

Master parcours Langue et Informatique

**Mention
Sciences du langage**

Spécialité
Recherche ou Professionnelle



Présentation de la mention

Sciences du langage

Responsable de la mention : Franck NEVEU franck.neveu@sorbonne-universite.fr

Offre de formation

L'offre de formation de la mention « sciences du langage » du master se compose de trois parcours : **Linguistique française et générale**, **Langue française appliquée**, **Langue et informatique**.

- **Parcours Linguistique française et générale (orientation recherche)**

Ce parcours à finalité recherche propose une offre de formation dans laquelle un lien étroit est établi entre l'étude linguistique du domaine français et la linguistique générale, qui est orientée vers la réflexion sur les universaux du langage, vers les concepts universels de la science du langage et vers l'étude comparative et typologique des langues.

Responsable : André THIBAUT andre.thibault@sorbonne-universite.fr

- **Parcours Langue française appliquée (orientation professionnelle)**

Ce parcours est à finalité professionnelle. Il est destiné à former aux métiers de professeur de français langue étrangère et, plus spécifiquement, en fonction des options en master 2, aux métiers touchant à la politique linguistique (postes de coopération éducative, linguistique et culturelle) ou à l'ingénierie en didactique des langues (conception de programmes de formation pour les établissements d'enseignement et les entreprises) et à la diffusion de français de spécialités (tourisme, arts, affaires).

Responsables : Philippe MONNERET philippe.monneret@sorbonne-universite.fr

Inès SFAR ines.sfar@sorbonne-universite.fr

- **Parcours Langue et informatique (orientation recherche ou professionnelle)**

Ce parcours est à finalité indifférenciée (recherche et professionnelle). Il offre une formation pluridisciplinaire et interdisciplinaire proposant des enseignements théoriques et appliqués sur la modélisation informatique et mathématique d'une langue naturelle et la résolution de problèmes d'ingénierie de la Langue.

Responsables : Claude MONTACIÉ : claudemontacie@sorbonne-universite.fr

Laurence DEVILLERS : devil@limsi.fr

Organisation

Chaque semestre comporte cinq UE dont deux sont communes à l'ensemble des parcours de la mention.

- L'UE 1, « **Méthodologie et épistémologie en linguistique** », est une UE fondamentale qui propose un socle de connaissances indispensables à la maîtrise du domaine linguistique : linguistique théorique et appliquée (domaine français), terminologie linguistique, conférences de linguistique.

- L'UE 5, « **Langue vivante** », est une UE proposant un ensemble d'enseignements permettant un approfondissement de la formation dans la pratique et dans l'approche réflexive des langues.
- Les UE 2, 3, et 4 sont des UE spécifiques à chaque parcours.

Présentation du parcours

Langue et Informatique

Responsables de la spécialité : Claude Montacié, claudemontacie@sorbonne-universite.fr

Laurence Devillers, devil@limsi.fr

Enseignants

Devillers Laurence, Professeur

Montacié Claude, Professeur

Eyharabide Victoria, MCF

Lejeune Gaël, MCF

Hayet-Robert Sophie, MCF

Ben Chaabene Nour El Houda, ATER

Petkovic Ljudmila, ATER

Chambat Anais, ATER

Olivier François, ATER

Sbissi Abderrahman Samia, ATER

Enseignants extérieurs

Intervenants professionnels

Boca Laurent, Ingénieur d'études, Informatique et Télécommunications, Groupe EDF

Fontan Christophe, Consultant SI, Direction des Services IT, Groupe EDF

Objectifs

La spécialité Langue et Informatique vise à la formation de professionnels de haut niveau maîtrisant la modélisation et la résolution mathématique, logique et informatique de problèmes qui relèvent de l'Ingénierie de la Langue et de la Communication Homme-Machine pour la Gestion de l'Information. L'objectif visé, grâce à cette synergie de compétences, est d'offrir une offre concurrentielle pour le développement de solutions informatiques pour ces problématiques.

Mise en place avec le concours de professionnels du secteur, elle permet aussi de donner une formation professionnelle de haut niveau dans le domaine en pleine expansion de l'ingénierie linguistique (recherche d'information, extraction et construction des connaissances, traduction assistée par ordinateur, compréhension et génération de textes, reconnaissance, synthèse et compréhension de la parole, système de dialogues). Elle répond à de vrais besoins et ouvre sur de réels débouchés professionnels. Les métiers actuellement identifiables sont : chef de projet d'ingénierie linguistique, linguiste informaticien, informaticien linguiste, Ingénieur technologies vocales, Ingénieur serveur vocal interactif, veilleur stratégique, knowledge manager, ingénieur support linguistique, terminologue d'entreprise, lexicographe, ergolinguiste.

Conditions d'admission

En M1 la formation est ouverte pour 19 étudiants titulaires d'une licence.

Le flux d'entrées de cette spécialité provient de la spécialité informatique de la licence de Sciences du Langage de la Faculté Lettres Sorbonne-Université ou de parcours équivalents. Des étudiants motivés provenant de parcours en Sciences du Langage ou en Informatique pourraient suivre ces enseignements.

L'admission se fera via la plateforme nationale des masters pour les étudiants possédant un diplôme français et par la plateforme Etudes en France pour les autres.

Contenu

La première année est consacrée à des cours théoriques sanctionnés par des examens. L'étudiant doit présenter un premier mémoire, centré sur les enseignements. Le passage en seconde année dépend des résultats obtenus en première année.

La deuxième année comprend un tronc commun sur les fondamentaux de la spécialité au premier semestre (8 semaines) et au second semestre (8 semaines) où la finalité « recherche » se fait par le choix de la soutenance du mémoire et la finalité « professionnelle » par le choix de la soutenance du stage en entreprise.

Modalités de contrôle des connaissances (MCC)

Les modalités de contrôle des connaissances des enseignements de l'UFR de Sociologie et d'Informatique des Sciences Humaines sont celles de la faculté des lettres (cf. site web). Il s'agit en première année d'un contrôle mixte pondérant note de contrôle continu et note d'examen, et en seconde année d'un contrôle continu intégral. Il faut se renseigner pour les enseignements d'autres UFR.

CALENDRIER FACULTAIRE 2025 - 2026

ayant reçu des avis favorables du conseil de la Faculté des Lettres le 13 mai 2025, et de la Commission de la Formation

1er semestre : du 15 septembre 2025 au 16 janvier 2026 inclus / 2nd semestre : du 26 j

Fin d'année universitaire : le 30 septembre 2026 (stages non crédités d'ECTS des cursus Licence, Master.

2025				2026						
SEPTEMBRE	OCTOBRE	NOVEMBRE	DÉCEMBRE	JANVIER	FEVRIER	MARS	AVRIL	MAI	JUIN	JUILLET
1 L	1 M	1 S	1 L	1 J	1 D	1 D	1 M	1 V	1 L	1
2 M	2 J	2 D	2 M	2 V	2 L	2 L	2 J	2 S	2 M	2
3 M	3 V	3 L	3 M	3 S	3 M	3 M	3 V	3 D	3 M	3
4 J	4 S	4 M	4 J	4 D	4 M	4 M	4 S	4 L	4 J	4
5 V	5 D	5 M	5 V	5 L	5 J	5 J	5 D	5 M	5 V	5
6 S	6 L	6 J	6 S	6 M	6 V	6 V	6 L	6 M	6 S	6
7 D	7 M	7 V	7 D	7 M	7 S	7 S	7 M	7 J	7 D	7
8 L	8 M	8 S	8 L	8 J	8 D	8 D	8 M	8 V	8 L	8
9 M	9 J	9 D	9 M	9 V	9 L	9 L	9 J	9 S	9 M	9
10 M	10 V	10 L	10 M	10 S	10 M	10 M	10 V	10 D	10 M	10
11 J	11 S	11 M	11 J	11 D	11 M	11 M	11 S	11 L	11 J	11
12 V	12 D	12 M	12 V	12 L	12 J	12 J	12 D	12 M	12 V	12
13 S	13 L	13 J	13 S	13 M	13 V	13 V	13 L	13 M	13 S	13
14 D	14 M	14 V	14 D	14 M	14 S	14 S	14 M	14 J	14 D	14
15 L	15 M	15 S	15 L	15 J	15 D	15 D	15 M	15 V	15 L	15
16 M	16 J	16 D	16 M	16 V	16 L	16 L	16 J	16 S	16 M	16
17 M	17 V	17 L	17 M	17 S	17 M	17 M	17 V	17 D	17 M	17
18 J	18 S	18 M	18 J	18 D	18 M	18 M	18 S	18 L	18 J	18
19 V	19 D	19 M	19 V	19 L	19 J	19 J	19 D	19 M	19 V	19
20 S	20 L	20 J	20 S	20 M	20 V	20 V	20 L	20 M	20 S	20
21 D	21 M	21 V	21 D	21 M	21 S	21 S	21 M	21 J	21 D	21
22 L	22 M	22 S	22 L	22 J	22 D	22 D	22 M	22 V	22 L	22
23 M	23 J	23 D	23 M	23 V	23 L	23 L	23 J	23 S	23 M	23
24 J	24 V	24 L	24 M	24 S	24 M	24 M	24 V	24 D	24 M	24
25 S	25 S	25 M	25 J	25 D	25 M	25 M	25 S	25 L	25 J	25

Organisation de la formation

Master 1 - Semestre 1

	Enseignements	CM & TD Par semaine	Modalités de contrôle des connaissances	ETCS
UE 1	Méthodologie et épistémologie en linguistique			7
	Linguistique théorique et appliquée domaine français	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
	Terminologie linguistique	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
	Conférences de l'équipe	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
	Mise à niveau en programmation	16 TD	Contrôle Continu	
UE 2	Informatique fondamentale			7
	Epistémologie des Sciences Humaines et de l'Informatique	1h CM 1,5 h TD	Contrôle Mixte	
	Paradigme de la programmation objet	1h CM 1,5 h TD	Contrôle Mixte	
	Méthodologie de la recherche en langue et Informatique	1h CM 1,5 h TD	Contrôle Mixte	
UE 3	Ingénierie des connaissances			5
	Représentation et modélisation des connaissances	1h CM 1,5 h TD	Contrôle Mixte	
	Représentation et traitement des documents électroniques	1h CM 1,5 h TD	Contrôle Mixte	
UE 4	Linguistique computationnelle			8
	Modèles de linguistique computationnelle	1h CM 1,5 h TD	Contrôle Mixte	
	Grammaires formelles	1h CM 1,5 h TD	Contrôle Mixte	
	Sémantique computationnelle	1h CM 1,5 h TD	Contrôle Mixte	
UE 5	Langues vivantes SIAL (langue au choix)	Voir avec le SIAL	Contrôle Continu	3
UE99	UE Enseignement sans crédit			

Master 1 - Semestre 2

	Enseignements	CM & TD Par semaine	Modalités de contrôle des connaissances	ETCS
UE 1	Méthodologie et épistémologie en linguistique			7
	Linguistique théorique appliquée domaine français	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
	Modélisation des productions langagières orales	1h CM 1,5 h TD	Contrôle Mixte	
	Terminologie Linguistique	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
UE 2	Informatique avancée pour les Sciences Humaines			6
	Programmation générique et conception objet	1h CM 1,5 h TD	Contrôle Mixte	
	Internet et bases de données	1h CM 1,5 h TD	Contrôle Mixte	
	Web sémantique et big-data	1h CM 1,5 h TD	Contrôle Mixte	
UE 3	Traitement automatique du langage I			5
	Programmation pour le TAL	1h CM 1,5 h TD	Contrôle Mixte	
	Introduction au traitement automatique des langues TAL	1h CM 1,5 h TD	Contrôle Mixte	
UE 4	Mémoire			9
	Soutenance du mémoire			
UE 5	Langues vivantes SIAL (langue au choix) Voir avec le SIAL		Contrôle Continu	3
UE99	UE Enseignement sans crédit			

Master 2 - Semestre 3

	Enseignements	CM Par semaine	Modalités de contrôle des connaissances	ETCS
UE 1	Méthodologie et épistémologie			7
	Linguistique théorique et appliquée	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
	Terminologie linguistique	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
	Conférences de l'équipe	Voir l'UFR de Langue Française	Voir l'UFR de Langue Française	
UE 2	Informatique, Langue			8
	Production et perception de la parole	2.5 h	Contrôle Continu	
	Dialogues et serveurs bots	2.5 h	Contrôle Continu	
	Indexation sémantique et recherche d'information	2.5 h	Contrôle Continu	
	Rationalité dans les interactions homme-machine	2.5 h	Contrôle Continu	
UE 3	Traitement automatique			5
	Paradigmes d'évaluation du traitement automatique des langues TAL	2.5 h	Contrôle Continu	
	Traduction automatique	2.5 h	Contrôle Continu	
UE 4	Société numérique			7
	Management de projets	2.5 h	Contrôle Continu	
	Génie Logiciel	2.5 h	Contrôle Continu	
	Conférence de Professionnels		Contrôle Continu	
UE 5	Langues vivantes – SIAL au choix		Contrôle Continu	3
UE 99	UE Enseignement sans crédit			

Master 2 - Semestre 4

	Enseignements	CM Par semaine	Modalités de contrôle des connaissances	ETCS
UE 1	Méthodologie et épistémologie			4
	Analyse de graphes et bases de données NoSQL	2.5 h	Contrôle Continu	
	Structures de connaissances et raisonnement	2.5 h	Contrôle Continu	
UE 2	Paralinguistique			6
	Reconnaissance et compréhension de la parole	2.5 h	Contrôle Continu	
	Synthèse de la parole et caractéristiques individuelles	2.5 h	Contrôle Continu	
UE 3	Sémantiques, connaissances et corpus			6
	Représentations vectorielles et similarités	2.5 h	Contrôle Continu	
	Annotation collaborative de corpus	2.5 h	Contrôle Continu	
UE 4	Finalité recherche ou professionnelle			11
	Soutenance du mémoire de recherche			
	Soutenance du stage en entreprise			
UE 5	Langues vivantes SIAL au choix (note de l'UE 5 du semestre 3)		Contrôle Continu	3
UE 99	UE Enseignement sans crédit			

Descriptifs des enseignements

Master 1 - Semestre 1

UE 1 : Méthodologie et épistémologie (7 ects)

M1LF13FA Linguistique Théorique appliquée

Évaluation :
cf Langue Française

Volume horaire :
2h par semaine

Descriptif : Cf UFR de Langue Française

Bibliographie : Cf UFR de Langue Française

M1LF15FA Terminologie linguistique

Évaluation :
cf Langue Française

Volume horaire :
1h par semaine

Descriptif : Cf UFR de Langue Française

Bibliographie : Cf UFR de Langue Française

M1SOL012 Conférences de linguistique

Volume horaire :
26h sur le semestre

Descriptif :

Les conférences du STIH (Sens, Texte, Informatique, Histoire), EA 4509, qui ont lieu d'octobre à décembre, portent sur des sujets variés de linguistique française ; elles sont assurées par des professeurs et des professeures invités, et précédées de séances préparatoires qui permettent d'en faciliter la compréhension. 6 conférences sont prévues. Le programme sera donné à la rentrée.

M1SOL013 Mise à niveau en programmation

Évaluation :
CCI
Exercices pratiques réalisés en séance : 70%
QCM de validation à la fin de chaque partie : 30%

Volume horaire :
16h TD

Descriptif : Ce module de mise à niveau a pour objectif de consolider, de manière intensive, les compétences en programmation Python nécessaires à l'intégration en Master 1. Il se structure

en trois parties :

- (I) La reprise des fondamentaux du langage : rappels sur types de données usuels, opérations et opérateurs, structures conditionnelles et boucles, structures de données, fonctions. Pour les étudiants expérimentés en programmation, des approfondissements avancés sur ces concepts seront au programme.

- (II) La pratique des statistiques (notamment textuelles), revue des méthodes et librairies les plus usuelles et des formats associés (csv, Json et Xml). Utilisation sur des corpus multilingues, avec visualisation. (histogrammes et loi de Zipf notamment) et réalisation d'une

application pratique de reconnaissance de langue. Pour les étudiants de niveau avancé, un complément sur la vectorisation et les bases de machine learning sera enseigné.

- (III) La réalisation de projets aboutis impliquant la factorisation de code, l'organisation de scripts structurés et la création de fonctions réutilisables. Pour les étudiants de niveau avancé, une initiation au versionnage de code (Github) et à la publication de librairies (via pip) sera proposée.

À la fin de chaque partie, un QCM et des scripts Python à rendre permettront de valider les compétences acquises et d'identifier les points à renforcer. Ce cours vise ainsi à donner

l'aisance en programmation dans une perspective orientée vers le traitement de données et les applications fondamentales de l'intelligence artificielle.

Bibliographie :

Cordeau, Bob & Pointal Laurent , Apprendre à programmer dans l'écosystème Python, Dunod, 2025

Documentation officielle Python : <https://docs.python.org/fr/3/>

Jurafsky, Daniel & Martin, James, Speech and Language Processing, 3e édition, draft en ligne : <https://web.stanford.edu/~jurafsky/slp3/> – référence majeure.

Kurdi, Mohamed Zakaria, Traitement automatique des langues et linguistique informatique, Vol. 1, ISTE Éditions, 2017.

Tanguy, Ludovic, « Approches statistiques et linguistiques : quelles complémentarités pour le TAL ? », Traitement automatique des langues, 2005.

UE 2 : Informatique fondamentale (7 ects)

M1SOL020

Epistémologie des Sciences Humaines et de l'Informatique

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

Ce cours propose de traiter des concepts généraux, selon une approche épistémologique, de l'informatique et de la modélisation dans les sciences humaines et sociales et dans la simulation du vivant. Les modèles scientifiques sont des constructions destinées à prédire certains aspects d'une réalité objective qui existe indépendamment de l'observateur. Le comportement de l'observateur est lui aussi modélisé. On insistera sur la différence entre les modèles, qui sont construits par les scientifiques, et la réalité, qui existe indépendamment des modèles. Les modèles sont des approximations successives de la réalité. Ce cours portera également sur les conséquences éthiques et sociétales de ces modèles sur des questions comme les décisions algorithmiques, la protection de la vie privée, les relations affectives avec les machines et l'évolution-humain-machine.

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M1SOL022

Paradigme de la programmation objet

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

Paradigme de la programmation objet (objet, classes et héritage), Gestion des exceptions et des flots d'entrées-sorties, Traitement des textes et des documents XML, Interface graphique et conception d'interfaces, Développement et test en équipe sous Eclipse (Git, Junit)

Bibliographie :

MEYER Bertrand, Conception et Programmation orientées objet, Eyrolles, 2000. BERTHIE V. et BRIAUD J.-B., Swing la synthèse : Développement des interfaces graphiques en Java, Dunod, 2003. DAUZON S., Git - Maîtrisez la gestion de vos versions, Epsilon, 2016.

M1SOL023

Méthodologie de la recherche en Langue et Informatique

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

Ce cours propose une introduction à l'apprentissage machine et la reconnaissance des formes, partie importante de l'intelligence artificielle. Les systèmes d'apprentissage sont de plus en plus utilisés dans de nombreux domaines, de la santé, du transport, de l'éducation ou de La Défense : traduction automatique, reconnaissance visuelle d'objet, reconnaissance de la parole, détection des opinions, etc. Comment valider ces modèles ? La logique de la modélisation est de confronter des données empiriques issues de l'observation aux sorties des simulations afin de valider les modèles. La problématique des bases de données de référence et des protocoles d'évaluation sera largement abordée. Un projet de reconnaissance des formes sera demandé pendant ce cours.

Bibliographie :

Pattern Classification, R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2001

M1SOL041

Représentation et traitement des documents électroniques

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

L'accès à l'information présente dans les documents disponibles sous forme numérisée n'est pas automatique pour la machine. En effet, un certain nombre d'informations sont computationnellement opaques, c'est à dire qu'elles ne sont pas directement accessibles pour la machine même si elles sont évidentes pour les humains. Par exemple, un humain est capable de dater un document même si l'information n'est pas explicitement mentionnée.

Les documents électroniques visent à standardiser, à normaliser la manière dont on encode les documents de manière à faciliter l'accès à l'information. C'est par exemple le cas des flux RSS au format XML où les informations telles que la date ou l'auteur vont être explicitement encodées dans le document de manière à faciliter le traitement automatique ultérieur.

Dans ce cours seront présentés différents formats de structuration des données dont notamment XML et JSON ainsi que des modalités pratiques de modélisation et d'utilisation exploitant ces formats.

Bibliographie :

Harold, E. R., Means, W. S. « XML in a Nutshell (XML en concentré) », 3rd Edition. O'reilly, 2009

SALVADOR X.L. « XML pour les linguistes », 2016

XML Tutorial. <<https://www.w3schools.com/xml/>>

Basset, L. « Introduction to JavaScript Object Notation », O'reilly, 2015

Griffiths, D. F., Higham, D. J. « Learning LaTeX », 2nd Edition, SIAM, 2016.

M1SOL021

Représentation et modélisation des connaissances

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

Qu'est qu'une connaissance ? Que signifie « représentation » ? Pourquoi représenter et modéliser des connaissances ? Ce cours part de ces questions qui, à première vue, pourraient sembler un peu naïves. Mais les réponses qu'on apporte à ces questions tout au long de ce cours sont très importantes pour l'informatique en général et pour l'intelligence artificielle symbolique en particulier. On s'intéresse à des modèles et des formalismes de représentation de connaissances dont le but est de permettre à des programmes informatiques de traiter des informations au niveau sémantique. On se familiarise notamment avec des réseaux sémantiques et des logiques de description. Les dernières séances du cours sont consacrées à Neo4j qui est un système de gestion de base de données orienté graphe.

Bibliographie :

Quillian, M. R. (1968). Semantic memory. Semantic information processing, 227–270.

Kayser, D. (1997). La représentation des connaissances. Hermès.

Stefik, M. (1993). Introduction to Knowledge Systems. Morgan Kaufmann.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., & Nardi, D. (Eds.). (2003). The description logic handbook: Theory, implementation and applications. Cambridge university press.

Eifrem, E., Webber, J., Robinson, I. (2015). Graph Databases, 2nd Edition. O'reilly.

Baton, J., Van Bruggen, R. (2017). Learning Neo4j 3.x, 2nd Edition. Packt.

UE 4 : Linguistique computationnelle (8 ects)

M1SOL030

Modèles de linguistique computationnelle

Évaluation :
Contrôle mixte

Volume horaire :
1h CM + 1,5h TD

Descriptif :

Le cours porte sur les bases du traitement automatique du langage. Les sujets abordés sont : les modèles de langue par n-grammes, l'analyse automatique de parties du discours (chaîne de Markov cachée), l'analyse syntaxique automatique, l'usage et la normalisation des corpus et les métriques d'évaluation de systèmes. L'évaluation du cours se passe par le rendu d'exercices TD, un projet de groupe et un examen à la fin du semestre.

Bibliographie :

Speech and Language Processing (2nd edition). Dan Jurafsky and James H. Martin

M1SOL032

Grammaires formelles

Évaluation :
Contrôle mixte

Volume horaire :
1h CM + 1,5h TD

Descriptif :

L'objectif du cours est de familiariser les étudiants avec les notions de langage et de grammaire formelle. Les grammaires formelles seront présentées en détails (types de règles, types de grammaires, types de langages, hiérarchie des langages, éléments récurrents d'une grammaire, structures, langages et grammaires) et replacées dans leur contexte (historique et thématique). Les notions théoriques présentées seront complétées par de nombreux exercices les illustrant. Au moins un algorithme d'analyse syntaxique sera détaillé (CKY).

Bibliographie :

Chomsky, N. Aspects of the Theory of Syntax MIT, 1965

Church, K. A Pendulum Swung Too Far Linguistic Issues in Language Technology - LiLT, 2011, 6.

M1SOL050

Sémantique computationnelle

Évaluation :
Contrôle mixte

Volume horaire :
1h CM + 1,5h TD

Descriptif :

Ce cours explore les notions théoriques et pratiques de la modélisation du sens en langage naturel, en articulant approches linguistiques, informatiques et traitement automatique du langage. Le cours s'appuie sur le logiciel GATE (General Architecture for Text Engineering) pour introduire des outils de traitement sémantique, d'annotation linguistique et d'extraction d'information. Après avoir présenté l'architecture de GATE, le cours approfondira l'environnement de programmation graphique suivi de l'API permettant le développement de composants personnalisés. Partant de l'utilisation des analyseurs lexicaux, dictionnaires et étiqueteurs morphosyntaxiques; le cours poursuivra vers l'annotation sémantique, l'extraction d'entités nommées, l'apprentissage du langage JAPE et l'utilisation avancée de la plateforme GATE pour l'analyse et l'annotation sémantique de textes. L'objectif du cours est de permettre aux étudiant(e)s de comprendre et d'utiliser les méthodes de sémantique computationnelle dans des applications concrètes comme la recherche d'information ou la désambiguïsation du sens des mots.

Bibliographie :

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

Cunningham, et al. Developing Language Processing Components with GATE Version 8. University of Sheffield Department of Computer Science. 17 November 2014..

UE 5 : Langues vivantes (3 ects)

Un cours de langue au choix. Se renseigner auprès des UFR de Langues ou du SIAL
<http://sial.paris-sorbonne.fr>

UE Enseignement sans crédit

Master 1 - Semestre 2

UE 1 : Méthodologie et épistémologie (7 ects)

M2SOL031

Modélisation des productions langagières orales

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

Synthèse des acquis concernant les différences entre les caractéristiques de l'écrit et celles de l'oral.

Approches des différents courants (pragmatique, ethnométhodologie, anthropologie de l'oral, ethnologie, sociologie et linguistique) qui ont conduit à une analyse linguistique et paralinguistique de l'oral.

Typologie des interactions verbales.

Analyse des parties constitutives des interactions verbales à partir de corpus réels (échanges, tours de parole, études des différentes séquences etc.).

Etude des disfluences (hésitations, pauses, amorces et reprises)

Analyse syntaxique et lexicale avec modélisation (analyse en grille, arbres marcottés).

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M2LF13FA

Linguistique Théorique appliquée

Évaluation :

cf Langue Française

Volume horaire :

2h par semaine

Descriptif : Cf UFR de Langue Française

Bibliographie : Cf UFR de Langue Française

M2LF15FG

Terminologie linguistique

Évaluation :

cf Langue Française

Volume horaire :

2h par semaine

Descriptif : cf Langue Française

Bibliographie : cf Langue Française

UE 2 : Informatique avancée (6 ects)

M2SOL020

Programmation générique et conception objet

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

Programmation à base de structures de données abstraites, polymorphisme (ad-hoc, inclusion, patrons de classe et foncteurs), Librairie standard de composants génériques (conteneurs séquentiels et itérateurs, relation d'ordre et conteneurs associatifs, algorithmes et foncteurs, traitement de textes, expressions régulières), Développement sous Eclipse en (debug, makefile) et interopérabilité logicielle, Méthodologies et heuristiques de conception (classes abstraites, interfaces et exceptions)

Bibliographie :

DELANNOY C. Programmer en langage C++, Eyrolles.

M2SOL021 Internet et bases de données

Évaluation :
Contrôle mixte

Volume horaire :
1h CM + 1,5h TD

Descriptif :

Réseaux informatiques (couches OSI), Programmation réseau (sockets, processus légers)
Notions de protocoles (messagerie, Langage HTML et transfert hypertexte HTTP)
Bases relationnelles XML; Modèle de conception client-serveur (Appel de procédures à distance, Accès distant aux bases de données JDBC)
Services Web et plateforme J2EE

Bibliographie :

DJAFAR K., Développement J2EE avec Eclipse et WSAD, Eyrolles

M2SOL033 Web sémantique et Big-data

Évaluation :
Contrôle mixte

Volume horaire :
1h CM + 1,5h TD

Descriptif :

Le cours est composé de deux parties distinctes et complémentaires. La première partie concerne le Web sémantique et l'étude des principes et systèmes de représentation des connaissances, les réseaux sémantiques, les logiques de description la construction de représentations conceptuelles et d'ontologies ainsi que les premiers formalismes des systèmes de représentation de connaissances sur le Web (Web sémantique). Des techniques et processus d'acquisition des connaissances seront présentés, suivis de notions d'inférence ontologique pour l'extraction de nouvelles connaissances. La deuxième partie se centre sur le Big Data et les principes et pratiques des systèmes de données et calcul distribués. Des outils comme Hadoop et Apache Spark seront présentés.

Bibliographie :

Le web sémantique : Comment lier les données et les schémas sur le web ?

Mai, 2012. Fabie Gandon, Catherine Faron-Zucker, Olivier Corby

Big Data: Principles and best practices of scalable realtime data systems. Nathan Marz, James Warren

Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale. Tom White

Spark: The Definitive Guide: Big Data Processing Made Simple. Bill Chambers, Matei Zaharia.

UE 3 : Traitement automatique du langage (5 ects)

M2SOL024 Programmation pour le TAL

Évaluation :
Contrôle mixte

Volume horaire :
1h CM + 1,5h TD

Descriptif :

Ce cours approfondit les bases acquises en licence en traitement automatique du langage (TAL), en se concentrant sur l'utilisation et la personnalisation des modèles récents d'intelligence artificielle, notamment les grands modèles de langage (LLMs), dans des contextes de recherche en SHS.

Il articule des compétences de programmation Python avancée, la maîtrise de bibliothèques modernes comme spaCy, transformers (Hugging Face) et scikit-learn, avec une réflexion sur l'usage critique de ces outils dans l'analyse de corpus.

Les étudiant·e·s apprendront à :

Revenir sur les pipelines classiques de TAL (tokenisation, POS-tagging, parsing, NER).

Mettre en œuvre des modèles pré-entraînés de type BERT/CamemBERT pour des tâches de classification de texte, d'étiquetage de séquences (sequence labeling) ou d'extraction d'information.

Effectuer un fine-tuning léger (entraînement d'un modèle existant sur un corpus spécialisé).

Analyser la performance, les biais et les limites des modèles utilisés dans un cadre SHS.

Une attention particulière sera portée à l'interprétation des sorties modèles, à l'équité algorithmique, à l'explicabilité, et à la reproductibilité des expériences.

Le cours sera structuré autour de TP en Python (Jupyter Notebooks) avec projet final individuel ou en binôme sur un corpus choisi.

Bibliographie :

- BENDER, E. M., DEBRU, T., MCMILLAN-MAJOR, SHMITCHELL. On the Dangers of

Stochastic Parrots. FaccT, 2021
 - EINSENSTEIN, J., Natural Language Processing. MIT Press., 2019
 - HUGGING FACE documentation : <https://huggingface.co/docs>, 2025
 - JURAFSKY, D. & MARTIN, J. Speech and Language Processing. <https://web.stanford.edu/~jurafsky/slp3/>, 2023

M2SOL035 Introduction au traitement automatique des langues TAL

Évaluation :
 Contrôle mixte

Volume horaire :
 1h CM + 1,5h TD

Descriptif :

L'objectif de ce cours est triple : 1) développer la maîtrise d'un outil d'analyse statistique textuelle (textométrie), TXM 2) faire découvrir le TAL et les questions qu'il soulève dans d'autres langues que le français et l'anglais et 3) montrer les possibilités d'analyse offertes par les grammaires locales, en particulier à travers l'outil Unitex. Les étudiants seront amenés à réaliser, en groupes, des recherches et des présentations sur des sujets liés. Tous les outils et ressources utilisées sont librement disponibles et en sources ouvertes.

Bibliographie :

TXM, <http://textometrie.ens-lyon.fr/?lang=fr>. Unitex : <https://unitexgramlab.org/fr>

UE 4 : Mémoire (9 ects)

M2SOLMEM Soutenance du mémoire

UE 5 : Langues vivantes (3 ects)

Un cours de langue au choix. Se renseigner auprès des UFR de Langues ou du SIAL
<http://sial.paris-sorbonne.fr>

UE Enseignement sans crédit

Master 2 - Semestre 3

UE 1 : Méthodologie et épistémologie (7 ects)

M3LF13FG Linguistique Théorique appliquée

Évaluation :
 cf Langue Française

Volume horaire :
 2h par semaine

Descriptif : Cf UFR de Langue Française

Bibliographie : Cf UFR de Langue Française

M3LF15FA Terminologie linguistique

Évaluation :
 cf Langue Française

Volume horaire :
 1h par semaine

Descriptif : Cf UFR de Langue Française

Bibliographie : Cf UFR de Langue Française

M3LF00ST

Conférences de linguistique

Volume horaire :
26h sur le semestre

Descriptif : Cf UFR de Langue Française

UE 2 : Informatique, langage (8 ects)

M3SOL030

Production et perception de la parole

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours présente les mécanismes physiologiques, acoustiques et cognitifs qui sous-tendent la perception et la production des sons, des mots et des énoncés oraux de la langue parlée. Articulateur et acoustique (communication parlée, anatomie du cortex cérébral, phonétique et phonologie, phonématique, introduction à la phonétique articulaire et acoustique), Prosodie (définitions et généralités, caractérisations de la substance prosodique, étiquetage prosodique, accentuation, intonation, méthodes d'analyse), Phonétique acoustique (onde acoustique, transduction, numérisation, analyse fréquentielle, modélisation de la production, analyse spectrale, coarticulation, lecture de spectrogrammes), Phonétique articulaire, Perception auditive (psychoacoustique, sonie des sons purs, masquage fréquentiel et temporel, physiologie du système auditif, application d'un modèle psychoacoustique)

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M3SOL035

Dialogues et serveurs bots

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours a pour objectif de fournir aux étudiants les bases théoriques et pratiques nécessaires à la conception, au développement et au déploiement d'un agent conversationnel finalisé, fonctionnant sur un serveur bot. Il s'appuie sur les principes de l'interaction homme-machine, de la modélisation linguistique et de l'intelligence artificielle appliquée au dialogue. Après une introduction aux modèles de tâches et de dialogue (HTA, GOMS), le cours présente une typologie des situations de dialogue (dialogues dirigés, mixtes, collaboratifs), permettant de distinguer les besoins fonctionnels des différents types d'agents. Une attention particulière est accordée à la méthode intent/slot, utilisée dans la majorité des systèmes de dialogue actuels. Cette approche permet d'identifier l'intention d'une requête utilisateur (intent) et d'en extraire les informations pertinentes (slots), en vue d'exécuter une tâche ou d'engager une réponse adaptée. Les étudiants apprendront à mettre en œuvre cette méthode dans des environnements personnalisés. Parallèlement, le cours introduit les modèles de langage, depuis les grammaires stochastiques et modèles N-Grammes jusqu'aux grands modèles de langage (LLM). Ces derniers sont étudiés dans une perspective critique, avec un accent sur leur intégration dans des systèmes dialogiques hybrides et leur évaluation (perplexité, exactitude, cohérence, etc.). À l'issue du cours, les étudiants seront capables de concevoir un gestionnaire de dialogue robuste, combinant méthodes symboliques (règles, AIML, intent/slot) et modèles statistiques ou neuronaux (LLM), dans une architecture client-serveur fonctionnelle.

Bibliographie :

McTear, M. (2020). Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Springer.
Pieraccini, R. (2021). AI Assistants. Morgan & Claypool Publishers.
López, Á. J. G., & Fernández, A. R. (2019). Building Chatbots with Python. Apress.

M3SOL034	Indexation sémantique et recherche d'information
-----------------	---

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours introduit les principes fondamentaux de la recherche d'information, en mettant l'accent sur les techniques d'indexation sémantique. Le cours commence par une introduction aux principales technologies de recherche d'information, comprenant les algorithmes, structures de données, moteurs de recherche, extraction et évaluation. L'accent sera mis sur les concepts d'indexation, de classification et de catégorisation. Ensuite, le cours se focalise sur l'identification des étapes-clé du processus de veille, les outils logiciels, les sources d'information utiles et les méthodes d'analyse des informations collectées. Sur le plan technique, les étudiant·e·s conçoivent un projet concret de moteur de recherche en utilisant Apache Lucene (moteur d'indexation) et Apache Solr (plateforme de recherche), afin de mettre en œuvre un système complet d'indexation et de recherche sémantique.

Bibliographie :

Recherche d'information, M.-R. Amini , E. Gaussier, 2013.
Solr in Action. Trey Grainger and Timothy Potter. March 2014. ISBN 9781617291029
Lucene in Action, Second Edition. Michael McCandless, Erik Hatcher, and Otis Gospodnetić. July 2010. ISBN 9781933988177
Information Retrieval: Implementing and Evaluating Search Engines (The MIT Press). Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack.

M3SOL047	Rationalité dans les interactions homme-machine
-----------------	--

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

La problématique des sciences affectives fait aujourd'hui un retour en force dans la communauté scientifique que ce soit en sciences humaines et sociales ou en sciences de l'information et de la communication, cela en partie grâce aux récentes découvertes en neurosciences. Les sciences humaines, la psychologie, l'éthologie et la sociologie ne sont plus les seules disciplines concernées par cette problématique : les linguistes, qu'ils soient spécialistes du discours, des interactions sociales ou de la parole, s'intéressent depuis peu aux émotions. Les émotions appellent un traitement pluridisciplinaire, qu'il s'agisse d'analyser les processus neurophysiologiques qui les sous-tendent ou leurs manifestations verbales, vocales et non verbales, qu'il s'agisse de les appréhender en laboratoire ou dans des conditions écologiques, qu'il s'agisse de les théoriser ou d'analyser leur rôle dans les processus cognitifs, les activités d'apprentissage ou les conduites de communication ou d'interaction sociale. Ce cours s'attache à montrer l'aspect pluridisciplinaire des sciences affectives et propose une introduction aux modèles computationnels utilisés pour reconnaître, raisonner et simuler des processus affectifs en interaction sociale. Cette nouvelle discipline s'appelle l'affective computing et prend ces sources dans les travaux de Rosalind Picard en 1997. Des outils de détection, raisonnement et simulation (détection des émotions, interaction affective, chatbot affectif, robot social...) seront présentés à travers des réalisations en R&D pour diverses applications au niveau international.

Bibliographie :

L'erreur de Descartes, Antonio Damasio, 1994
Affective computing, Rosalind Picard, 1997
Des robots et des hommes: mythes, fantasmes et réalité, Laurence Devillers, Plon 2017.

UE 3 : Traitement automatique (5 ects)

M3SOL025	Paradigmes d'évaluation du traitement automatiques des langues TAL
-----------------	---

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours vise à fournir aux étudiant·e·s une compréhension avancée, critique et pratique des méthodes d'évaluation en traitement automatique du langage (TAL). L'évaluation est un volet central du développement et de la mise en œuvre des modèles, qu'ils soient symboliques, statistiques ou neuronaux.

Les paradigmes d'évaluation abordés selon différents angles :

- Performance des modèles supervisés (classification, étiquetage, extraction).

- Evaluation des tâches non supervisées (clustering, embeddings).
- Mesure de la qualité des corpus et des annotations (accord inter-annotateur, biais, bruit).
- Evaluation des LLMs avec et sans tâche définie
- Discussion de la reproductibilité, de l'interprétabilité et des enjeux éthiques de l'évaluation.

Une série d'études de cas et un mini-projet ponctueront le cours.

Bibliographie :

- ARSTEIN & POESIO Inter-Coder Agreement for Computational Linguistics. CL., 2008
- ETHAYARAJH, How Contextual Are Contextualized Word Representations? ACL, 2020
- JURAFSKY, D. & MARTIN, J.. Speech and Language Processing. <https://web.stanford.edu/~jurafsky/slp3/> 2023, (sections sur sur l'évaluation)
- LIANG. et al. . Holistic Evaluation of Language Models (HELM). Stanford CRFM., 2023
- **POWERS. . Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2020**

M3SOL024

Traduction automatique

Évaluation :

CCI

Volume horaire :

2,5h CM

Descriptif :

L'objectif de ce cours est de comprendre les fondements théoriques et historiques de la traduction automatique (TA), maîtriser les approches symbolique, statistique et neuronale, identifier les limites linguistiques des systèmes de TA face aux idiomatismes, aux figures de style (métaphores, hyperboles, euphémismes, etc.), aux registres de langue et aux contextes culturels, d'expérimenter avec des outils de TA (DeepL, Google Translate, OpenNMT, ChatGPT, etc.), de mettre en œuvre une évaluation critique, tant automatique qu'humaine, des traductions générées. Ce contenu s'intègre dans une optique de formation à la traduction littéraire assistée, à la post-édition avancée et à l'évaluation linguistique fine.

Bibliographie :

Birke, J., & Sarkar, A. (2006). A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. EACL.

Shutova, E. (2011). Design and Evaluation of Metaphor Processing Systems. Computational Linguistics, 39(1).

L'Homme, M.-C., & Tutin, A. (2021). Lexical Approaches to Figurative Language in Multilingual Contexts. Lexis, 16.

Mathur, N., Bhattacharyya, P., & Specia, L. (2020). Results of the WMT20 Metrics Shared Task: Evaluating Metrics with Human Judgments.

Klein, G. et al. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. ACL System Demonstrations.

UE 4 : Société Numérique (7ects)

M3SOL043

Management de projets

Évaluation :

CCI

Volume horaire :

2,5h CM

Descriptif :

Sensibiliser les auditeurs au management de projet. Promouvoir le management de type pro-actif. Démontrer la nécessité de la gestion des risques comme moyen de maîtrise de l'incertitude. Sensibiliser les auditeurs à la qualité du logiciel. Accroître le professionnalisme des auditeurs à travers un partage d'expérience. Fournir les moyens nécessaires à une auto-formation.

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M3SOL045	Génie logiciel
-----------------	-----------------------

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

L'enseignement de génie logiciel présente, au travers de cours magistraux et travaux dirigés, les méthodes de travail permettant le développement de logiciels de manière industrialisée et orienté qualité. Le cours aborde le langage UML, la méthodologie de tests et la conception logicielle au travers de modèles. Le cours introduira également les notions et concepts largement portés par les nouvelles méthodes de travail en lien avec la transformation numérique (philosophie et principes des pratiques agiles et zoom sur la méthode Scrum).

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M3SOL046	Conférences de professionnels
-----------------	--------------------------------------

UE 5 : Langues vivantes (3 ects)

Un cours de langue au choix. Se renseigner auprès des UFR de Langues ou du SIAL
<http://sial.paris-sorbonne.fr>

UE Enseignement sans crédit

Master 2 - Semestre 4

UE 1 : Méthodologie et épistémologie (4 ects)

M4SOL304	Analyse de graphes et bases de données NoSQL
-----------------	---

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours présente les concepts clés et outils pratiques pour l'analyse de graphes et l'utilisation des bases de données NoSQL. Il aborde la modélisation des données sous forme de graphes ainsi que les techniques d'exploration et de visualisation des relations entre entités. La première partie se concentre sur les bases de données NoSQL, avec un focus particulier sur les bases orientées graphes, notamment Neo4j, reconnue pour son architecture flexible adaptée aux données non structurées ou semi-structurées. La seconde partie porte sur les méthodes d'analyse de graphes : détection de communautés, mesures de centralité, parcours et algorithmes de propagation. Ces techniques sont mises en pratique avec Gephi, un logiciel open source de visualisation et d'analyse de réseaux, et illustrées par des exemples issus des réseaux sociaux, de la bioinformatique ou de la recommandation.

Bibliographie :

Bruchez, R. (2015). *Les bases de données NoSQL et le BigData*: Comprendre et mettre en œuvre. Editions Eyrolles. 02/12/2021 (3eme édition)
Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

M4SOL301

Structure de connaissances et raisonnement

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours vise à fournir aux étudiants les compétences nécessaires pour : Maîtriser la programmation logique pour la représentation de connaissances et l'automatisation du raisonnement, Comprendre et concevoir des systèmes d'argumentation à partir de données multimodales, Développer des modèles hybrides combinant apprentissage automatique et raisonnement symbolique, Modéliser le raisonnement sur les actions dans des environnements dynamiques et incertains, Concevoir des systèmes multimodaux intelligents intégrant texte, image, son, etc, Intégrer les enjeux éthiques, sociaux et épistémologiques liés à la modélisation des connaissances et à l'IA. Le plan du cours sera

Programmation logique et représentation des connaissances (Logique des prédicats, Horn clauses, Prolog, Bases de connaissances, règles, faits, inférence déductive, Raisonnement non-monotone et logique abductive), Raisonnement automatisé dans des environnements dynamiques (Logique temporelle, logique d'actions (Situation Calculus, Event Calculus), Raisonnement sur les effets indirects et incertains, Intégration de capteurs/sources de données externes), Systèmes d'argumentation (Modèles formels d'argumentation (Dung, ASPIC+), Raisonnement sur les conflits et la persuasion), Apprentissage et raisonnement hybride (Symbolic AI vs. Deep Learning : limites et complémentarité, Intégration de réseaux de neurones avec logique (neuro-symbolic systems), Méthodes inductives : ILP (Inductive Logic Programming)), Raisonnement multimodal (Fusion de sources : texte, image, son, vidéo, Raisonnement croisé sur des modalités hétérogènes, Architectures de systèmes intelligents multimodaux (VQA, image captioning + logique)

Bibliographie :

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson (Ch. 7–9 : Logic and knowledge representation)

Thielscher, M. (2011). *A Brief Introduction to Fluent Calculus*. AI Communications.

Cabrio, E., & Villata, S. (2018). *Five Years of Argument Mining: A Data-driven Analysis*. IJCAI.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). *Multimodal Machine Learning: Survey and Taxonomy*. IEEE TPAMI.

UE 2 : Paralinguistique (6 ects)

M4SOL201

Reconnaissances de la parole

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours présente les paradigmes de la reconnaissance de la parole et des caractéristiques individuelles du locuteur et les algorithmes associés. Reconnaissance des formes vocales (tâches et objectifs en reconnaissance, principe des applications, espace de représentation des formes, concept de classe, apprentissage des classes, reconnaissance/identification), Techniques de reconnaissance des formes vocales (stratégies de reconnaissance, fusion des classificateurs, adaptation par maximum a posteriori), Tâches de reconnaissance (facteurs de complexité du signal vocal, facteurs de complexité de la tâche), Reconnaissance vocale (reconnaissance de continuums acoustiques, programmation dynamique, comparaison dynamique, application à la reconnaissance d'entités vocales isolées), modèles de Markov cachés (théorie de l'information, processus de Markov, modélisation du canal acoustique, méthodes d'apprentissage, application à la reconnaissance en mots isolés et en parole continue). Biométrie physique et caractéristiques individuelles (authentification auditive et visuelle), Reconnaissance du locuteur (identification, vérification, suivi, détection de changement)

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M4SOL202	Synthèse de la parole
-----------------	------------------------------

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours présente les paradigmes de la synthèse de la parole à partir du texte, de l'analyse synthèse et des algorithmes associés, les effets de facteurs physiologiques, pathologiques, psychologiques et comportementaux sur la parole. Synthèse à partir du texte (conversion graphèmes-phonèmes et interface syntaxe-prosodie) Prédiction linéaire (modèle linéaire source/filtre, méthode d'autocorrélation, vocodeur) Codage de la parole, analyse-synthèse et synthèse acoustique ; Algorithmes de synthèse par corpus. Voix pathologiques (dysphonies, dysarthries, dysprosodie) et qualité de la parole (caractérisation), Stress et parole stressée (indicateurs du stress et de la fatigue, analyse de la voix stressée ; Etats paralinguistiques et méthodes de reconnaissance par Deep Neural Network

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

UE 3 : Sémantique, Connaissances et Corpus (6 ects)

M4SOL305	Représentations vectorielles et similarités
-----------------	--

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours vise à comprendre l'évolution des représentations linguistiques vectorielles (distributionnelles, embeddings, contextualisées) au fil du temps et leur relation avec les mesures de similarité indispensables à des tâches de TAL (retrieval, appariement, clustering, scoring). Ce cours est en liaison directe avec le cours PETAL (Paradigmes d'Evaluation pour le TAL du premier semestre).

L'objectif premier est de bien distinguer le procédé de vectorisation (Bag of Words par exemple) des métriques de comptage et de pondération (one hot encoding, Tf-Idf, BM 25 ...) afin de pouvoir les combiner et les comprendre dans des contextes industriels ou de recherche.

Bibliographie :

Pas de bibliographie spécifique, nous reprendrons avec une autre perspective des éléments figurant dans les cours M2SOL024 (Programmation pour le TAL) et M3SOL025 (Paradigmes d'évaluation du traitement automatique des langues TAL)

M4SOL303	Annotation collaborative de corpus
-----------------	---

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Avec l'essor de l'apprentissage supervisé, la formalisation linguistique, au sein du TAL, s'est déplacée des grammaires locales à l'annotation manuelle. Or, celle-ci reste un processus mal maîtrisé. Ce cours présente les enjeux de l'annotation manuelle de corpus pour le TAL, les problématiques liées à sa formalisation et à son évaluation et les solutions proposées, notamment la myriadisation (crowdsourcing), en particulier via des jeux ayant un but (GWAP).

Bibliographie :

Fort, K. Collaborative Annotation for Reliable Natural Language Processing. ISTE Wiley, 2016

Guillaume, B.; Fort, K. & Lefebvre, N. Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. Proc. of the 26th International Conference on Computational Linguistics (COLING), 2016, 3041-3052

Fort, K.; Adda, G. & Cohen, K. B. Amazon Mechanical Turk: Gold Mine or Coal Mine? Computational Linguistics (editorial), 2011, 37, 413-420

Fort, K.; Nazarenko, A. & Rosset, S. Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis. Proc. of the International Conference on Computational Linguistics (COLING), 2012, 895-910

UE 4 : Finalité : recherche ou professionnelle (11 ects)

M4SOLMEM	Soutenance du mémoire de recherche
M4SOLSTG	Soutenance du stage en entreprise

UE 5 : Langues vivantes (3 ects)

Un cours de langue au choix. Se renseigner auprès des UFR de Langues ou du SIAL
<http://sial.paris-sorbonne.fr>

CONTACTS

UFR de Sociologie et Informatique pour les Sciences Humaines

Secrétariat

Maison de la Recherche 28, rue Serpente 75006 Paris

Téléphone : 01 53 10 58 30

Courriel : Lettres-Sociologie-Informatique-secretariat@sorbonne-universite.fr

Ouverture de 9h30 à 12h30 et de 13h30 à 16h30

Fermeture les Lundis et Vendredis après-midi