant leur intégration dans les métiers du numérique. Une semaine de formation intensive en pré-rentrée est programmée pour la mise à niveau des candidats n'ayant pas suivi de formation informatique au préalable,
- à des personnes ayant une formation en mathématiques ou informatique qui souhaitent développer une compétence spécifique en traitement des données textuelles (écrit et oral).

Conditions d'admission

Les candidats sont admis sur dossier. En M1, les étudiants de l'inalco doivent justifier d'un niveau suffisant dans une langue enseignée à l'Inalco et faire la preuve de compétences en linguistique ou en informatique. Les étudiants étrangers doivent en outre justifier d'une connaissance suffisante du français.

Les candidats de nationalité française ou diplômés d'une université française doivent s'inscrire via la plateforme https://monmaster.gouv.fr/
Les ressortissants d'autres pays doivent s'inscrire via Campus France https://www.campusfrance.org ou e-candidat pour l'Inalco https://admissions.inalco.fr/

Débouchés professionnels

Les métiers des industries de la langue sont multiples. Des possibilités s'offrent dans les

entreprises du numérique ou ayant des secteurs spécialisés dans le développement d'outils TAL (NLP) et IA pour la conception et la maintenance de logiciels. Ces mêmes entreprises et bien d'autres (petites et moyennes, *startups* organismes nationaux et internationaux), pour les services qu'elles proposent ou pour leurs besoins propres, peuvent également faire appel à des spécialistes dans les domaines de la production, de l'organisation et de la gestion de l'information et de la connaissance. Les entreprises qui recrutent nos alumni sont par exemple Orange, Thales, Dassault System, Google, EDF, Ubisoft, Systran, Acolad, Viasema, Mondeca, SématiWeb, Aday, etc. Les métiers relèvent essentiellement de l'ingénierie informatique : ingénieur linguiste, ingénieur NLP, développeur NLP, ingénieur LLMs, ingénieur IA, data scientists et ingénieur text mining,
etc.

2- Présentation rapide du Master

Master 1

L'emploi du temps pour les Master 1 est figé avec assez peu de choix hormis pour les langues et la linguistique. Vous trouverez ci-dessous un lien avec les cours fixes :

https://docs.google.com/spreadsheets/d/1NPydiKWIS9-UA_f_jgbY22SFSniofWMI/edit?usp=sharing&ouid=109368815789528169658&rtpof=true&sd=true

MASTER TRAITEMENT AUTOMATIQUE DES LANGUES
M1

INTITULÉS	ECTS	H	T	Enseignant
MASTER 1RE ANNÉE	**60**			
Semestre 7	**30**			
UE 1 : LANGUES VIVANTES (1 choix)	**3**			
Langue de l'Inalco, 1 (cours de la langue niveau M1, ou initiation à une langue via MOOC)	3	24	TD	
Paper readings, 1 (Reading of papers in English)	3	24	TD	Kaeta Gabor (Inalco)

~~NLP in English, 1 (NLP tutorials in English)~~	~~3~~	~~24~~	~~TD~~	(fermé en 2025-2026)
Langue (LANSAD USN ou UPN)	3	24	TD	
UE 2 : LINGUISTIQUE COMPUTATIONNELLE (3 EC à choisir)	**9**			
Linguistique pour le TAL	3	24	CM TD	Kata Gabor (Inalco)
Corpus multilingues et traduction	3	24	CM TD	Maud Bénard (Inalco)
Corpus arborés et parsing	3	24		Aleksandra Miletic (UPN) + Ioana Madalina Silai (UPN)
Analyse linguistique des modèles de langues	3	24	CM TD	Andrea Briglia (USN)
Cours de Linguistique au choix dans les 3 établissements	3	24	CM TD	
UE3 : FONDAMENTAUX (3 EC)	9			
Mathématiques pour le TAL, 1 (algèbre linéaire, statistiques et méthodes d'évaluation, Damien Nouvel, Inalco)	3	24	CM TD	Damien Nouvel (Inalco)
Algorithmique et programmation, 1 (Introduction à l'algorithmique, la gestion de données, la programmation)	3	24	CM TD	Eric Jordan (Inalco) + Iris Taravella (UPN) (2 groupes)
Traitement symbolique des données (automates)	3	24	CM TD	Anna Colli (Inalco) + Delphine Battistelli (UPN) (2 groupes)
UE4 : INGENIERIE (3 EC)	9			
Introduction au TAL (histoire et épistémologie du TAL, segmentation, concepts et objets : tokens, lemmes, n-grams	3	24	UPN	Iris Taravella (UPN)
Construction de données, 1 (outils pour la constitution de corpus, formats de données, structurations des données, gestion informatique du multilinguisme)	3	24	CM TD	Ilaine Wang (Inalco)
Projet de programmation encadré, 1 (mise en place d'une chaîne de traitement TAL de la conception au développement)	3	36	CM TD	Pierre Magistry (Inalco) + Yoann Dupont (USN)
Semestre 8	**30**			
UE 1 : LANGUES VIVANTES (1 EC)	**3**			
Langue de l'Inalco, 2 (cours de la langue niveau M1, ou initiation à une langue via MOOC)	3	24	TD	
Paper readings, 2	3	24	TD	UPN
~~NLP in English, 2 (course of NLP in English)~~	~~3~~	~~24~~	~~TD~~	(fermé en 2025-2026)
UE 2 : LINGUISTIQUE COMPUTATIONNELLE (2 EC)	**6**			
Analyse statistique des données textuelles	3	24	CM TD	Mathieu Valette (Inalco)
Sémantique distributionnelle	3	24	CM TD	Pascal Amsili (USN)
UE3 : FONDAMENTAUX (3 EC)	**9**			
Mathématiques pour le TAL, 2 (modèles mathématiques pour la représentation, algèbre linéaire, introduction aux modèles de langues	3	24	CM TD	Loic Grobol (UPN)
Algorithmique et programmation, 2 (arbres, optimisation, complexité)	3	24	CM TD	Intervenant extérieur CC Inalco

Traitement statistique des données (méthodes statistiques et probabilistes, apprentissage automatique : statistique bayésiennes, classifieurs linéaires, réseaux de neurones)	3	24	CM TD	Damien Nouvel (Inalco) + EC USN (2 groupes)
UE 4 : INGENIERIE (4 EC)	**12**			
Applications (Éthique du TAL & présentation des principales applications en TAL – exemples : fouille de textes, fouilles d'opinions, extraction d'information, reconnaissances d'entités nommées, IA générative traitement de la parole)	3	24	CM TD	Intervenants extérieurs CC Inalco (9 heures assurées par l'Inalco)
Construction de données, 2 (enrichissement de corpus, base de données SQL, XML)	3	24	CM TD	Iris Taravella (UPN)
Projet de programmation encadré, 2 (mise en place d'une chaîne de traitement TAL de la conception au développement)	3	36	CM TD	Pierre Magistry (Inalco) + Yoann Dupont (USN)
Parole / oralités	3	24		Cédric Gendrot (USN)

Séminaires linguistique USN : https://www.sorbonne-nouvelle.fr/medias/fichier/brochure-master-sdl-15-juillet-et-apres-2025-2026_1753448321598.pdf

Séminaires linguistique INALCO :
https://www.inalco.fr/sites/default/files/2025-07/Brochure_master_SDL_2025-2026.pdf

Séminaires linguistique UPN : https://master-fldl.parisnanterre.fr/

Master 2 R&D

INTITULÉS	ECTS	Volume horaire semestriel	Typologie envisagée : TD/CM /CMTD	Enseignant
MASTER 2ème ANNÉE parcours R&D	**60**			
Semestre 9	**30**			
UE 1 : Ingénierie (7 choix)	**3**			
Réseaux de Neurones pour l'oral	3	24	CMTD	Audibert & Gendrot, USN
Analyse sémantique automatique	3	24	CMTD	Amsili, USN
Modélisation des langues	3	24	CMTD	Kahane, UPN
La subjectivité dans le langage : applications en TAL	3	24	CMTD	Battistelli, UPN
Apprentissage supervisé : méthodes, modèles, exemples	3	24	CMTD	Taravella, UPN
Ingénierie des connaissances 1 : des	3	24	CMTD	Taravella &

réseaux sémantiques vers les ontologies				Battistelli, UPN
Linguistique outillée et traitements statistiques	3	24	CMTD	Battistelli, UPN
Ingénierie des connaissances 2 : ontologies et technologies du Web sémantique	3	24	CMTD	Taravella, UPN
Apprentissage automatique et réseaux de neurones 1	3	24	CMTD	Grobol, UPN
Interfaces web pour le TAL	3	24	CMTD	Grobol, UPN
Apprentissage automatique et réseaux de neurones 2	3	24	CMTD	Grobol, UPN
Langues peu dotées : typologie quantitative et traitement automatique	3	24	CMTD	Grobol, Milletic, Kahane, Faghiri, UPN
Enjeux majeurs et avancées récentes du TAL	3	24	CMTD	Taravella & Battistelli, UPN
TALA526C - Extraction - fouille de textes, extraction d'information	3	24	CMTD	Nouvel, INALCO
TALA536B - Apprentissage, réseaux de neurones profonds, modèles de langues	3	24	CMTD	Magistry, INALCO
TALA516B – TAL pour les langues peu dotées	3	24	CMTD	Wang, INALCO
TALA526A - Techniques Web, programmation Web, réseaux et applications mobiles	3	24	CMTD	Jourdain, INALCO
TALA526B – Génération - traduction automatique	3	24	CMTD	Fily, INALCO
TALA516C - Corpus, langues, cultures	3	24	CMTD	Valette, INALCO
TALA516A - Linguistique pour le TAL multilingue : sémantique de corpus	3	24	CMTD	Valette, INALCO
TALA536C - Acquisition, modélisation et représentation des connaissances	3	24	CMTD	Darenne, INALCO
TALA536A - Programmation et	3	36	CMTD	Gabor, INALCO

programmation objets				

UE 2 : Linguistique (3 EC à choisir)	3	24	CMTD	
A choisir parmi la liste des cours du Master de SDL de l'USN ou INALCO ou UPN	3	24	Séminaire	
A choisir parmi la liste des cours du Master de SDL de l'USN ou INALCO ou UPN	3	24	séminaire	
A choisir parmi la liste des cours du Master de SDL de l'USN ou INALCO ou UPN	3	24	séminaire	

Séminaires linguistique USN : https://www.sorbonne-nouvelle.fr/medias/fichier/brochure-master-sdl-15-juillet-et-apres-2025-2026_1753448321598.pdf

Séminaires linguistique INALCO :
https://www.inalco.fr/sites/default/files/2025-07/Brochure_master_SDL_2025-2026.pdf

Séminaires linguistique UPN : https://master-fldl.parisnanterre.fr/

Semestre 10	**30**			
Stage en entreprise ou en laboratoire de recherche	**30**			

Master 2 Ingénierie Multilingue

INTITULÉS	ECTS	Volume horaire semestriel	Typologie envisagée : TD/CM /CMTD	Enseignant
MASTER 2ème ANNÉE parcours Ingénierie Multilingue	**60**	336		
Semestre 9	**30**			
UE 1 : Linguistique	**9**			
TALA510a – Linguistique pour le TAL multilingue, 1 (sémantique de corpus)	3	24	CMTD	Valette, INALCO
TALA510a – Écritures multilingues (encodage, langues peu dotées)	3	24	CMTD	Wang, INALCO

TALA510c – Corpus, langues, cultures	3	24	CMTD	Valette, INALCO
UE 2 : Ingénierie	**11**			
TALA526A - Techniques Web, programmation Web, réseaux et applications mobiles	3	24	CMTD	Jourdain, INALCO
TALA526B – Génération, 1 - traduction automatique	4	24	CMTD	Fily, INALCO
TALA526C – Extraction, 1 - fouille de textes, extraction d'information	4	24	CMTD	Nouvel, INALCO
UE 3 : Modèles et formalismes	10			
TALA536A - Programmation et programmation objets	4	36	CMTD	Gabor, INALCO
TALA536B - Apprentissage, réseaux de neurones profonds, modèles de langues, 1	3	24	CMTD	Magistry, INALCO
TALA536C - Acquisition, modélisation et représentation des connaissances	3	24	CMTD	Darenne, INALCO

Semestre 10	**30**			
UE 1 : Linguistique	**3**			
TALB510a – Linguistique pour le TAL multilingue, 2 (lexicologie)	3	24	CMTD	Gabor, INALCO
UE 2 : Ingénierie	**9**			
TALB520a –Traitement de la parole	3	24	CMTD	Fedchenko, INALCO
TALB520b – Génération, 2 (système de questions-réponses résumé automatique	3	24	CMTD	Nouvel, INALCO
TALB520C– Extraction, 2 (Fouille de textes, extraction d'information)	3	24	CMTD	Nouvel, INALCO
UE 3 : Modèles et formalismes	3			
TALA536B - Apprentissage, réseaux de neurones profonds, modèles de langues, 2	3	24	CMTD	Fedchenko, INALCO
UE 4 : Stage et Mémoire	**15**			

3- Descriptif des cours du tronc commun (master 1)

3-1. UE LANGUES VIVANTES

Langue de l'Inalco (48 heures)

- Résumé de 5 lignes (cours de la langue niveau M1, ou initiation à une langue via MOOC)
- Bibliographie indicative
- Intervenant(s)

Paper readings (48 heures)

- Résumé de 5 lignes Reading of papers in English
- Bibliographie indicative

Intervenant(s)

NLP in English (48 heures)

Non dispensé cette année

3-2. UE LINGUISTIQUE COMPUTATIONNELLE

Corpus multilingues et traduction (24 heures)

- Résumé de 5 lignes
- Bibliographie indicative
- Intervenant(s) ; Maud Bénard, Inalco

Linguistique pour le TAL (24 heures)

- Résumé de 5 lignes
- Bibliographie indicative
- Intervenant(s) ; Kata Gobor

Corpus arborés et parsing (24h)

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. Les schémas 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, https://arborator.grew.fr/

Universal Dependencies Treebanks, universaldependencies.org

Grew-match, match.grew.fr

Modalités de contrôle

Contrôle continu : La moyenne du cours est composée de 2 DS de 1h chacun, espacés sur le semestre.

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

Enseignantes : Aleksandra Miletić (Paris Nanterre) & Ioana Madalina Silai (Paris Nanterre)

Analyse statistique des données textuelles (24 heures)

Ce cours porte sur les méthodes et heuristiques de l'analyse des données textuelles (ADT) dans la perspective de son école européenne (lexicométrie, textométrie, herméneutique numérique, etc.)

dont les débouchés sont principalement l'analyse du discours et les humanités numériques. Dans une perspective critique et par comparaison avec le TAL, Il s'agit d'opposer à la notion de « données » celle d'« observables » , à la notion de *dataset*, celle de construction de corpus et au principe d'automatisation, celui d'itération humain-machine. Le cours s'appuiera sur un ensemble d'outils d'ADT et la mise en place d'une analyse de corpus.

Bibliographie indicative
LEBART, Ludovic, PINCEMIN, Bénédicte et POUDAT, Céline, 2019. Analyse des données textuelles Québec : Presses de l'Université du Québec. ISBN 978-2-7605-5052-0

Intervenant(s) ; Mathieu Valette, Inalco

Sémantique distributionnelles (Pascal Amsili, USN)

- Partie "fondamentale" : sémantique distributionnelle, un peu d'algèbre linéaire (réduction de dimensionalité, opérations sur les vecteurs…), et plongements lexicaux (de Word2vec à BERT)
- Partie "applicative" (selon temps) : tâches de sémantique computationnelle : résolution de coréférences, détection des inférences naturelles (RTE/NLI)

- Intervenant : Pascal Amsili

3-3. UE FONDAMENTAUX

Programmation et algorithmique 1 et 2

Enseignant : Iris Taravella (Paris Nanterre), Eric Jordan (Inalco), Mathieu Dehouck & Pascal Amsili (Sorbonne nouvelle),
Lieu : Semestre 1 : Paris Nanterre, L210 et L128
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 & Eric Jordan)

Ce cours aborde les notions de base du langage Python 3 : types de données (données numériques, chaînes de caractères, listes, dictionnaires, tuples), boucles, fonctions, modules intégrés, 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 exercices 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%)

Mathématiques pour le TAL

- Résumé de 5 lignes
  - statistiques élémentaires (moyenne, écart-type, corrélation)
  - évaluation des modèles (précision, rappel, etc.)
  - algèbre linéaire (vecteurs, matrices)
  - combinatoire et probabilités
- Bibliographie indicative
- Intervenant(s) : Damien Nouvel, Inalco

modèles mathématiques pour la représentation, algèbre linéaire, introduction aux modèles de langues.

Traitement symbolique des données

Lieu : USN, salles 308 et 418, Batiment L (Ricoeur)

Horaire : jeudi 10h30-12h30 au 1er semestre
Intervenantes : Delphine Battistelli (USN), Anna Colli (Inalco)

Le cours propose une initiation aux questions méthodologiques posées par l'approche dite symbolique en TAL. Deux cas d'étude seront explorés, posant chacun un challenge différent du

point de vue de la modélisation mais aussi du point de vue des enjeux applicatifs sous-jacents : le premier concernera les unités adverbiales temporelles ; le second des unités impliquant le verbe 'dire'. Seront mis à contribution des mathématiques de manière directe ou indirecte, notamment via une modélisation sous la forme d'automates (utilisation du logiciel Unitex).

Bibliographie indicative

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

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 sera calculée à partir de 2 DST.
Contrôle dérogatoire et rattrapage : Examen sur table de 2h.

Traitement statistique des données

Yoann Dupont (USN) + Damien Nouvel (Inalco) (intitulé à modifier ? Traitement probabiliste des données ?)

- principes d'apprentissage automatique
- inférence bayésienne
- arbres de décision
- entropie et régression logistique
- réseaux de neurones simples (perceptron)

3-4. INGÉNIERIE

Introduction au TAL

Enseignante : Iris Taravella (UPN)

Lieu : Université Paris Nanterre, salle L318
Horaire : Semestre 1 : Vendredi 10h30-12h30

Résumé :
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, de définir les bases et notions de cette discipline ainsi que de montrer les applications concrètes.

Modalité de contrôle : QCM de 1 heure

Bibliographie

Cori, M. (2020). « Le traitement automatique des langues en question, Des machines qui comprennent le français ? » Cassini, 248 pages.
F-R Chaumartin (2020). « Le traitement automatique des langues. Comprendre les textes grâce à l'intelligence artificielle ». InfoPro, Dunod.
Cori, M. (2008). « Des méthodes de traitement automatique aux linguistiques fondées sur les corpus », Langages n° 171, 2008, pp. 95-110.
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.

Construction de données

Lieu : Semestre 1 : Inalco, PLC, 5.08
Semestre 2 : Université Paris Nanterre, salle à préciser

Horaire : Semestre 1 : mardi 13h30-15h30
Semestre 2 :

Enseignants : S1 : Johanna Cordova, Ilaine Wang, (Inalco) et S2 : Iris Taravella (UPN)

S1 : Deux parties :
- Outils pour la constitution de corpus, formats de données, structurations des données
  - gestion informatique du multilinguisme
  -
S2 : Deux parties :
- Enrichissement de corpus : Cette partie sera consacrée 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.
- Bases de données relationnelles : Cette partie 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. Des exercices sont systématiquement associés à la présentation des concepts. Les étudiants travailleront ensemble par petits groupes de 2-4 personnes sur la création d'une base de donnée. A la fin, ils présenteront à l'oral les résultats du travail et rendront un petit rapport écrit.
Le cours ne suppose pas de connaissances informatiques préalables.

Modalités de contrôle

Contrôle continu : La moyenne de l'année est composée d'un projet portant sur une des deux parties et d'un examen sur table de 2h.

Projet de programmation encadré (2 semestres)

Mise en place d'une chaîne de traitement TAL de la conception au développement, Pierre Magistry, Inalco & Yoann Dupont, USN & Loic Grobol UPN)
« bash git et du python orienté génie logiciel/comment on s'organise en groupe et pour une base de code saine »

Applications (2e semestre)

Ce cours sera l'occasion d'interventions ponctuelles de spécialistes du TAL qui présenteront quelques applications en TAL (fouille de textes, fouilles d'opinions, extraction d'information, reconnaissances d'entités nommées, IA générative traitement de la parole, etc.).

Parole / oralités

Cédric Gendrot, USN
Lieu : Semestre 2 : USN

Mise en place d'un système de synthèse de parole. Rappels des principes de phonétique et de traitement du signal. Historique de la synthèse. Utilisation de synthèse neuronale et de clonage. Langages utilisés : python et praat.
Validation : un devoir sur table à mi-semestre et un devoir maison individuel

4- Descriptif des cours des parcours (master 2)

Corpus, langues, cultures

Lieu : Maison de la recherche de l'Inalco, 2 rue de Lille, salle L0.01
Horaire : lundi 13h-15h00 au 1er semestre
Intervenant : Mathieu Valette (Inalco)

Résumé. Ce cours s'organise comme un séminaire de recherche et a pour objet les relations entre les langues, les cultures et les nouvelles problématiques du TAL : IA générative, grands modèles de langues (LLMs). Nous questionnons les points de contact et les divergences des différentes disciplines en jeu (linguistique, IA), dans une perspective théorique et historique, de manière à évaluer en quoi elles peuvent se féconder mutuellement. Nous évaluerons la place octroyée à la culture dans les travaux industriels et académiques portant sur la constitution de LLMs, où les cultures sont appréhendées sous le seul angle axiologique (valeurs morales). Nous nous interrogerons dans ce cadre sur le statut donné par les acteurs de l'IA générative aux objets culturels (narratifs, filmiques, picturaux, etc.), dont la fonction est d'être choisis, contextualisés, interprétés, et transmis pour construire une culture ou permettre le dialogue entre les cultures. Nous ouvrirons le débat sur les risques encourus par les sociétés humaines dans ce contexte (fragmentation culturelle) et les solutions possibles. Ce cours peut être vu comme une introduction épistémologique au cours *Linguistique pour le TAL multilingue* (TALA516A).

Evaluation : exposés, rapport

Bibliographie indicative
Léon J. (2015). Histoire de l'automatisation des sciences du langage, Paris : ENS Éditions.
Offert F., Dhaliwal, R.S. (2025). The method of Critcal AI Studies, A Propaedeutic, arXiv:2411.18833v3 [cs.CY] 23 Mar 2025.
Valette, M. (2025). Culture et acculturation des grands modèles de langue », EvalLLM@ CORIA-TALN-RJCRI-RECITAL 2025, 30 juin 2025, Marseille, France, 68-76.

Linguistique pour le TAL multilingue : sémantique de corpus

Lieu : Maison de la recherche de l'Inalco, 2 rue de Lille, salle L0.01
Horaire : lundi 15h-17h00 au 1er semestre
Intervenant : Mathieu Valette (Inalco)

Résumé. Ce cours s'organise en TD et vise à mettre en applications une sémantique de corpus pour les applications. Cette année, nous nous focaliserons sur le concept de « récit civilisationnel » élaborés en sciences politiques qui n'est que marginalement liées à des problématiques narratologiques. Nous étudierons les méthodes permettant de modéliser ces récits et de les identifier dans des corpus de textes multilingues.

Evaluation : rapport

Bibliographie indicative
Rastier, F. (2011). La mesure et le grain. Sémantique de corpus. Paris : Champion, Collection Lettres numériques.

Charon, P. (2024.) Lire la désinformation comme un récit sériel : pour une approche littéraire des manipulations de l'information, Le rubicon,
Elfes, J. (2024) Mapping News Narratives Using LLMs and Narrative-Structured Text Embeddings, preprint https://doi.org/10.48550/arXiv.2409.06540

Modélisation des langues

Lieu : Paris Nanterre, salle BFC 408

Horaire : mercredi 13h30-16h30

Enseignant : Sylvain Kahane (Paris Nanterre

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 : unité linguistique, lexème, collocation, relation prédicat-argument vs dépendance syntaxique, etc.

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, 2022, *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.

Apprentissage supervisé : méthodes, modèles, exemples

Lieu : Université Paris Nanterre, salle BFC205
Horaire : lundi 13h30-16h30
Enseignant : Iris Taravella

Résumé. Ce cours a pour objectif d'initier les étudiants au processus de l'apprentissage supervisé. Le cours abordera les thématiques suivantes : modèles de l'apprentissage de surface, la représentation de données, l'apprentissage profond, LLM, etc. Suite à cette introduction théorique, les étudiants travailleront ensemble par petits groupes de 2-4 personnes sur un article scientifique choisi par l'enseignant / les étudiants . A la fin, ils présenteront à l'oral une recherche présenté dans l'article ainsi que les résultats du test de la méthodologie proposée dans l'article sur les données nouvelles et comparables. Ils rendront un petit mémoire écrit.

Ingénierie des connaissances 1 : des réseaux sémantiques vers les ontologies

Enseignant : Delphine Battistelli, Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre, bâtiment BFC, salle 205
Horaire : lundi 13h30-14h30

L'Ingénierie des Connaissances 1 (IC1) 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 : contrôle continue la moyenne des exercices + un devoir sur table de 3h

Espace cours en ligne : oui.

Ingénierie des connaissances 2 : ontologies et technologies du Web sémantique

Enseignant : Iris Taravella (Paris Nanterre)
Lieu : Paris Nanterre, salle BFC205
Horaire : lundi 13h30-16h30

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 IC1 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.

Enjeux majeurs et avancées récentes du TAL

Enseignant : Iris Taravella et Delphine Battistelli (Paris Nanterre)

Lieu : Paris Nanterre, Bâtiment BFC, salle 205
Horaire : dates à préciser

Le cours vise à faire découvrir aux étudiants les recherches actuelles dans les domaines du TAL à travers une série de présentations d'enseignants et de chercheurs du domaine.

Conférences professionnelles

Enseignant : Iris Taravella et Delphine Battistelli (Paris Nanterre)
Lieu : Paris Nanterre, bâtiment BFC, salle 205
Horaire : mardi 9h30-12h30

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

La subjectivité dans le langage : applications en TAL

Enseignant : Delphine Battistelli (UPN)

Lieu : Paris Nanterre, Bâtiment BFC, salle 205

Horaire : lundi, 10h-13h

Ce cours présente des méthodes, modèles et applications propres à appréhender un niveau d'analyse et d'annotation de la composante dite subjective du langage, largement explorée dans divers domaines du TAL (analyse d'opinions, d'émotions, de sentiments, fact-cheking, hate speech detection, ...), mais où la notion de subjectivité en elle-même est prise dans d'importantes difficultés définitoires (recouvrant tour à tour des concepts linguistiques telles que modalité, évidentialité, évaluation, ...). L'enjeu du cours consistera entre autres à démontrer l'intérêt de modèles théoriques linguistiques robustes pour des approches performantes dans ce domaine.

Bibliographie indicative

Aline Étienne, Delphine Battistelli, and Gwénolé Lecorvé. 2024. Emotion Identification for French in Written Texts: Considering Modes of Emotion Expression as a Step Towards Text Complexity Analysis. In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 168–185, Bangkok, Thailand. Association for Computational Linguistics.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and

Viviana Patti. 2022. APPReddit: a Corpus of Reddit Posts Annotated for Appraisal. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 3809–3818, Marseille, France. European Language Resources Association.

Modalités de contrôle

Contrôle continu : 1 DM + 1 DST + 1 exposé en groupe .
Contrôle dérogatoire et rattrapage : Un dossier de projet.

Langues peu dotées : typologie quantitative et TAL

Intervenants : Pegah Faghiri (Interdisciplinary Laboratory of Numerical Sciences (LISN)), Sylvain Kahane (UPN), Loïc Grobol (UPN), Aleksandra Miletić (UPN)
Lieu : UPN

L'existence de corpus annotés linguistiquement permet de dépasser les classifications binaires de la typologie traditionnelle en adoptant une approche quantitative qui considère non seulement l'existence mais aussi la fréquence de phénomènes linguistiques dans une langue donnée. Or, les corpus annotés de taille suffisante ne sont disponibles que pour une centaine de langues sur les 7000 langues recensées dans le monde. De surcroît, les outils qui sont à la pointe de l'état de l'art en TAL connaissent souvent d'importantes pertes de performances quand ils sont portés à des langues peu dotées. Dans un premier temps, le cours propose une introduction à la typologie quantitative illustrant l'intérêt critique de ce type d'approches pour la linguistique. La deuxième partie du cours propose un tour d'horizon des méthodes du TAL aptes à faciliter la création de corpus pour les langues moins dotées. Ces méthodes sont mises en pratique sur un échantillon de corpus écologique.

TAL pour les langues peu dotées (= « Ecritures multilingues »)

Enseignants : Johanna Cordova, Ilaine Wang
Lieu : Inalco, 2 rue de Lille (Paris 7e)
Horaire : mardi, 9h-11h

Dans ce cours, nous discuterons de ce qu'on appelle les langues « peu dotées » en TAL, et nous verrons comment les doter en ressources (corpus oral et écrit) et en outils (analyse syntaxique, reconnaissance d'entités nommées, speech-to-text et text-to-speech etc.).

Modalités de contrôle : 1 projet de groupe sur une langue peu dotée et une tâche de TAL au choix

---

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

Enseignant : Cédric Gendrot, Nicolas Audibert
Lieu : USN
Horaire : mercredi 17h-19h 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.

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

Apprentissage automatique et réseaux de neurones

Enseignant : L. Grobol (Paris Nanterre)

Lieu : Paris Nanterre, salle BFC205

Horaire : Mercredi 9h30-12h30

Ce cours présente un panorama de l'apprentissage automatique, en donnant des bases théoriques et formelles solides aux algorithmes et aux modèles utilisés. Il se présente en alternance de CM théoriques sur des sujets précis (algorithme de descente de gradient, structure des réseaux de neurones…) peu aisés à apprendre en autonomie et de TP mettant en application ces notions dans des cas concrets. Le cours est également l'occasion de solidifier les acquis en programmation en Python.

Le cours est construit en deux parties de 8 semaines, théoriquement indépendantes, mais il est recommandé de suivre les deux. Bien que des rappels soient fait au fil de l'eau, des notions de bases en mathématiques en général et en algèbre linéaire en particulier aident grandement pour pouvoir appréhender les parties théoriques et les descriptions des algorithmes étudiés.

Pages des cours :

- Partie 1 : https://loicgrobol.github.io/apprentissage-artificiel/
- Partie 2 : https://loicgrobol.github.io/neural-networks/

Évaluations : pour chaque partie, un TP noté individuel et un mini-projet en groupe.

Interfaces web pour le TAL

Enseignant : L. Grobol (Paris Nanterre)

Lieu : Paris Nanterre, salle BFC205

Horaire : Mercredi 13h30-16h30

Ce cours propose un renforcement des acquis en programmation Python et en ingénierie logicielle, orientée vers le développement d'interfaces réseaux et dans une moindre mesure de sites web servant d'interfaces pour des modèles de TAL. Le cours est principalement pratique, consistant en des TP explorant différentes techniques et bibliothèque logicielles ainsi que des outils tels que les débugueurs ou les analyseurs statiques de programmes.

Évaluation: Un TP noté individuel et un mini-projet en groupe.

Titre : Extraction, 1 - fouille de textes, extraction d'information

Enseignant : D. Nouvel (Inalco)

Lieu : rue de Lille (à vérifier sur le planning)

Horaire : Lundi 10h-12h (à vérifier sur le planning)

Ce cours présente différentes méthodes permettant d'extraire des informations depuis un corpus textuel. La première partie du cours portera sur les méthodes lexicales (sélection de documents, topic modeling, extraction de termes, fouille de données, motifs, concordanciers). Nous verrons ensuite les méthodes d'extraction d'informations par annotation (reconnaissance d'entités nommées, extraction d'événements).

TALA536C - Acquisition, modélisation et représentation des connaissances

Enseignant : L. Darenne (Inalco)

Lieu : rue de Lille, LO.01

Horaire : Jeudi 14h30-16h30

Evaluation : Examen sur table + deux TPs notés

La première partie du cours sera consacré à la compréhension de ce qu'est un connaissance, ce qu'est d'acquérir, structurer, modéliser et représenter la connaissance à travers différentes méthodes et langages. Nous aborderons en deuxième partie de semestre les ontologies et les graphes de connaissance, notamment avec des TPs.

TALA526A - Techniques Web, programmation Web, réseaux et applications mobiles

Enseignant : Louis Jourdain