Journée d'études organisée par LIDILE EA 3874
(Axes : DiLeM, TRASILT, ELIA)

Langue 2.0 : Recherche, développement et exploitation du numérique en linguistique appliquée

Langue 2.0 :
Recherche, développement et exploitation du numérique en linguistique appliquée

Descriptif de la journée d'études

Dans ce XXI^{ème} siècle, que l'on peut qualifier comme étant l'ère du *numérique* ou du *digital*[1] selon l'emploi que l'on veut attribuer à ces mots, la communication et l'activité professionnelle, tant au niveau public que privé, s'effectuent par et grâce au numérique. Presque tous les secteurs d'activité - social, politique, économique et éducatif - se développent au rythme des innovations qui jalonnent les évolutions et les usages du numérique.
Tous ces domaines d'activité humaine ont comme dénominateur commun la présence massive de données langagières. En effet, en tant que véhicules centraux de la communication, les productions langagières (qu'elles soient écrites ou parlées) sont les principales porteuses d'information au sein de nos sociétés. Dès lors que ces données langagières se présentent sous une forme numérique ou numérisable et dans des quantités importantes, leurs possibilités d'exploitation sont sans limites.
L'unité de recherche LIDILE EA 3874 (Linguistique, Ingénierie et Didactique des Langues) souhaite s'interroger, par le biais de cette journée d'étude, sur le potentiel du numérique selon trois axes d'investigation : la linguistique, l'ingénierie et la didactique des langues.
Cette journée d'étude se propose de recenser certaines avancées, projets et réalisations dans le contexte actuel.
On s'interroge plus précisément sur :
- le rapport entre les systèmes et/ou mécanismes numériques mis en place et le fonctionnement de la langue dans un environnement qui comprend les mondes virtuels, le traitement automatique des langues et de la parole ;
- l'exploitation des données numériques en linguistique pour un usage professionnel et didactique ;
- l'utilisation des outils numériques dans l'enseignement-apprentissage de la langue maternelle et/ou étrangère ;
- les projets de recherches collaboratifs et innovants du numérique en rapport avec l'étude de la langue maternelle et/ou étrangère.

Chercheurs, enseignants-chercheurs, doctorants, étudiants et praticiens de langues diverses interviendront dans cette journée d'étude qui vise à intéresser tout professionnel dans le domaine des langues.

---

1 Il est à noter que les termes *digital* et *numérique* n'ont pas encore reçu une distinction nette quant à leur signification auprès des usagers. Marlène Duretz (« Dilemme numérique », Le Monde, le 14 janvier 2014. <https://www.lemonde.fr/economie/article/2014/01/14/dilemme-numerique_4347680_3234.html>) et Fabian Ropars (« Faut-il dire numérique ou digital ? », BDM/media, culture web, le 11 février 2015 à 10h49, mis à jour le 28 juin 2018 à 11h29, <https://www.blogdumoderateur.com/numerique-ou-digital/>) affirment, par exemple, que l'on attribue la catégorie de substantif ou d'adjectif, selon le secteur professionnel et l'usage auquel les termes *numérique* et *digital* correspondent: on dira plutôt *industrie numérique* que *industrie digitale* et *digitalisation d'une marque* plutôt que *numérisation d'une marque*. Fabian Ropars précise que « Numérique tend à renvoyer de fait au technologique, à la dimension discrète de la technologie, celle que manipulent les ingénieurs et qui reste intangible. Digital semblerait concerner plutôt l'usager dans son expérience de cette technologie numérique ».

Responsables scientifiques : Aura Luz Duffé Montalván et Christine Evain.

Comité d'organisation :

Marie Baize-Varin
Marie-Françoise Bourvon
Marie Chatelain
Nicole Cloarec
Griselda Drouet
Aura Luz Duffé Montalván
Octavia Efraim
Christine Evain
Emmeline Isaïa

Conférences plénières
Amphi M Bât 42

- **Thierry Morineau**, Université Bretagne Sud, Laboratoire de Psychologie, Cognition, Communication, Comportement (LP3C EA 1285)

Thierry Morineau est Professeur de Psychologie cognitive et Ergonomie à l'Université Bretagne Sud et membre du laboratoire LP3C. Ses travaux actuels concernent l'ingénierie cognitive des systèmes sociotechniques, avec comme objet d'étude principal, l'activité des soignants en contexte d'urgence médicale.

« La digitalisation des environnements de travail et ses conséquences sur l'individu »

Résumé :
L'impact de la digitalisation des environnements de travail peut être appréhendé à travers la distinction classique, proposée par F. De Saussure en linguistique, entre Référent, Signifié et Signifiant. Dans les environnements faiblement technologiques, l'interaction entre l'individu et son environnement est fondée sur un continuum entre ces trois éléments composant la signification. Ce continuum vécu physiquement et mentalement dans l'action fournit une information directe à l'individu sur l'état de l'environnement. Avec le développement technologique, une mise à distance toujours plus grande de l'environnement de travail génère des effets significatifs, tant sur le plan du fonctionnement cognitif que sur l'état psychologique global de l'individu.

- **Émilie Née**, Université Paris-Est Créteil, Centre d'étude des discours, images, textes, écrits, communication (Céditec EA3119)

Émilie Née s'intéresse à l'écrit et à l'écriture dans une perspective discursive et phraséologique : elle travaille ainsi sur différents genres d'écrits (rapports, écrits de communication, par exemple) émanant de sphères professionnelles ou de sphères académiques pour en faire émerger des régularités, des normes, des routines discursives. Dans ses travaux portant sur la répétition ou sur la phraséologie, Émilie Née convoque de manière privilégiée les méthodes de la statistique textuelle (ou lexicométrie / textométrie) et de la linguistique de corpus, pour mettre au jour des phénomènes discursifs.

« Des données et des outils numériques pour quoi faire ? Le cas de l'analyse du discours »

Résumé :
Quel intérêt présente l'outil informatique pour l'analyse des discours ? Quels sont les apports des études en analyse du discours qui ont mobilisé l'informatique, du point de vue des observables, des méthodes ou encore de l'herméneutique des textes et des discours ?
C'est à ces questions que nous proposons de répondre dans notre intervention. Nous aborderons aussi la question des humanités numériques, qui orientent l'analyse du discours vers de nouvelles lectures du texte et vers la mobilisation de ressources numériques à côté des corpus traditionnels. Un premier temps de l'intervention reviendra sur les raisons épistémologiques et politiques qui expliquent la rencontre précoce entre informatique et

analyse du discours. Nous présenterons ensuite un panorama des recherches francophones en analyse du discours outillée, du côté des sciences du langage, mais aussi de l'histoire, de la sociologie, des sciences de l'information et de la communication, et de la science politique en mettant l'accent sur :

- les points d'entrée dans le discours qu'autorisent les approches informatisées ;
- les types et genres de discours analysés (politiques, médiatiques…) ;
- les méthodes privilégiées par ces recherches ;
- les résultats produits.

Un troisième temps de l'intervention portera sur la question des corpus en analyse du discours outillée. D'un côté nous montrerons comment les corpus ont pu évoluer depuis les débuts de cette approche, passant de corpus « réservoirs » à des ressources textuelles structurées. De l'autre, nous proposerons une typologie des corpus que l'on mobilise généralement en fonction des dynamiques que l'on souhaite mettre en évidence. Ce troisième temps de l'intervention sera l'occasion de rappeler un certain nombre de principes méthodologiques et de montrer quelques illustrations d'analyses multidimensionnelles.

Nous clôturerons cette intervention sur les pratiques interprétatives en analyse du discours outillée et sur les limites d'une approche informatisée des discours.

Communications
Amphi M Bât 42

- **Marie Chatelain**, PRAG d'espagnol, Écoles de Saint-Cyr Coëtquidan, Guer (56), Ministère des Armées, LIDILE EA 3874

Formatrice en pédagogies numériques, pratique la classe inversée depuis 4 ans, chercheuse associée à LIDILE

« Changement de posture dans le supérieur : enseigner l'espagnol en s'appuyant sur le numérique et les pédagogies actives »

Résumé :
Les recherches récentes en pédagogie (Lebrun 2015) ont démontré que, plus l'élève était impliqué et actif dans son apprentissage, plus l'acquisition des compétences cognitives se révélaient être efficaces. D'autre part, Michel Serres l'évoquait dans son ouvrage *Petite Poucette* (2012) : la société évolue et le rapport au savoir n'est plus le même : il est à portée de main. L'enseignement actuel se doit de répondre aux exigences contemporaines en travaillant des savoirs-faires permettant l'acquisition de compétences transverses.
Aux Écoles de St Cyr-Coëtquidan (Écoles visant à former les futurs officiers de l'Armée de Terre), le défi, en espagnol, de proposer une formation de qualité aux élèves militaires de St Cyr est double : s'adapter aux contraintes particulières de leur formation et réussir à maintenir éveillés leur motivation et leur intérêt à travailler.
Ce postulat nous a amené à diversifier nos méthodes d'enseignement en s'inspirant très largement des pédagogies actives. L'apprenant peut ainsi travailler la langue à travers des « tâches complexes » pour lesquelles il devra manier différentes compétences : travail de groupe, collecte d'informations, autonomie. Les savoirs sont co-construits (élève – élève et élève – professeur).
Nous présentons une séquence travaillée en autonomie pour montrer de façon concrète l'utilisation de pédagogies actives. Nous en ferons le bilan après un feed-back des élèves et présenterons l'évolution de l'enseignement de l'espagnol à travers l'année.

- **Emmeline Isaïa**, Doctorante, Université Rennes 2, membre de LIDILE EA 3874.

Emmeline Isaïa est actuellement en deuxième année de Doctorat en didactique des langues, au sein de LIDILE. Sa recherche doctorale, réalisée sous la direction d'Aura Luz Duffe Montalván et de Thierry Morineau, s'intègre dans le domaine de la didactique des langues et dans l'ergonomie numérique IHM. Il s'agit de créer un outil numérique facilitant l'acquisition du vocabulaire par les apprenants en langue espagnole. Après un passage par le monde de l'entreprise, en tant que chargée de communication et responsable des projets web, Emmeline a souhaité mettre ses compétences numériques au service de l'enseignement et de la recherche. En parallèle de son travail de recherche, elle enseigne l'espagnol au collège Saint Jean Lasalle de Guidel et est professeur-documentaliste en collège Les Saints Anges de Pontivy.

« L'ontologie sous forme numérique : une aide pour l'acquisition du lexique en langue étrangère ? Le cas de la langue espagnole »

Résumé :
L'objectif de cette communication est de présenter le début de notre travail de recherche. Notre thèse de départ est que l'acquisition lexicale d'une langue peut être facilitée grâce à un système ontologique numérique qui s'appuie notamment sur des images. Cette étude s'intègre dans le champ de la didactique de la langue espagnole auprès d'un public francophone. Elle a pour objectif final de créer un outil numérique pour faciliter la mémorisation du lexique d'une langue étrangère.
L'hypothèse de départ s'appuie, sur deux idées fortes : l'existence d'un lexique mental, selon laquelle les mots sont ancrés dans la mémoire grâce à leurs relations avec d'autres mots (Van der Linden 2006); ainsi que sur la théorie selon laquelle, l'apprentissage des mots en contexte est largement supérieur à un apprentissage des mots sans contexte (Mondiria 1996). Il s'agit donc de s'appuyer sur les recherches en sciences du langage et en didactiques des langues.
Cependant, la création de l'ontologie doit aussi prendre appui sur la recherche en neuropsychologie. Pour faciliter l'acquisition du lexique à travers une ontologie, il faut prendre en compte les processus de mémorisation ainsi que le contexte d'utilisation de l'outil numérique. En effet, Neisser (1997) explique que l'acte de se souvenir et donc le processus de mémorisation se pratique dans un contexte particulier, dans un temps particulier et dans une place particulière.
La création d'un tel outil, doit donc prendre en compte l'utilisateur, ses motivations et le contexte. Il doit également être interactif pour favoriser la mémoire de l'action. L'apprenant qui utilisera l'outil devra le faire avec l'objectif d'enrichissement de son vocabulaire. Pour ces raisons, l'interface numérique devra avoir une approche écologique et s'intégrer parfaitement à la démarche de l'utilisateur.
Dans un premier temps, il s'agira donc de rappeler la définition du terme ontologie, puis d'expliquer l'intérêt d'une ontologie sous forme numérique dans le cas de l'apprentissage du lexique. Enfin, une présentation de la méthodologie de recherche ainsi qu'une maquette de l'outil à créer, sera réalisée.

- **Griselda Drouet** et **Élisabeth Richard** (LIDILE EA 3874), **Laurent Rousvoal** et **Sandrine Turgis** (IODE UMR 6262, Rennes 1), **Sébastien Motta** (Université de Nantes), **Vincent Etien** (Irisa, Rennes 1).

Griselda Drouet est maître de conférences en linguistique française à l'Université Rennes 2. Elle mène ses recherches en analyse de discours au sein de l'équipe LIDILE et s'intéresse en particulier à l'étude de la langue en contexte, à l'organisation du discours oral et aux marqueurs discursifs.

Élisabeth Richard est professeur des universités en Linguistique et Didactique des langues, directrice de l'équipe LIDILE. Ses travaux de recherche en linguistique sont axés sur le français parlé, l'énonciation et la reformulation. Elle est également spécialiste de Didactique du français langue étrangère.

Laurent Rousvoal est maître de conférences en droit privé et sciences criminelles à l'université Rennes 1. Il fait partie de l'UMR CNRS (6262) IODE. Il est responsable scientifique du projet LaNoPale (La notion de matière pénale. Analyse interdisciplinaire d'un objet de droit et de science du droit, séminaire de recherche porté par la MSHB).

Sandrine Turgis est maître de conférences en droit public à l'université Rennes 1, elle fait également partie du laboratoire IODE. Elle est spécialiste de droit européen et international des droits de l'homme.

Sébastien Motta est maître de conférences en philosophie à l'université de Nantes, il est membre de l'équipe CAPHI (EA2163). Ses thèmes de recherches sont axés autour de la philosophie de la logique, du langage et de l'esprit, et la métaphysique.

Vincent Etien est étudiant à l'université Rennes 1, il est actuellement en Master à l'IRISA. Il a été stagiaire sur le projet de cartographie des renvois dans le corpus étudié par l'équipe LaNoPale.

« Des données et des outils numériques : comment faire ? Un exemple de visualisation d'arrêts de la Cour Européenne des Droits de l'Homme »

Résumé :
Le projet LaNoPale, sous la direction de Laurent Rousvoal et porté par la MSHB, est un séminaire de recherche interdisciplinaire qui s'est attaché à définir la notion de « matière pénale » à travers un corpus de 70 000 mots, constitué d'une sélection d'arrêts de la Cour Européenne des Droits de l'Homme, allant de 1976 à 2014. L'approche interdisciplinaire de cette étude propose un éclairage inédit de ces textes et permet d'en affiner la compréhension et l'évaluation. À cette fin, l'analyse linguistique a pu relever un certain nombre de phénomènes discursifs récurrents et mettre en évidence l'importance des renvois de références entre les textes dans la construction-même des arrêts. Les différents arrêts sont le plus souvent rendus en s'appuyant sur un jeu de renvois intertextuels et extratextuels qui semblent construire arrêts après arrêts, années après années, le fil rouge de la légitimité du discours de la Cour et de son recours à la notion de "matière pénale". Ce tissage référentiel complexe nécessitait, au-delà d'un relevé précis, une mise en visualisation cartographique active afin de mieux en apprécier l'ampleur et l'importance. Cette communication vise à présenter les enjeux juridiques, linguistiques et techniques de la réalisation d'une telle visualisation numérique et viendra approfondir le lien entre linguistique appliquée et numérique dans le cadre spécifique de la recherche interdisciplinaire.

- **Octavia Efraim**, Doctorante à LIDILE EA 3874

Octavia Efraim est doctorante en traitement automatique des langues au sein de LIDILE. Ses travaux portent sur l'équivalence sémantique textuelle, avec application à la relation client. Dans ses travaux antérieurs, Octavia s'est intéressée notamment aux outils automatiques au service de la traduction. Elle enseigne également l'informatique à l'université Rennes 2.

« L'apprentissage automatique à la rescousse de la relation client : vers un modèle de compréhension polyvalente de messages clients »

Résumé :
Dans le cadre de la gestion des relations avec les clients (ou, plus couramment, gestion de la relation client), le traitement de demandes formulées librement par les clients joue un rôle de plus en plus important. Les possibilités toujours plus poussées de traitement automatisé de ces données non-structurées que procurent les avancées constantes dans le domaine du traitement automatique du langage naturel, ont transformé radicalement la relation client.

Un système qui met en relation l'usager et le service client de l'entreprise doit répondre aux besoins de deux utilisateurs : le client et le service client. Ces deux besoins sont différents : alors que le client souhaite recevoir une réponse à sa demande, l'objectif du service client est de choisir l'action la plus pertinente qui permette de traiter au mieux la demande du client selon la stratégie de l'entreprise.
Dans cette communication nous proposerons des éléments pour la conception d'un tel système, en nous plaçant du côté du service client. Son choix d'action à mettre en œuvre passe par une compréhension du message du client. Nous nous proposons d'aborder une approche conjointe de deux tâches qui sont traditionnellement traitées séparément : la détection, dans un message formulé par un usager, de l'intention et de l'émotion de ce dernier. Pour cela nous nous inscrivons dans une perspective multitâche, où l'on vise à résoudre des problèmes différents par le biais d'un modèle unique. Nous définirons également une notion d'équivalence fonctionnelle, qui nous permettra d'aborder la problématique de la similarité sémantique de messages d'usagers sous un angle original.

Atelier
Salle informatique B27110

- **Thomas Gaillat**, LIDILE EA 3874, Université de Rennes 2

Thomas Gaillat, en poste à l'Université de Rennes 2, est enseignant-chercheur en linguistique. Il enseigne l'anglais de spécialité à l'université depuis 1999 ; il est membre de LIDILE. Sa thèse soutenue en 2016 à l'Université Paris-Diderot porte sur l'interopérabilité des corpus d'apprenants en anglais et la modélisation des formes référentielles this, that *et* it *chez les apprenants. Ses travaux portent sur des problématiques de linguistique anglaise au croisement du Traitement Automatique des Langues et de la linguistique de corpus. Son premier axe de recherche se place dans une perspective acquisitionnelle d'une langue étrangère avec l'exploitation de méthodes quantitatives pour l'analyse automatique de la langue. Son deuxième axe porte sur l'analyse automatique de sentiments par apprentissage machine.*

« Des corpus annotés aux jeux de données : un enjeu de la modélisation de la langue d'apprenants »

Résumé :
Les corpus de langues sont une source de données empiriques exploitées pour l'étude de la langue (Wynne 2005). Leur construction répond à des critères précis garantissant la représentativité de la langue étudiée. Une multitude de corpus existe aujourd'hui et permet l'analyse de différentes variétés de langue dans une multitude de domaines. Leur structuration inclut, dans de nombreux cas, des annotations en fonction de schémas reflétant les objectifs des études à mener (Leech, 2005), *a fortiori* pour les corpus d'apprenants (Gilquin 2015 ; Tono 2003). Une fois constitués, la question de leur exploitation se pose. Les textes de corpus annotés permettent des requêtes textuelles par motifs pouvant combiner différents critères de recherche (mots, PoS et autres). Les résultats obtenus sont présentés en termes de fréquences des motifs recherchés. Ces approches apportent des réponses sur les usages quantitatifs bruts. Cependant elles ne permettent pas d'aborder ces usages dans un cadre multifactoriel. On compte les formes répondant à des critères, mais on ne mesure pas l'influence des critères sur les formes.
Cet atelier propose d'aborder la question à partir de la constitution de jeux de données. Les jeux de données peuvent être des représentations multi-variées des corpus. Ces structures de données permettent de mettre en regard les motifs recherchés avec les caractéristiques de leurs contextes (co-texte et annotations). On peut alors utiliser des méthodes probabilistes (Baayen 2008 ; Gries 2010) pour mesurer l'effet produit par un groupe de variables sur l'occurrence d'un phénomène observable.
À partir de corpus, nous procéderons à l'échantillonnage de données textuelles annotées pour les placer dans des matrices. Nous montrerons comment les choix opérés dépendent des questions de recherche. Sous R (R Core Team 2012), les participants mettront en œuvre des procédures d'annotation et d'extraction avec Quanteda (Benoit et al. 2018) pour constituer un jeu de données type.

Références:

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Gries, S. Th. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Frankfurt am Main: Peter Lang.

Leech, G. (2005). Adding Linguistic Annotion. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 17–29). Retrieved from http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

R Core Team. (2012). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

Tono, Y. (2003). Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson, & A. M. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference UCREL Technical paper* (pp. 800–809). Lancaster, UK: Lancaster University.

Wynne, M. (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Retrieved from http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm