# CPER LCHN
# Langues, Connaissances et Humanités Numériques

———————

## Bilan 2018

Ce document présente un bilan des activités de recherche qui ont été menées dans les laboratoires de l'Université de Lorraine et qui sont liées aux financements obtenus dans le cadre du projet CPER LCHN 2015—2020.

Ce bilan ne prétend pas être exhaustif, car il est souvent difficile de faire un lien direct entre des financements obtenus à un instant donné et les retombées potentielles qui peuvent intervenir plusieurs mois ou plusieurs années après. Nous considérons cependant qu'il est assez représentatif de la palette des activités que ce CPER a générées jusqu'ici.

L'essentiel du texte de ce bilan a été produit par nos collègues impliqués dans différents projets. Nous les remercions pour leur aide dans l'élaboration de ce rapport et nous nous excusons auprès du lecteur si la formulation des diverses sections n'est pas homogène.

## Contexte, présentation générale de l'opération

Au sein de la thématique Sciences du numérique, le projet Langues, Connaissances et Humanités Numériques (LCHN), complémentaire du projet Cyber-Entreprise, a pour objectif de conforter la Lorraine dans les domaines de la gestion et de l'accès aux contenus numériques, dont la plus grande partie demeure sous forme langagière. Il propose de mettre en place des plateformes d'expérimentation scientifique pour conforter les coopérations entre acteurs lorrains qui ont montré au cours des dernières années leur capacité à travailler ensemble que ce soit lors du précédent CPER (Projet « Traitement Automatique des Langues et des Connaissances » du CPER « Modélisation, Information et Simulation Numérique » et « Langues, Textes et Documents » du PRST « Homme et Société ») ou dans le cadre de projets ANR, permettant ainsi à la Lorraine d'acquérir une visibilité marquée au travers de plateformes nationales de diffusion de ressources dans le cadre des PIA : Equipex ORTOLANG, pour la langue et les ressources langagières, et Idex national ISTEX, pour des ressources en Information scientifique et technique (IST).

Ce sous-programme est par essence même fortement pluridisciplinaire (Informatiques et Sciences Humaines et Sociales) et réunit des compétences diverses sur les aspects ingénierie des langues (informatique et linguistique), extraction et structuration de connaissance (informatique, IST, linguistique), humanités numériques (linguistiques, information et communication, histoire, philosophie, littérature, psychologie, sociologie et informatique) et E-éducation (informatique, information et communication, linguistique, sciences de l'éducation, psychologie).
Ce projet se veut aussi contribuer à l'axe Ingénierie des langues et de la connaissance du projet I-Site Lorraine Université d'excellence.

## Objectifs recherchés

Dans le cadre du projet LCHN, nous proposons de structurer quatre plateformes matérielles et logicielles complémentaires et fortement interconnectées :

- Une plateforme d'expérimentation en Ingénierie des langues,
- Une plateforme d'expérimentation en Extraction et structuration de connaissances,

- Une plateforme d'expérimentation en Humanités Numériques,
- Une plateforme d'expérimentation en E-Éducation,

dont trois s'appuyant sur des matériels spécifiques pour, entre autres, permettre le traitement de corpus de grand volume.

Ces plateformes serviront de soutien au développement d'actions scientifiques avec comme objectif de conforter ou mieux positionner la Lorraine au plan national et international dans les quatre domaines cités ci-dessus qui nous apparaissent de plus en plus incontournables dans les domaines de la gestion, de l'accès et de l'exploitation des contenus numériques. En particulier, en cohérence avec le projet I-Site Lorraine Université d'excellence, ces plateformes serviront de support d'expérimentation pour, cf. dossier I-Site de l'Université de Lorraine, *développer le traitement automatique des langues, l'extraction et le traitement des connaissances, la consolidation de ressources lexicales et textuelles, la veille et l'intelligence économique.*

# Bilan des plateformes matérielles

## 1. Cluster GPU pour l'apprentissage profond

Les GPU installés dans Grid5000 ont été utilisés pour tous les travaux liés au *deep-learning* du laboratoire LORIA. Ceci inclut en particulier tous les projets des équipes Multispeech et Synalp. Le cluster GPU a été également utilisé à l'ATILF. Les recherches ont porté sur :

- reconnaissance automatique et alignement de la parole ;
- séparation de sources sonores et analyses de scènes auditives ;
- génération automatique de résumés ;
- analyse des actes de dialogue et analyse des sentiments ;
- analyse syntaxique de textes.

Même s'ils sont utilisés en priorité pour les travaux liés au CPER, les GPU sont aussi ponctuellement utilisés dans d'autres projets de l'équipe Capsid du LORIA, en particulier ceux nécessitant des simulations numériques dans le domaine du calcul des structures moléculaires ou de l'équipe Biscuit du LORIA, sur l'apprentissage par renforcement.

Les GPU sont utilisés à environ 80% dès leur installation. Ils sont donc devenus des outils indispensables aux travaux de toutes les équipes qui travaillent en *deep-learning* et en simulation numérique.

Publications récentes liées à la plateforme :

- Badr Abdullah, Irina Illina, Dominique Fohr. **Dynamic Extension of ASR Lexicon Using Wikipedia Data**. *IEEE Workshop on Spoken and Language Technology (SLT)*, Dec 2018, Athènes, Greece. 2018, Proceedings of IEEE SLT.
- Matthieu Zimmer, Yann Boniface, Alain Dutech. **Developmental Reinforcement Learning through Sensorimotor Space Enlargement**. *ICDL-EPIROB 2018 - 8th joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, Sep 2018, Tokyo, Japan. pp.1-6
- Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr. **Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment.** *Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association*, Sep 2018, Hyderabad, India. 2018.
- Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, Hoa Le. **Multi-task dialog act and sentiment recognition on Mastodon.** *COLING*, Aug 2018, Santa Fe, United States.
- Lauréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin. **CRNN-based joint**

**azimuth and elevation localization with the Ambisonics intensity vector.** *IWAENC 2018 - 16th International Workshop on Acoustic Signal Enhancement*, Sep 2018, Tokyo, Japan.

- Michel Vacher, Emmanuel Vincent, Marc-Eric Bobillier Chaumon, Thierry Joubert, François Portet, et al. **The VocADom Project: Speech Interaction for Well-being and Reliance Improvement.** MobileHCI 2018 - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, Sep 2018, Barcelona, Spain. 2018,
- Jon Barker, Shinji Watanabe, Emmanuel Vincent, Jan Trmal. **The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines.** Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association, Sep 2018, Hyderabad, India. 2018.
- Emmanuel Vincent, Tuomas Virtanen, Sharon Gannot. **Audio source separation and speech enhancement**. Wiley, pp.504, 2018, 9781119279860, comme éditeurs et les chapitres suivants comme auteurs :
- Emmanuel Vincent, Sharon Gannot, Tuomas Virtanen. **Introduction**.
- Tuomas Virtanen, Emmanuel Vincent, Sharon Gannot. **Time-frequency processing - Spectral properties.**
- Timo Gerkmann, Emmanuel Vincent. **Spectral masking and filtering**.
- Emmanuel Vincent, Sharon Gannot, Tuomas Virtanen. **Acoustics - Spatial propertie**
- Emmanuel Vincent, Tuomas Virtanen, Sharon Gannot. **Perspectives**.
- Lauréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin. **CRNN-based multiple DoA estimation using Ambisonics acoustic intensity features**. Submited to the IEEE Journal of Selected Topics in Signal Processing, Special Issue on Acoustic S.. 2018.
- Ken Déguernel, Emmanuel Vincent, Gérard Assayag. **Probabilistic Factor Oracles for Multidimensional Machine Improvisation**. *Computer Music Journal*, Massachusetts Institute of Technology Press (MIT Press): Arts & Humanities Titles etc, 2018, 42 (2), pp.52-66.
- Nathan Libermann, Frédéric Bimbot, Emmanuel Vincent. **Exploration de dépendances structurelles mélodiques par réseaux de neurones récurrents**. *JIM 2018 - Journées d'Informatique Musicale*, May 2018, Amiens, France. pp.81-86, 2018,
- Denis Jouvet, David Langlois, Mohamed Menacer, Dominique Fohr, Odile Mella, et al.. **Adaptation of speech recognition vocabularies for improved transcription of YouTube videos.** *Journal of International Science and General Applications*, ISGA, 2018, 1 (1), pp.1-9.

## 2. Système Mocap d'acquisition de données multimodales

Les activités récentes en lien avec l'équipement sont les suivantes :

- Acquisition d'un corpus audiovisuel grâce au système de Mocap. Le corpus est composé de 5 000 phrases. C'est le plus grand corpus que nous avons enregistré à ce jour.
- Le corpus est utilisé pour le développement d'une technique de prédiction des mouvements du visage à partir de l'audio (thèse de Théo Biasutto-Lervat).
- Le même corpus est utilisé pour le développement d'un système de synthèse audiovisuelle ; c'est-à-dire, animer une tête parlante avec la parole à partir du texte (thèse de Sara Dahmani).
- La technique de prédiction des mouvements du visage à partir de l'audio a été intégrée dans le système de lipsync Dynalips, qui a pour but d'animer des avatars à partir de l'audio. À ce stade, nous avons développé un démonstrateur qui a été présenté au festival international d'animation à Annecy (audience : spécialistes de l'animation 3D), en juin 2018. L'objectif des travaux sur le projet de maturation Dynalips est la création d'une startup pour commercialiser cette solution.
- Au mois de novembre 2018, nous avons enregistré un corpus Mocap en allemand (dans le cadre du projet e-fran METAL : apprentissage de l'allemand langue seconde).
- Au mois de décembre 2018, nous avons enregistré un corpus Mocap en anglais.

## 3. Plateforme « Humanités numériques »

Ci-dessous, le matériel financé récemment par le CPER dans cette plateforme :

- 1 NAS de type Compelent DELL pour le stockage sécurisé des données ;
- 1 serveur Dell de type R720 pour compléter le cluster de serveurs ;
- 12 extensions mémoire de 32Go pour ajouter 128Go sur chacun des 3 serveurs du cluster LCHN ;
- 1 licence VMWare 6 Standard pour virtualiser les serveurs.

Cette plateforme est utilisée pour mettre à disposition des chercheurs des machines virtuelles qui peuvent être utilisées pour des recherches en interne ou pour mettre en place des sites internet. Nous énumérons ci-dessous une partie des projets utilisant ces machines qui sont visibles de l'extérieur.

**Projet « Grew » :**

- hébergement du site http://grew.fr de présentation et de documentation du logiciel Grew (réécriture de graphes pour le traitement automatique des langues) ;
- application en ligne match.grew.fr permettant d'effectuer des requêtes sur des corpus d'analyses syntaxiques – plus d'une centaine de corpus disponibles (notamment tous les corpus du projet Universal Dependencies en 70 langues différentes), environ 18000 requêtes servies en 2018 ;
- hébergement du site http://deep-sequoia.inria.fr de distribution des corpus Sequoia et Deep-sequoia ;
- Publications 2018
- Guillaume Bonfante, Bruno Guillaume, Guy Perrier. **Application de la réécriture de graphes au traitement automatique des langues**. *ISTE éditions,* pp.242, 2018, Série Logique, linguistique et informatique. (https://hal.inria.fr/hal-01930591)
- Bruno Guillaume, Guy Perrier. **La réécriture de graphes au service de l'annotation de corpus et de l'exploitation de corpus annotés**. Grammar and Corpora 2018, Nov 2018, Paris, France (https://hal.inria.fr/hal-01930651).

**Projet « zombilingo » :**

- hébergement du GWAP (Game With A Purpose ou Jeu ayant un but) ZombiLingo (zombilingo.org) dans lequel les joueurs doivent trouver des relations syntaxiques dans les phrases – en février 2019, 1400 joueurs sont inscrits et ont produit 485 000 annotations ;
- hébergement du GWAP RigorMortis (rigor-mortis.org) dans lequel les joueurs doivent annoter des expressions polylexicales ;
- Publications 2018
- Karën Fort, Bruno Guillaume. **Produire des données pour la recherche en jouant aux zombies**. *Interstices*, INRIA, 2018 (https://hal.inria.fr/hal-01827612)
- Karën Fort, Bruno Guillaume, Mathieu Constant, Nicolas Lefèbvre, Yann-Alan Pilatte. **"Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform.** *LAW-MWE-CxG 2018 - COLING 2018 Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions,* Aug 2018, Santa Fe, United States. pp.207 - 213 (https://hal.inria.fr/hal-01912706)

**Projet « verbenet » :**

- hébergement du site verbenet.inria.fr qui présente la ressource Verb∋Net, qui est une traduction du VerbNet de l'anglais vers le français – l'interface permet également aux experts de valider la ressource et de la mettre à jour.

**Projet « hp-papers » :**

- hébergement du site web hp-papers.lchn.fr, qui présente un outil de recherche sur la correspondance d'Henri Poincaré, un outil exploitant le web sémantique (une base de données RDF) ; l'utilisateur saisit des requêtes en langage sparql et l'application lui renvoie des résultats – e projet s'intègre dans la refonte du site utilisé pour la correspondance d'Henri Poincaré.

**Projet « Demonette » :**

- Hébergement du site web demonette.atilf.fr. Démonette est une base lexicale morphologique du français organisée en réseau dérivationnel, dont chaque entrée est un couple (Mot1, Mot2) appartenant à la même famille morphologique. Chaque entrée est décrite par 31 champs (dont la catégorie morphosyntaxique et le type sémantique de chaque mot, ainsi que la définition de Mot1 par rapport à Mot2). La version distribuée Démonette-1.2 comporte 96 027 entrées, dont les données initiales ont pour origine le TLFnome et Verbaction (cf. infra pour plus de détails).

**Projet « Fleuron » :**

- Le projet Fleuron (fleuron.atilf.fr) propose des ressources multimédias illustrant un ensemble de situations de la vie d'un étudiant en France ainsi que des outils destinés à préparer des étudiants étrangers à organiser leur séjour en France. Ces ressources permettent d'observer différentes situations, telles que :
- s'inscrire ou se réinscrire auprès d'un service administratif ;
- obtenir sa carte d'étudiant et d'autres documents universitaires officiels ;
- s'informer sur sa réussite à un diplôme ou sur ses notes ;
- obtenir des renseignements sur sa situation administrative et pédagogique ;
- discuter avec des étudiants sur des sujets divers ;
- discuter avec un enseignant sur son programme ou sur son travail.

**Projet « Metal » :**

- Le projet METAL (metal.loria.fr) se propose de concevoir, développer et évaluer un ensemble d'outils de suivi individualisé destinés aux élèves ou aux enseignants (Learning Analytics), et des technologies innovantes pour un apprentissage personnalisé des langues à l'écrit (grammaire française) et à l'oral (prononciation de langues vivantes). Il participe ainsi à l'amélioration de la qualité de l'apprentissage et au développement de la maîtrise des langues par les élèves.
  La machine virtuelle (metal.lchn.fr) n'est pas encore en production mais elle sera utilisée pour mettre en ligne un exerciseur (génération automatique d'exercices de grammaire pour les collégiens).

Voici quelques exemples de machines virtuelles utilisées en interne.

**Projet « Coread » :**

- Il existe de plus en plus de ressources de types variés disponibles, en particulier sur ANNODIS ou ORTOLANG. Parallèlement, la communauté scientifique a mis à disposition de plus en plus de documents textuels de tous genres avec notamment, outre ANNODIS et ORTOLANG, les plateformes ISTEX, SCIENTEXT, ORFEO (à venir). Partant de cet état de fait, l'opération CoReA2D vise à contribuer à la projection de ressources existantes de niveaux lexical et/ou phraséologique sur des données textuelles disponibles afin de les enrichir et d'y rendre possible une grande variété d'explorations visant par exemple à l'extraction et à la structuration de connaissances, ou encore à la classification et l'indexation de documents.

**Projet « ItsyBitsy » Editor :**

- Ce projet vise à remplacer l'éditeur lexicographique Dicet (utilisé jusqu'à présent pour construire les Systèmes Lexicaux, tels que le Réseau Lexical du Français) par ItsyBitsy, un éditeur de nouvelle génération qui aura les caractéristiques suivantes par rapport à Dicet.
- ItsyBitsy permettra de gérer les connexions interlangues via une base pivot modélisant les universaux linguistiques. Cette base pivot incorporera notamment le modèle des fonctions lexicales standard de la théorie Sens-Texte, qui servent à tisser l'ossature des Systèmes Lexicaux.
- ItsyBitsy sera utilisable soit en mode « expert » – qui présuppose une très bonne maîtrise des principes de la Lexicographie Explicative et Combinatoire –, soit en mode « grand public » – adapté pour les applications pédagogiques ou professionnelles (terminologie, etc.) de cette lexicographie.
- ItsyBitsy permettra la mise en place future d'un mode de production participative (crowdsourcing), pour l'enrichissement des données des Systèmes Lexicaux.
- ItsyBitsy est programmé sous forme d'application Web, compatible avec la majorité des navigateurs courants, afin de permettre une meilleure accessibilité aux bases lexicales.

**Projet « Franparse » :**

- Franparse, est une application client/serveur permettant de piloter un ensemble d'outils pour l'analyse en dépendances de Frantext. Ces outils pourront également être combinés à l'aide de modèles linéaires ou non linéaires afin d'améliorer les résultats des analyses.

## 4. Plateforme « E-éducation »

I. Équipement Game Lab (PC Gamer, TV 4K, enceintes et casque de réalité virtuelle)

L'équipement acquis au cours de l'année 2018 pour l'Expressive Game Lab (http://www.expressivegame.com/fr/, en partenariat avec les laboratoires CREM et LORIA de l'université de Lorraine) avait pour vocation d'offrir un environnement d'expérimentation permettant la réalisation d'analyses qualitatives de contenu quant aux formes de narration mises en œuvre dans différents genres vidéoludiques. Afin de prendre en compte la dimension sensorielle multimodale de la narration vidéoludique (conjuguant son, visuel et retour haptique) et l'évolution constante des plateformes de jeu, le matériel acquis visait la mise en œuvre d'un dispositif permettant la restitution d'un affichage très haute résolution (4K) avec son spatialisé sur 5 points d'écoute (amplification et enceintes), et d'un dispositif de réalité virtuelle (nécessitant notamment un ordinateur avec carte graphique dernière génération orientée *gaming*). Les expérimentations et analyses ont notamment pris place dans le cadre du nouveau contrat quinquennal du CREM intitulé « Narrations de la société, sociétés de la narration », sur un corpus étendu de jeux issus à la fois de la scène indépendante du jeu vidéo – notamment le genre des *walking simulators*) et des

productions à haut budget (dites *AAA*) sur consoles de salon et ordinateur individuel. Ces analyses ont permis d'élaborer une réponse à l'appel à manifestation d'intérêt de la Région Grand Est dans le cadre du dispositif « Aide aux projets collaboratifs de Recherche & Développement et d'Innovation ». Le projet retenu, intitulé « Goblinz Story », est issu d'une collaboration entre le CREM et l'entreprise mosellane Goblinz Studio (114 000 euros pour 12 mois en 2019/2020). Il permettra le recrutement d'un chercheur contractuel qui poursuivra les expérimentations à partir du matériel acquis avec le CPER.

Les expérimentations menées ont abouti à la publication d'un numéro de revue consacré aux liens entre narratologie et ludologie au sein de la revue en ligne *Sciences du jeu* (https://journals.openedition.org/sdj/894), incluant 2 articles de chercheurs du CREM :

- Genvo, Sébastien (2018). **Présentation**. *Sciences du jeu, 9*. Disponible à : http://journals.openedition.org/sdj/896
- Bazile, Julien (2018). **Ludoformer Lovecraft : Sunless Sea comme mise en monde du mythe de Cthulhu**. *Sciences du jeu, 9.* Disponible à : http://journals.openedition.org/sdj/996

Enfin, les travaux menés au sein de la plateforme ont été valorisés à l'occasion du premier symposium du réseau des laboratoires francophones sur le jeu, organisé à Montréal (14-16 mai 2018), où chaque acteur présentait les thématiques de recherche, défis scientifiques et obstacles relatifs à la création d'un Game Lab. Le premier résultat de cette collaboration sera la mise en œuvre d'une collection dédiée aux sciences du jeu aux Presses Universitaires de Liège.

II. Laboratoire d'observation portable Noldus (valise de captation vidéo et de traitement qualitatif des données filmées) et Tobii Glasses (lunettes pour l'*eye tracking*)

Le laboratoire d'observation portable Noldus est une valise « tout-en-un » comportant un équipement de captation audiovisuelle (caméras portables, micros, câbles et logiciel de pilotage) et un ordinateur mobile avec le logiciel The Observer de codage qualitatif des données recueillies. Il a été utilisé durant l'année 2018 pour des observations en contexte réel (dit *écologique*) dans le cadre du projet e-TAC « Environnements Tangibles Augmentés pour l'Apprentissage Collaboratif » (http://e-tac.univ-lorraine.fr/, associant les laboratoires PErSEUs, CREM et LCOMS de l'Université de Lorraine) sur les territoires numériques éducatifs. Il a permis de filmer plusieurs séquences de codesign et de conception collaborative réalisées en cycles 3 et 4 dans deux établissements scolaires (école et collège) de Moselle, qui ont ensuite fait l'objet d'analyse de contenus via le logiciel The Observer, pour catégoriser les activités ainsi filmées par types d'interactions, outils/matériels utilisés et comportements des élèves. L'objectif du projet e-TAC est ensuite d'évaluer l'impact des interfaces tangibles augmentées, qui sont actuellement en cours de développement au sein du projet et dont les premiers prototypes seront testés en 2019, sur les processus d'apprentissage collaboratif en groupe. La particularité de ces interfaces émergentes est de ne plus faire appel à des actions via un clavier-souris-écran d'ordinateur, mais de manipuler des objets numériques tangibles en interaction avec des objets matériels présents dans la classe et disposés sur une table.

Ces expérimentations ont fait l'objet de plusieurs actions de valorisation, parmi lesquelles :

- -Olry, Alexis (2018). Les Interfaces Tangibles Augmentées en contexte scolaire : favorable à la collaboration ? In : *Actes des Septièmes Rencontres Jeunes Chercheurs en EIAH (RJC EIAH 2018)*.

- -Humbert, Pierre et Roussel, Benoit (2018). Interfaces tangibles et augmentées pour la co-conception en classe. Présenté à GT EduIHM, 30e conférence francophone sur l'Interaction Homme-Machine (IHM 2018).
- -Co-organisation durant la conférence IHM'17 et IHM'18 de workshops dans le cadre du groupe de travail « IHM et éducation ».

Dans le cadre de ce projet, l'utilisation des Tobii Glasses (lunettes permettant l'enregistrement de données en *eye-tracking*) a été reportée à l'année 2019 pour les observations et analyses d'usage des premiers prototypes d'interfaces tangibles augmentées, qui seront testés en classe.

III. Equipement Tobii Bar (*eye tracker* mobile, de type « barre »)

La Tobii Bar a été utilisée dans le cadre de l'intégration d'un module d'oculométrie dans l'application Evalyzer (http://www.evalyzer.com/fr/), conçue avec le soutien de la SATT Grand Est et de chercheurs du laboratoire PErSEUs (université de Lorraine). Evalyzer est une plateforme Web qui permet de réaliser des tests utilisateurs à distance. Ainsi, des personnes ayant accepté de participer à une étude peuvent réaliser des tâches de recherche d'information sur sites Web sans avoir à se déplacer dans un laboratoire d'usage, en utilisant leur propre environnement matériel. La plateforme Evalyzer permet donc de définir le protocole de test, d'inviter les participants à prendre part à l'étude et d'enregistrer les comportements de l'internaute. Les données recueillies lors de ces tests utilisateurs sont : le temps de réalisation de chacune des tâches, les parcours dans le site Web, l'enregistrement vidéo de la session, les pages consultées, la distance parcourue par la souris, les scrolls des pages, etc. De plus, des questionnaires peuvent être administrés à l'aide de la plateforme. Toutes ces données sont envoyées sur les serveurs du projet et différentes métriques sont calculées, comme les taux de revisites des pages Web. Afin de compléter ces données recueillies à distance par des données oculométriques recueillies en laboratoire, l'équipe projet a décidé d'intégrer un module d'oculométrie à Evalyzer. En réalisant cette intégration, elle évite l'utilisation d'autres logiciels comme Tobii Studio, qui permet aussi de réaliser des tests utilisateurs sur sites Web, mais qui ne permet pas d'intégrer les données d'un test aux données recueillies par Evalyzer. Pour le développement de ce module, toujours en cours et réalisé avec le soutien financier de la SATT Grand Est, l'équipe a utilisé le SDK de Tobii et la Tobii Bar.

# Bilan du travail des ingénieurs

## 1. Travail effectué par Simon Méoni

Documentation et code produit (CoReA2D et Démonette) : https://simonmeoni.github.io/documentation-atilf

**CoReA2D**
L'objectif de ce projet est de bâtir un environnement d'annotation manuelle afin de produire des corpus de référence à grande échelle à destination des algorithmes basés sur un apprentissage sur corpus et à destination des algorithmes non supervisés comme élément de comparaison en vue de la mesure des performances de tous les algorithmes

par apprentissage.

Dans cette perspective, la première phase du projet est d'évaluer l'existant en termes d'environnement d'annotation. Trois environnements ont été choisis : BRAT pour sa simplicité et son accès web, GLOZZ pour sa position dominante dans la communauté de l'annotation, GATE pour sa position dans la communauté du TAL. Les données utilisées pour cette évaluation sont issues du projet TERMITH actuellement déposé sur Ortolang.

Les tâches effectuées par M. Méoni ont consisté à :

- extraire les données utilisées pour le test des environnements existants ;
- assurer l'interopérabilité entre le format des données utilisées (XML-TEI-P5-STDF-TBX) et les formats d'entrée des trois environnements ;
- assurer le développement de toutes les briques logicielles nécessaires à la réalisation d'une campagne d'annotation en conditions réelles :
  - accès au logiciel d'annotation ;
  - formation des annotateurs recrutés par Evelyne Jacquey sur chaque environnement ;
  - calcul de l'accord interannotateur et report des annotations après arbitrage ;
  - accès à des données externes utilisées lors de l'annotation :
    - bases de données lexicales (mises à jour régulières en fonction de l'avancement d'équipes partenaires qui amendaient régulièrement les données) ;
    - documentations sur les environnements ;
    - consignes d'annotation.

***Publications***

Soumis : **Annotation manuelle de candidats termes et écrit scientifique**, *Revue TAL*, N° thématique sur *Corpus annotés*, sciencesconf.org:tal-60-2:233634

**Démonette**

La base de données lexicale Démonette réunit sous forme tabulée l'ensemble des descriptions pertinentes servant à identifier les propriétés morphologiques, sémantiques formelles et structurelles de chaque entrée, qui est une relation dérivationnelle entre deux mots du français. Le travail de S. Méoni a consisté en l'élaboration d'une interface permettant à un utilisateur d'interroger la base à distance, de formuler des combinaisons de requêtes portant sur les différents types d'information contenue, et de pouvoir visualiser les résultats sous une forme graphique.

Tout au long du projet, M. Méoni a rendu compte de ses progrès, par des rapports écrits et des versions de travail de l'interface. Celle-ci est désormais accessible au public : https://demonette.atilf.fr/

## 2. Travail effectué par Nabil Gader

I. Mise en place d'une plateforme d'analyse syntaxique

Une annotation syntaxique de corpus de qualité implique souvent l'interaction de deux processus: une annotation automatique suivie d'une vérification et correction manuelle. Ce type de procédure est très commune dans la communauté du traitement des langues et il existe un besoin d'outils facilitant le travail.

Depuis 2018, Nabil Gader est chargé de développer une application web d'annotation syntaxique de textes. Cette application permettra aux utilisateurs ayant un compte d'appliquer des analyseurs syntaxiques sur les textes de leur choix puis de corriger manuellement l'annotation. L'application donnera la possibilité aux utilisateurs d'améliorer incrémentalement les modèles d'analyse à partir des textes corrigés. Il est encadré par Mathieu Constant (PR Université de Lorraine, ATILF) et travaille en étroite collaboration avec le service informatique de l'ATILF, et en particulier Cyril Pestel (IE CNRS).

Durant la première année du projet, Nabil Gader a mis en place l'architecture générale de l'application avec la sélection et l'implantation de diverses solutions technologiques (ex. docker, angular, …) et composants logiciels (ex. Bratt). La deuxième année sera consacrée à l'intégration finale de l'application :

- mise en place effective des analyseurs syntaxiques ;
- finalisation de la chaîne de communication entre les différents composants logiciels ;
- gestion des comptes utilisateur ;
- ajout de fonctionnalités de prétraitement intégrant différents outils de segmentation, lemmatisation et étiquetage grammatical.

Il est prévu une période de test auprès d'un ensemble d'utilisateurs sélectionnés au sein de l'ATILF, notamment des tests sur les textes de Frantext.

En termes de débouchés, le code source de l'application sera distribué sous une licence libre. Il est prévu de publier un article sur le sujet en fin de projet et de participer à diverses sessions de démonstration dans conférences comme TALN.

## II. Construction de l'éditeur lexicographique ItsyBisty

La lexicographie des Systèmes Lexicaux repose sur l'utilisation d'un éditeur lexicographique spécialement conçu pour le tissage des réseaux lexicaux. Le Réseau Lexical du Français (RL-fr), notamment, a pu être développé grâce à un éditeur spécialement conçu à cette fin dans le cadre du projet majeur RELIEF (2011–2014). Ce dernier est une application Java conçue pour donner accès à tous les Systèmes Lexicaux du type RL-fr, pour toutes les langues concernées, stockés sous forme de bases de données SQL.

Le projet Itsy Bitsy Editor vise à remplacer l'éditeur Dicet par un éditeur de nouvelle génération qui aura les caractéristiques suivantes par rapport à Dicet :

- gestion de connexions interlangues via une base pivot modélisant les universaux linguistiques ;
- possibilité d'éditer les réseaux lexicaux soit en mode « expert » soit en mode « grand public » ;
- application Web compatible avec la majorité des navigateurs courants ;
- système *Open Source* avec architecture modulaire, conçu pour un éventuel travail collaboratif ;
- intégration graduelle d'un mode de production participative (*crowdsourcing*) contrôlé ;
- intégration avec un navigateur graphique de réseaux lexicaux.

Durant l'année 2018–2019, l'Ingénieur d'Étude (IE) engagé pour effectuer le développement informatique de l'éditeur, Nabil Gader, a mis en place la nouvelle structure informatique de la base lexicale permettant le mode de fonctionnement caractérisé ci-dessus. Le travail s'est effectué sous la direction d'Alain Polguère (PR Université de Lorraine, ATILF), avec la collaboration de deux IE permanents du CNRS : Sandrine

Ollinger et Cyril Pestel.

Durant la seconde année du projet, Nabil Gader va réaliser la programmation de l'interface web permettant l'édition lexicographique proprement dite sur la nouvelle structure de base de données qui vient d'être mise en place.

Publications et débouchés scientifiques attendus

Une publication est attendue à la fin du projet, ainsi qu'au moins une présentation en séminaire de recherche. Pour l'instant, une page web a été conçue pour documenter le projet sur le site ATILF des Systèmes Lexicaux :
https://lexical-systems.atilf.fr/lexicographie/

En facilitant le travail lexicographique à distance, l'éditeur ItsyBitsy est notamment appelé à jouer un rôle crucial dans le cadre de nos collaborations avec des laboratoires extérieurs, comme l'OLST de l'Université de Montréal. Il sera également exploité dans le cadre des applications pédagogiques des travaux sur les grands réseaux lexicaux – cf. la convention signée entre le CNRS et la DSDEN (Direction des Services Départementaux de l'Éducation Nationale) de Meurthe-et-Moselle encadrant la collaboration LELREP avec la REP+ La Fontaine.

# 3. Travail effectué par Mamadou Diallo

## I. Projet Needle

Needle est un outil de navigation Web fondé sur la collaboration et la contribution, par sélection de pages jugées intéressantes par ses usagers, sous forme d'extensions pour Firefox et Chrome. Depuis juillet 2018, il a fait l'objet de plusieurs tâches et réalisations dont :

- développement d'un système de recherche sur la base de mots clés ou de *hashtags* correspondants aux descripteurs des pages sélectionnés par les usagers ;
- ajout d'un filtre dédié à la sélection de navigation par mots clés ou usagers ;
- module de suppression de références ;
- possibilité de mettre à jour automatiquement la légende d'une référence ;
- ouverture au public et débogages des problèmes associés à cette mise à disposition ;
- intégration d'un système d'accès à Needle sur d'invitation ;
- déploiement de nouvelles versions pour Firefox et Chrome ;
- correction de divers bogues.

Valorisation du projet Needle

La plateforme Needle est désormais mise à disposition de la communauté universitaire depuis la page http://needle.univ-lorraine.fr. Elle fait l'objet d'un dépôt de droits d'usages et d'exploitation sous licence libre. Les données qu'elle recueille le sont de manière totalement anonyme et sont exploitables par les chercheurs de l'Université de Lorraine.

Ce projet a fait l'objet en 2018 de deux communications dans des conférences avec actes, d'une publication dans une revue scientifique (Hermès), de deux articles grand public sur

le site The Conversation France (dont l'un a dépassé les 18000 vues), qui ont permis d'attirer plus 1100 bêta-testeurs supplémentaires. Le projet Needle a été sélectionné pour être présenté les 15 et 16 mai 2019 sur le salon Innovatives SHS organisé par le CNRS pour promouvoir les innovations issues de la recherche en sciences humaines et sociales.

## II. Projet Publictionnaire

Le Publictionnaire est un dictionnaire encyclopédique et critique des publics disponible en ligne (http://publictionnaire.huma-num.fr). Depuis janvier 2019, il a fait l'objet :

- d'une mise à jour de l'interface de saisie des notices via Wordpress consistant, pour l'essentiel, en une consolidation et amélioration de la saisie des liens (gestion et affiliation des auteurs, élimination des redondances, relations entre notices) ;
- de l'intégration d'un moteur de recherche interne dédié (en cours de finalisation) ;
- d'une implémentation selon le protocole OAI-PMH des notices afin de les rendre plus facilement moissonnables par les moteurs académiques (dont https://isidore.science/).

### Valorisation du projet Publictionnaire

Le dictionnaire comporte désormais plus de 240 notices et a déjà fait l'objet d'une communication en novembre 2018 lors d'une journée d'étude sur Paris.

Publications :

- Falgas, Julien, 2018. **Needle, une innovation issue des sciences de l'information et de la communication face à la crise de l'inspiration**. In : *XXIe congrès de la Société française des sciences de l'information et de la communication, Création, créativité et médiations »* - Actes vol. 3 : objets techniques, dispositifs et contenus (pp. 221-236). SFSIC, MSH Paris Nord, Paris.
- Falgas, Julien, 2018. **Needle, la navigation web contributive comme modalité d'accès éthique aux documents numériques**. In : Balicco, Laurence, Broudoux, Évelyne, Chartron, Ghislaine, Clavier, Viviane et Pailliart, Isabelle, 2018. L'éthique en contexte info-communicationnel numérique : déontologie, régulation, algorithme, espace public – *Actes du colloque « Document numérique et société »* (p. 115-125). De Boeck Supérieur, Louvain-la-Neuve.
- Falgas, Julien, 2018. « **Needle, mettre la critique des GAFA à l'épreuve de l'expérience** », *Hermès*, vol. 3, 82.
- Falgas, Julien, 2018. **Crise de l'imagination : l'inventeur du web prend ses responsabilités, et vous ?** *The Conversation France*, octobre. Disponible à l'adresse : https://theconversation.com/crise-de-limagination-linventeur-du-web-prend-ses-responsabilites-et-vous-104353
- Falgas, Julien, 2019. **Ceux qui ont lu cet article ont participé à une expérience qui pourrait révolutionner leur manière de s'informer.** *The Conversation France*, janvier. Disponible à l'adresse : https://theconversation.com/ceux-qui-ont-lu-cet-article-ont-participe-a-une-experience-qui-pourrait-revolutionner-leur-maniere-de-sinformer-110492
- Walter, Jacques, 2018. **Présentation du Publictionnaire.** *Journée d'étude Musées & recherche 2018 - Le souci du public,* novembre, Paris. Disponible à l'adresse : https://ocim.fr/formation/musees-recherche-2018-le-souci-du-public/#

# Impacts des projets financés antérieurement

## PROSODCORPUS

Les premières années du CPER ont contribué à financer la préparation et l'annotation de données concernant les particules de discours, ainsi que quelques travaux préliminaires sur ce thème. En 2017, nous avons proposé à l'ATILF une thèse sur les particules de discours, qui a été retenue, et qui a débuté à l'automne 2017. Il s'agit de la thèse de Lou Lee ; c'est aussi elle qui avait travaillé auparavant à l'annotation des données.

Publications récentes :

- Lou Lee, Katarina Bartkova, Mathilde Dargnat, Denis Jouvet : **"Prosodic and Pragmatic Values of Discourse Particles in French"**; *ExLing 2018 – 9th Tutorial and Research Workshop on Experimental Linguistics,* Aug 2018, Paris, France. 2018
- Katarina Bartkova, Denis Jouvet : **"Analysis of prosodic correlates of emotional speech data"**; *ExLing 2018 – 9th Tutorial and Research Workshop on Experimental Linguistics*, Aug 2018, Paris, France. 2018

## ITL-DI-Œil – Interrelation Troubles du Langage, Discours et Processus Oculomoteurs

Publications récentes :

- Musiol, M ; Bocéréan, C ; Hoffmann, C ; Barthélémy, S ; Padroni, S ; Franck, N ; Demily, C. (2019)**. Do you pay attention with me when we talk? A double eye-tracking study in schizophrenia.** *Psychiatry Research* (soumis).
- Bocéréan, C ; Hoffmann, C ; Padroni, S ; Franck, N ; Demily, C ; Musiol, M. (2018). **Do you see what I mean? What eye movement rates (EMRs) can tell us about memory processes involved in clinical verbal interactions.** *BMC Psychology*. (soumis).
- Rebuschi, M., Musiol, M., Amblard, M. (2019). **Coherence and Incoherence, From Psychology to Linguistics and Back.** In M. Amblard, M. Musiol & M. Rebuschi (eds). (2019). *(In)coherence of Discourse. Formal and conceptual issues of language. Springer, series: Language, Cognition and Mind* (Chungmin Lee "Editor Springer book series). À par.
- Rebuschi, M., Musiol, M., Amblard, M. (2019). **Corpora et psychopathologie** (2019). *Corpus. (MSH-L USR UL)*, à par.
- Besche-Richard, C., Musiol, M. (2018). **Les troubles schizophréniques.** In C. Besche-Richard (éd). *Psychopathologie de l'adolescent et de l'adulte : perspectives cognitives et neuropsychologiques*. Paris : Dunod, Ch 6, 141-162.
- Musiol, M. (2019). **Les Troubles du Langage et de la Pensée chez le patient SCZ au risque de la réaction visuo-attentionnelle de son interlocuteur dans l'entretien clinique.** Communication invitée, *16e Congrès de Psychiatrie, « Les schizophrénies »,*
- Centre Hospitalo-Universitaire de Tizi Ouzou et Association des Psychiatres du Djurdjura, Auditorium de Tizi Ouzou, 13-14 mars 2019.
- Musiol, M. (2019). **Troubles psychopathologiques et troubles oculomoteurs.** *Communication invitée,* Laboratoire de psychologie clinique et neuropsychologie,

Université de Picardie, mai 2019.

- Musiol, M. (2018). **Attention visuelle, mouvements de regard et discours.** *Communication invitée,* Laboratoire de Psychologie Cognitive, Université de Paris 5. 3 sept 2018.

## Démonette-1.3

Actions menées en 2018

A) Acquisition de propriétés phonétiques et morpho-phonologiques : ce travail a consisté à dégager les régularités morpho-phonologiques du lexique morphologiquement construit. Il a comporté diverses étapes:

1- acquérir la transcription phonétique des lexèmes dont la graphie est conventionnellement représentée par M1 et M2. Deux sources se complémentent à cet effet : Lexique.org (www.lexique.org) et Glaff (http://redac.univ-tlse2.fr/lexiques/glaff.html). Un lexème n'ayant pas de forme phonologique en soi (c'est une entité abstraite qui se fléchit en un ensemble de mots, c.-à-d. son paradigme flexionnel, suivant les contraintes d'accord imposées par le contexte) sa représentation phonétique est la collection de ses radicaux utilisés en flexion et en construction, appelée espace thématique. La taille et le contenu de chaque espace dépendent de la partie du discours du lexème décrit. Les seuls radicaux enregistrés dans chaque entrée de Démonette sont ceux qui sont pertinents en dérivation (et qui donc sont utilisés dans la relation entre M1 et M2). La valeur de chaque radical est reconstituée à partir de la transcription phonétique du mot-forme approprié réalisant le lexème et présent dans la source.

Cette tâche a été réalisée sur les 167 369 entrées de Démonette1.3, au moyen des transcriptions phonémiques du Glaff (un lexique formé de 1 406 857 entrées du français et construit à partir du Wiktionnaire, cf. http://redac.univ-tlse2.fr/lexiques/glaff.html), codées en SAMPA.

2- Identifier la séquence phonique commune à M1 et M2 et le type de variation radicale, qui accompagne, le cas échéant, la construction morphologique de M2 à partir de M1, et vice versa.

3- regrouper les relations M1, M2 en fonction du modèle formel abstrait auquel elles appartiennent.
On distingue les relations régulières (pas de variation), les relations de substitution de phonème (t/s), les relations faisant intervenir l'adjonction (--/at) ou la suppression d'un segment (at/--). Enfin, certaines relations sont supplétives : les radicaux n'ont rien en commun (ni/negas).

Les tâches 2 et 3 ont été réalisées (codage, validation) avec l'aide d'un étudiant en M1 SDL (avril-septembre 2018) dont le travail (2 mois de vacations) a servi de base à la rédaction de son mémoire de maîtrise.

Publication et mise à disposition des résultats

1) Les résultats obtenus ont directement fait l'objet de deux présentations orales :

- conférences internationales DeriMo (http://derimo2017.marginalia.it/index.php/

programme) et ISMO (https://colloque-ismo.univ-lille3.fr/index.php), avec publication dans les actes :

- Namer, F., Hathout, N. and Lignon, S. (2017). **"Adding morpho-phonological features into a French morpho-semantic resource: the Demonette derivational database"**. *Proceedings of the First International Workshop on Resources and Tools for Derivational Morphology (DeriMo),* Milan, Italy: 49-61.

2) Ces résultats ont également été exploités dans les présentations suivantes :

- Lignon, S., Huguin, M., Hathout, N. and Namer, F. (2018). **''Between morphophonology and paradigms: the choice of the form of the base in French derivation"**. *Revisiting Paradigms in Word-Formation (Workshop organized in the framework of the conference WORD-FORMATION THEORIES III TYPOLOGY AND UNIVERSALS IN WORD-FORMATION IV),* Košice, Slovakia.
- Hathout, N. and Namer, F. (2018). **"ParaDis: a Families-and-Paradigms model for derivation".** *Revisiting Paradigms in Word-Formation (Workshop organized in the framework of the conference WORD-FORMATION THEORIES III TYPOLOGY AND UNIVERSALS IN WORD-FORMATION IV)*, Košice, Slovakia.

Des nouvelles collaborations ont été rendues possibles suite à la diffusion des résultats de Demonette1.3 :

- Dans le cadre de la conférence DeriMo, avec l'équipe du Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE), Università Cattolica del Sacro Cuore, Milan, et notamment les responsables du projet Word Formation Latin Team (Eleonora Litta Modignani Picozzi, Marco Passarotti). Dans le cadre de cette nouvelle collaboration, il est prévu d'organiser l'édition 2021 DeriMo à Nancy.
- Toujours dans le cadre de DeriMo Milan, mais également lors de la conférence à Košice, des liens se sont renforcés entre l'ATILF et l'Institute of Formal and Applied Linguistics de l'université Charles, à Prague. Dans ce cadre, je suis oratrice invitée à Prague en septembre prochain (édition 2019 de la conférence DeriMo).
- A l'occasion de la conférence ISMO, des contacts ont été pris avec Pr. Lior Laks de l'université de Bar Ilan, spécialiste de la morphologie des langues sémitiques, en vue de développer un Démonette1.3 pour l'hébreu moderne et l'arabe palestinien. Ce projet s'est concrétisé par l'invitation de L Laks en tant que Professeur invité à l'Atilf, invitation concrétisée dans le cadre du dernier AAP émanant de l'UL (campagne 2018). L. Laks a séjourné à Nancy du 5 novembre au 5 décembre 2018. Un article en collaboration est en préparation pour faire connaître les premiers résultats de ce travail.

Nouveaux financements obtenus :

Le soutien financier et logistique du CPER a enfin été crucial pour faire avancer suffisamment la base de données et pouvoir présenter et obtenir une demande de financement auprès de l'ANR. Le dossier déposé a réuni un consortium de 4 UMR (STL, CLLE-ERSS, LLF, ATILF), l'ATILF étant partenaire principal.
Le dossier a été déposé en septembre 2016 (première phase de l'AAP). Un financement de 600 000 euros pour 4 ans a été accordé, sous le label ANR-17-CE23-0005. Le projet, intitulé Demonext "Dérivation Morphologique en Extension" a démarré le 2 avril 2018. Cette année-là, les travaux réalisés dans le cadre du CPER ont été complémentaires de ceux programmés dans le cadre du projet ANR.