

Master parcours Langue et Informatique

**Mention
Sciences du langage**

Spécialité
Recherche ou Professionnelle



SOMMAIRE

SOMMAIRE	3
Présentation de la mention Sciences du langage	4
Offre de formation	4
Organisation	4
Présentation du parcours Langue et Informatique	6
Enseignants	6
Enseignants extérieurs	6
Intervenants professionnels	6
Objectifs	7
Conditions d'admission	7
Contenu	7
Modalités de contrôle des connaissances (MCC)	7
Calendrier facultaire	8
Organisation de la formation	9
Master 1 - Semestre 1	9
Master 1 - Semestre 2	10
Master 2 - Semestre 3	11
Master 2 - Semestre 4	12
Descriptifs des enseignements	13
Master 1 - Semestre 1	13
UE 1 : Méthodologie et épistémologie (7 ects)	13
UE 2 : Informatique fondamentale (7 ects)	14
UE 3 : Ingénierie des connaissances (5 ects)	14
UE 4 : Linguistique computationnelle (8 ects)	15
UE 5 : Langues vivantes (3 ects)	16
Master 1 - Semestre 2	17
UE 1 : Méthodologie et épistémologie (7 ects)	17
UE 2 : Informatique avancée (6 ects)	17
UE 3 : Traitement automatique du langage (5 ects)	18
UE 4 : Mémoire (9 ects)	19
UE 5 : Langues vivantes (3 ects)	19
Master 2 - Semestre 3	19
UE 1 : Méthodologie et épistémologie (7 ects)	19
UE 2 : Informatique, langage (8 ects)	19
UE 3 : Traitement automatique (5 ects)	21
UE 4 : Société Numérique (7ects)	21
UE 5 : Langues vivantes (3 ects)	22
Master 2 - Semestre 4	22
UE 1 : Méthodologie et épistémologie (4 ects)	22
UE 2 : Paralinguistique (6 ects)	22
UE 3 : Sémantique (6 ects)	23
UE 4 : Finalité : recherche ou professionnelle (11 ects)	25
UE 5 : Langues vivantes (3 ects)	25
Contact	25

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, claude.montacie@sorbonne-universite.fr
Laurence Devillers, devil@limsi.fr

Enseignants

Devillers Laurence, Professeur
Montacié Claude, Professeur
Eyharabide Victoria, MCF
Guérin Françoise, MCF
Fort Karën, MCF
Lejeune Gaël, MCF
Pascal Boldini, PRAG
Ascencio Laurie, ATER
Ben Chaabene Nour El Houda, ATER
Ben Ltaifa Ibtihel, ATER
Zine Oumaima, ATER

Enseignants extérieurs

Caraty Marie-José, Professeur

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 2023 - 2024

(hors écoles internes : Celsa et INSPÉ)

adopté en conseil de la Faculté des Lettres le 18 Juin 2023, en CFVU le 28 Juin 2023 et en Conseil d'Administration de Sorbonne Université le 04 Juillet 2023

1er semestre : du 18 septembre 2023 au 18 janvier 2024 inclus / **2nd semestre** : du 29 janvier au 02 juillet 2024 inclus

Fin d'année universitaire : le 30 septembre 2024 (stages non crédités d'ECTS des cursus Licence, Master 1 et Master 2 R) ou le 16 décembre 2024 (stages crédités d'ECTS en Master 2 P)

2023				2024								DEBUT D'ANNÉE UNIVERSITAIRE SUIVANTE			
SEPTEMBRE	OCTOBRE	NOVEMBRE	DECEMBRE	JANVIER	FÉVRIER	MARS	AVRIL	MAI	JUIN	JUILLET	AOUT	SEPTEMBRE	OCTOBRE	NOVEMBRE	DECEMBRE
1 V	1 D	1 M	1 V	1 L	1 J	1 V	1 L	1 M	1 S	1 L	1 J	1 D	1	1	1
2 S	2 L	2 J	2 S	2 M	2 V	2 S	2 M	2 J	2 D	2 M	2 V	2 L	2	2	2
3 D	3 M	3 V	3 D	3 M	3 S	3 D	3 M	3 V	3 L	3 M	3 S	3 M	3	3	3
4 L	4 M	4 S	4 L	4 J	4 D	4 L	4 J	4 S	4 M	4 J	4 D	4 M	4	4	4
5 M	5 J	5 D	5 M	5 V	5 L	5 M	5 V	5 D	5 M	5 V	5 L	5 J	5	5	5
6 M	6 V	6 L	6 M	6 S	6 M	6 M	6 S	6 L	6 J	6 S	6 M	6 V	6	6	6
7 J	7 S	7 M	7 J	7 D	7 M	7 J	7 D	7 M	7 V	7 D	7 M	7 S	7	7	7
8 V	8 D	8 M	8 V	8 L	8 J	8 V	8 L	8 M	8 S	8 L	8 J	8 D	8	8	8
9 S	9 L	9 J	9 S	9 M	9 V	9 S	9 M	9* J	9 D	9 M	9 V	9 L	9	9	9
10 D	10 M	10 V	10 D	10 M	10 S	10 D	10 M	10 V	10 L	10 M	10 S	10 M	10	10	10
11 L	11 M	11 S	11 L	11 J	11 D	11 L	11 J	11 S	11 M	11 J	11 D	11 M	11	11	11
12 M	12 J	12 D	12 M	12 V	12 L	12 M	12 V	12 D	12 M	12 V	12 L	12 J	12	12	12
13 M	13 V	13 L	13 M	13 S	13 M	13 M	13 S	13 L	13 J	13 S	13 M	13 V	13	13	13
14 J	14 S	14 M	14 J	14 D	14 M	14 J	14 D	14 M	14 V	14 D	14 M	14 S	14	14	14
15 V	15 D	15 M	15 V	15 L	15 J	15 V	15 L	15 M	15 S	15 L	15 J	15 D	15	15	15
16 S	16 L	16 J	16 S	16 M	16 V	16 S	16 M	16 J	16 D	16 M	16 V	16 L	16	16	16
17 D	17 M	17 V	17 D	17 M	17 S	17 D	17 M	17 V	17 L	17 M	17 M	17 S	17	17	17
18 L	18 M	18 S	18 L	18 J	18 D	18 L	18 J	18 S	18 M	18 J	18 D	18 M	18	18	18
19 M	19 J	19 D	19 M	19 V	19 L	19 M	19 V	19 D	19 M	19 V	19 L	19 J	19	19	19
20 M	20 V	20 L	20 M	20 S	20 M	20 M	20 S	20 L	20 J	20 S	20 M	20 V	20	20	20
21 J	21 S	21 M	21 J	21 D	21 M	21 J	21 D	21 M	21 V	21 D	21 M	21 S	21	21	21
22 V	22 D	22 M	22 V	22 L	22 J	22 V	22 L	22 M	22 S	22 L	22 J	22 D	22	22	22
23 S	23 L	23 J	23 S	23 M	23 V	23 S	23 M	23 J	23 D	23 M	23 V	23 L	23	23	23
24 D	24 M	24 V	24 D	24 M	24 S	24 D	24 M	24 V	24 L	24 M	24 S	24 M	24	24	24
25 L	25 M	25 S	25 L	25 J	25 D	25 L	25 J	25 S	25 M	25 J	25 D	25 M	25	25	25
26 M	26 J	26 D	26 M	26 V	26 L	26 M	26 V	26 D	26 M	26 V	26 L	26 J	26	26	26
27 M	27 V	27 L	27 M	27 S	27 M	27 M	27 S	27 L	27 J	27 S	27 M	27 V	27	27	27
28 J	28 S	28 M	28 J	28 D	28 M	28 J	28 D	28 M	28 V	28 D	28 M	28 S	28	28	28
29 V	29 D	29 M	29 V	29 L	29 J	29 V	29 L	29 M	29 S	29 L	29 J	29 D	29	29	29
30 S	30 L	30 J	30 S	30 M	30 M	30 S	30 M	30 J	30 D	30 M	30 V	30 L	30	30	30
	31 M		31 D	31 M		31 D		31 V		31 M	31 S		31		31

* Ascension

11
ouverture inscriptions Pédagogiques
en ligne (IPweb) semestre 1

22
ouverture inscriptions Pédagogiques
en ligne (IPweb) semestre 2

30
fin d'année
universitaire
2022-2023 /
stages L,
M1 et M2 R

15
fin d'année
universitaire
2022-2023 /
stages M2 P
accès jurys

- Périodes de cours
- Périodes de cours et de révision tutorée
- Examens : session initiale (2 périodes : janvier et mai)
- Examens : session de rattrapage S1 et S2 (juin)
- Jurys de Licence (y compris dérogoatoires)
- Jurys de Master (y compris dérogoatoires)
- Publication des résultats
- Interruption d'études
- Dimanches / Jours fériés

30
fin d'année
universitaire
2023-2024 /
stages L,
M1 et M2 R

16
fin d'année
universitaire
2023-2024 /
stages M2 P
accès jurys

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	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 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			3
	SIAL (langue au choix)	Voir avec le SIAL	Contrôle Continu	

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	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			5
	Dictionnaire et néologismes	1h CM 1,5 h TD	Contrôle Mixte	
	Corpus, ressources et linguistique outillée	1h CM 1,5 h TD	Contrôle Mixte	
UE 4	Mémoire			9
	Soutenance du mémoire			
UE 5	Langues vivantes			3
	SIAL (langue au choix)	Voir avec le SIAL	Contrôle Continu	
UE 99	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 vocaux	2.5 h	Contrôle Continu	
	Indexation sémantique et recherche d'information	2.5 h	Contrôle Continu	
	Sciences affectives	2.5 h	Contrôle Continu	
UE 3	Traitement automatique			5
	Plates-formes logicielles pour le 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	
	Droit et responsabilité	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
	Logique et langues naturelles	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	
	Cognition, sentiment, opinion	2.5 h	Contrôle Continu	
UE 3	Sémantique			6
	Structures de connaissances et raisonnement	2.5 h	Contrôle Continu	
	Terminologie et stylométrie	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		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 :

Ce cours s'intéresse à la notion d'arbitraire linguistique. L'arbitraire dans la langue est souvent opposé à la motivation. Dans ce cours, on proposera l'idée que, loin d'être le contraire de la motivation, l'arbitraire est la condition préalable à partir de laquelle le code linguistique est libre de reproduire des structures indépendantes (phénomènes de motivation) ou bien de tracer des distinctions en l'absence de toute structure préalable (phénomènes de non-motivation). Dans ce cours, on examinera des manifestations de l'arbitraire à plusieurs niveaux de la grammaire française : - phono-symbolisme et structure de la syllabe ; - opposition modale indicatif/subjonctif ; - distinction massif/comptable et polysémie ; - rapport entre fonctions grammaticales et rôles sémantiques.

Bibliographie :

M. PRANDI, The building blocks of meaning, Amsterdam-Philadelphia, 2004.
M. PRANDI, Conceptual Conflicts in Metaphors and Figurative Language, Routledge, New York – Londres, 2017. D'autres indications bibliographiques seront fournies au fil du semestre.

M1LF15FA Terminologie linguistique

Évaluation :
cf Langue Française

Volume horaire :
1h par semaine

Descriptif :

Introduction générale à la terminologie linguistique et au fonctionnement du discours linguistique. Présentation des problématiques. Présentation de la conception normative et du fonctionnement de la terminologie linguistique.

Bibliographie :

S. AUROUX, La raison, le langage et les normes, Paris, PUF, 1998. • J.-C. MILNER, 1989, Introduction à une science du langage, Paris, Éditions du Seuil
F. NEVEU (dir.), Syntaxe & Sémantique, n°7 « La terminologie linguistique : problèmes épistémologiques, conceptuels et traductionnels », Caen, Presses Universitaires de Caen, 2006.
F. NEVEU, Dictionnaire des sciences du langage, Paris, A. Colin, 2011.

M1SOL012 Conférences de l'équipe STIH

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 jeudi de 18h30 à 20h30. Le programme sera donné à la rentrée.

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'aco-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

UE 3 : Ingénierie des connaissances (5 ects)

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 récursifs 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 :

Le but de ce module est d'introduire les principaux logiciels pour le TAL puis de se focaliser sur la plateforme GATE. Après avoir présenté l'architecture de GATE, le cours approfondira l'environnement de programmation graphique suivi de l'API permettant d'étendre ses fonctionnalités. 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.

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.

Linguistique théorique : Voir avec l'UFR de Langue Française

M2LF13FA Linguistique Théorique appliquée

Évaluation :
cf Langue Française

Volume horaire :
2h par semaine

Descriptif :

Ce cours s'intéresse aux relations conceptuelles comme cause, finalité, temporalité, etc. En philosophie, elles sont généralement ramenées à des notions abstraites. En linguistique, elles sont souvent réduites à la sémantique d'un ensemble de conjonctions de subordination. Ce cours propose une troisième voie. D'une part, on identifiera des structures conceptuelles indépendantes de la langue ; de l'autre, on étudiera les stratégies textuelles pour exprimer et manipuler ces structures. Après un examen des principales grammaires françaises, ce cours se concentrera sur la cause, la finalité, la condition et la concession. Par là nous serons amenés à discuter la définition de subordination et la classification des locutions prépositionnelles.

Bibliographie :

G. GROSS & alii, Sémantique de la cause, Paris, Peeters, 2009.

G. GROSS & M. Prandi, La finalité – Fondements conceptuels et genèse linguistique, Bruxelles, de boeck.duculot, 2004.

M. PRANDI, The building blocks of meaning, Amsterdam-Philadelphia, 2004

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 étudiés.

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)

M2SOL023

Dictionnaire et néologismes

Évaluation :

Contrôle mixte

Volume horaire :

1h CM + 1,5h TD

Descriptif :

On peut considérer que les dictionnaires sont des photographies de la langue (des Daguerrotypes pour Pierre Larousse), des représentations de l'état du stock lexical à un moment donné de l'histoire d'une langue. Ceci suppose d'effectuer des choix en mettant à jour ce stock en faisant sortir du dictionnaire des termes, par exemple, inusités ou au contraire en faisant rentrer des termes nouveaux, des néologismes, ou redécouverts. Ceci a une implication importante sur les ressources produites pour et utilisées par le Traitement Automatique des Langues : la langue est vivante. Pour Pierre Larousse: "Un dictionnaire [...] ne doit ni suivre de trop loin ni ouvrir la marche : c'est un laquais qui porte les bagages de [la langue] en le suivant par derrière".

Mais la nouveauté ce ne sont pas seulement les formes, les mots, mais aussi les sens. Des sens nouveaux apparaissent, des sens anciens disparaissent dans un processus continue d'adaptation du lexique aux besoins des locuteurs.

Bibliographie :

GUILBERT L. « La créativité lexicale », Paris, Larousse, 1975

SABLAYROLLES J.F. «Alain Rey, le Linguiste et les Mots Nouveaux », Lambert-Lucas, pp.87-98, 2011, La Lexicothèque.

M2SOL034

Corpus, ressources

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>

Master 2 - Semestre 3

UE 1 : Méthodologie et épistémologie (7 ects)

Sens et texte : voir à l'UFR de Langue Française

Linguistique théorique : voir à l'UFR de Langue Française

Terminologie linguistique : voir à l'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 articulateur 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 articulateur, 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.

M3SOL033

Dialogues et serveurs vocaux

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours a pour but la conception et mise au point d'un dialogue oral finalisé et son déploiement sur un serveur vocal interactif. Modèle de tâche (GOMS, CTT) et modèle de dialogue (Réseaux de Petri à objets, ...), Architecture de dialogue (ARCH, AMF) et processus de conception et d'adaptation (Plastic ML), Structures de dialogues et flux de commande en Voice XML, Ressources, entrées utilisateur et sorties du système en Voice XML, Mise en œuvre de solutions

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

M3SOL034

Indexation sémantique et recherche d'information

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

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.

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.

M3SOL041

Sciences affectives

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)

M3SOL021 Plate-formes logicielles

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Le cours est composé de deux parties distinctes et complémentaires. La première approfondira l'utilisation d'Unitex, en particulier les cascades de transducteurs pour l'annotation. La deuxième se centre sur l'utilisation avancée de la plateforme GATE pour l'analyse et l'annotation sémantique de textes. Des notions d'inférence ontologique pour l'extraction de nouvelles connaissances seront présentées.

Bibliographie :

Cascades de transducteurs avec Unitex : http://tln.li.univ-tours.fr/Tln_CasEN.html.

M3SOL024 Traduction automatique

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Bibliographie :

.

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.

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>

Master 2 - Semestre 4

UE 1 : Méthodologie et épistémologie (4 ects)

M4SOL120 Logique et langues naturelles

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Dans un premier temps, nous rappelons les formalisations en logique du 1er ordre des énoncés de la langue naturelle fondées sur la théorie vériconditionnelle de la signification, et leurs extensions dans les langages simplement typés d'ordre supérieur.

La présentation de la théorie inférentielle de la signification, nous conduira vers la mise en évidence de la correspondance de Curry-Howard, et sa mise en œuvre dans la théorie intuitionniste des types.

Ce dernier cadre formel sera utilisé pour résoudre les problèmes classiques de la compositionnalité, des contextes obliques, et plus généralement de l'intersubstituabilité des expressions coréférentielles.

Bibliographie :

P. Delsarte, A. Thayse, Logique pour le traitement de la langue naturelle, Hermès science, 2001.

J-P. Desclés, B. Djioua, F. Le Priol, Logique et langage, Hermann, 2010.

D. R. Dowty, Introduction to Montague Semantics, Kluwer Academic Pub., 1980.

P. Martin-Löf, Intuitionistic type theory, Bibliopolis, 1984.

A. Ranta, Type-theoretical grammar, Oxford: Clarendon, 1994.

G. Sundholm, "Proof theory and meaning", In D. Gabbay, F. Guenther, editors, Handbook of Philosophical Logic, Volume III, Alternatives to Classical Logic, D. Reidel, 1986.

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.

M4SOL203

Cognition, sentiment, opinion

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours porte sur la fouille d'opinion et l'analyse des sentiments et des émotions extraites automatiquement dans des textes. L'analyse de sentiments, aussi appelé opinion mining, est le processus qui permet de déterminer la tonalité émotionnelle d'un texte. Cette analyse est utilisée pour mieux comprendre la perception, les opinions et les jugements des utilisateurs. Les applications et outils de traitement et visualisation existants seront présentés, ainsi que le traitement des microtextes (ex. Twitter, Sina Weibo) provenant des réseaux sociaux

Bibliographie :

La bibliographie sera donnée au fil de ce cours.

UE 3 : Sémantique (6 ects)

M4SOL301

Structures connaissances

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

Ce cours est la suite des 3 cours d'ingénierie des connaissances (Représentation et traitement des documents électroniques, Représentation et modélisation des connaissances, Web sémantique et big-data) enseignés en première année de master. Le Web sémantique se veut un Web dont le contenu peut être appréhendé et exploité par des machines pour fournir des services plus intelligents à ses utilisateurs. Il propose une infrastructure qui complète le contenu du Web avec des connaissances formellement structurées. Ce cours à caractère applicatif vise justement à montrer comment exploiter les connaissances ainsi que leurs structures afin de développer des services intelligents. Pour mettre en pratique les connaissances acquises, les étudiantes et les étudiants développeront en groupes des applications utilisant les technologies du Web sémantique. Plusieurs pistes et plusieurs domaines d'applications seront proposés. Par exemple, un moteur de recherche sémantique à partir des textes non structurés ou semi-structurés, un système de calcul de similarité entre des textes écrits dans des langues différentes, un système de recommandation dans des domaines tels que les articles de presse, la culture et le tourisme, une application du type « Web des objets » manipulant des données recueillies des différents capteurs (ex. pollution de l'air, bruit, pression atmosphérique, température, luminosité, humidité, vent) et effectuant des raisonnements à base de règles.

Bibliographie :

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

Staab, S., & Studer, R. (Eds.). (2010). Handbook on ontologies. Springer Science & Business Media.

Gandon, F., Corby, O., & Faron-Zucker, C. (2012). Le web sémantique: Comment lier les données et les schémas sur le web?. Dunod.

Barnaghi, P., Wang, W., Henson, C., & Taylor, K. (2012). Semantics for the Internet of Things: early progress and back to the future. International Journal on Semantic Web and

Information Systems (IJSWIS), 8(1), 1-21.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In Recommender systems handbook (pp. 1-34). Springer, Boston, MA.

Bast, H., Buchhold, B., & Hausmann, E. (2016). Semantic search on text and knowledge bases. Foundations and Trends® in Information Retrieval, 10(2-3), 119-271.

M4SOL302

Terminologie et Stylométrie

Évaluation :
CCI

Volume horaire :
2,5h CM

Descriptif :

La terminologie est la discipline qui étudie la formation et l'utilisation du vocabulaire dans les domaines spécialisés. La terminologie c'est aussi l'ensemble des termes d'un domaine. Un terme est, dans un contexte approprié d'énonciation, mono-sémique par définition. Ceci facilite la compréhension entre les experts d'un domaine mais rend le sens beaucoup moins interprétable par les locuteurs ne connaissant pas le domaine.

La stylométrie elle s'intéresse à l'impact du style dans les textes. On peut étudier le style individuel, les spécificités du style d'un auteur, ou bien le style collectif, les spécificités d'un genre textuel.

Nous présenterons ces deux domaines d'application du Traitement Automatique des Langues, assez distants de prime abord, puis nous montrerons comment deux domaines peuvent être complémentaires : comment l'étude stylistique peut aider l'extraction terminologique et inversement.

Bibliographie :

DAILLE B. «Term variation in specialised corpora characterisation, automatic discovery and applications.», Terminology and Lexicography Research and Practice, 19 2017. John Benjamins

DAILLE B., JACQUEY E., LEJEUNE G., MELO L.F., TOUSSAINT Y. « Ambiguity Diagnosis for Terms in Digital Humanities », Language and Resources Conference (LREC) 2016

LONGREE D., LUONG X., MELLET S., « Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé », Lexicometrica 2004

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

SECRETARIAT

- Madame Anastazija Dzerdz-Gavrilov
- 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 10h à 12h et de 14h à 16h,
- Fermeture le Lundi et Vendredi après-midi

BIBLIOTHEQUE

- Maison de la Recherche bibliothèque Serpente

SCUIOIP

- Tél. 01 40 46 26 14
- Courriel : lettres-scuioip@sorbonne-universite.fr