CPER LCHN
Langues, Connaissances et Humanités Numériques

———————

Bilan 2019

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 qui interviennent plusieurs mois ou plusieurs années après. Nous considérons cependant que ce bilan est 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 les projets concernés. 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 ;
- analyse de scènes auditives ;
- reconnaissance de la parole robuste au bruit ;
- reconnaissance de la parole protégeant la vie privée ;
- rehaussement de parole ;
- séparation de sources de parole et localisation de locuteurs ;
- adaptation et extension de lexique pour la reconnaissance de la parole ;
- génération automatique de résumés ;
- analyse des actes de dialogue et analyse des sentiments ;
- analyse syntaxique de textes ;
- planification de trajectoires en robotique ;
- classification de textes et traitement des dialogues dans les microblogs ;
- synthèse audiovisuelle expressive.

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 :

- 4 articles dans des revues scientifiques ;
- 20 articles dans des conférences ;

- 1 livre ***Audio source separation and speech enhancement*** pour lequel un chercheur du LORIA fait partie des directeurs de publication et dont 6 chapitres sont liés aux activités du CPER.

La plateforme a également servi d'effet levier pour lever des fonds pour étendre ces travaux notamment au travers des nouveaux projets suivants :

- projet Impact LUE OLKi (1.5M€ du PIA LUE) ;
- projet H2020 Comprise (3M€) ;
- projets ANR LEAUDS, ROBOVOX, JCJC DiSCogs.

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 – 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, un démonstrateur 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.
- Dans le cadre du projet e-fran METAL, un corpus Mocap en allemand (pour l'apprentissage de l'allemand langue seconde) et un corpus Mocap en anglais ont été enregistrés.

3. Équipement pour l'acquisition d'IRM dynamique

L'équipement pour l'acquisition d'IRM dynamique a été installé le 14 juin 2019 sur le système IRM 3T Prisma Siemens du CHRU de Nancy Brabois par le Pr Jens Frahm du Max Planck Institute de Göttingen.

L'équipement d'acquisition d'IRM dynamique va être utilisé pour étendre le nombre de sujets de la base de données d'imagerie ArtSpeechMRIfr : une base de données d'imagerie par résonance magnétique statique et en temps réel (IRMrt, IRM 3D) du conduit vocal. La base de données contient également des données traitées : audio débruité, des annotations alignées phonétiquement, des contours articulatoires et informations sur le volume du conduit vocal, qui constituent une ressource précieuse pour la recherche sur la parole. En plus de la base de données déjà publiée qui repose sur les données de deux hommes parlant le français, les chercheurs ont déjà réalisé les acquisitions de quatre locuteurs supplémentaires avec le système d'acquisition d'IRM dynamique.

Le corpus de rtMRI inclus dans cette base de données comprend 79 phrases synthétiques construites à partir d'un dictionnaire phonétisé, qui permet de raccourcir la durée des acquisitions tout en gardant une très bonne couverture des contextes phonétiques existant en français.

L'arrivée réellement opérationnelle de l'IRM temps réel (à 50 Hz) permet d'acquérir de grandes bases de données[1] sur l'évolution temporelle du conduit vocal. Au-delà du simple progrès technologique, l'IRM temps réel bouleverse la manière d'envisager la modélisation temporelle de la géométrie du conduit vocal. L'IRM 3T Siemens de la plateforme d'IRM du CHRU de Nancy est équipé du seul système d'IRM temps réel en France (développé par l'équipe de Jens Frahm au Max Planck Institute de Göttingen). À ce jour, les chercheurs ont déjà inclus 25 sujets qui ont utilisé cette acquisition. Seule une coupe (en général dans le plan médio-sagittal) peut être acquise à l'aide de l'IRM temps réel (figure ci-contre) par acquisition. Les travaux de réalisation de cette base de données sont financés dans le cadre du projet ANR 2015 ArtSpeech.

La thèse de Ioannis Duros – coencadrée par Yves Laprie du Loria et Pierre-André Vuissoz du laboratoire IADI INSERM U1254 et financée dans le cadre de l'initiative Lorraine Université d'Excellence – utilise aussi l'acquisition d'IRM dynamique. L'objectif de cette thèse est de résoudre le problème de la création d'un atlas dynamique 3D du conduit vocal qui capture la dynamique des articulateurs dans les trois dimensions. La méthode développée utilise l'IRM 2D temps réel dans plusieurs plans sagittaux et pour plusieurs locuteurs. Après un alignement temporel, les acquisitions sont combinées pour constituer un espace de référence qui supprime les particularités anatomiques des différents locuteurs pour ne garder que la variabilité de la production de la parole afin de constituer un atlas.

De plus, dans le cadre de AAPG 2020 de l'ANR, deux propositions de projets de recherche nécessitant l'utilisation de l'acquisition d'IRM dynamique ont été soumises : « Fully 3D talking head with aero-acoustic simulations » et « AeroNaV : Modélisation Aérodynamique, articulatoire et acoustique de la Nasalité dans la Voix des sujets avec polyposes naso-sinusiennes ».

Par ailleurs, la Dr Karyna Isaieva travaille avec l'acquisition d'IRM dynamique pour développer une technique de suivi de la pointe de la langue pour calculer sa vitesse et comparer ces nouveaux résultats avec une technique de contraste de phase en ciné IRM qui permet de mesurer la vitesse des tissus directement.

4. 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 mémoires de 32Go pour ajouter 128Go sur chacun des 3 serveurs du cluster LCHN ;
- 1 licence VMWare 6 Standard pour virtualiser les serveurs.

[1] https://hal.inria.fr/hal-02167756

Cette plateforme permet de 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 « Franparse » :

- Un ingénieur financé par le FEDER dans le cadre de ce CPER a développé une plateforme d'annotation syntaxique de corpus. L'année précédente a permis de mettre en place les briques de la plateforme. L'application web est actuellement en test et sera prochainement déployée en production à l'URL publique suivante: https://franparse.atilf.fr.

Projet « PARSEME-FR » :

- Hébergement du démonstrateur https://mwedemonstrator.atilf.fr du projet ANR PARSEME-FR (https://parsemefr.lis-lab.fr). Cette plateforme permet de tester les outils d'identification d'expressions polylexicales du projet sur des textes. Elle permet également d'explorer un corpus annoté en expressions polylexicales verbales alignées à un lexique de telles expressions
- Développement d'un système automatique de lemmatisation des expressions polylexicales. Le système a été initialement développé pour le français, a été adapté pour l'italien, le polonais, et le portugais.

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 – plusieurs centaines de corpus disponibles (notamment les 157 corpus du projet Universal Dependencies en 90 langues différentes), environ 80 000 requêtes servies depuis 2018.

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, 1 400 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.

Projet « Encyclopédie Grammaticale du Français » :

- Mise à disposition de notices portant sur des notions de référence en Sciences du langage (encyclogram.fr) – un ingénieur du CPER a mis en place une plateforme OJS pour la gestion des notices et les liens avec les rédacteurs et relecteurs.

Projet « Cahiers de lexicologie » :

- Mise en ligne d'une revue de lexicologie (https://cahierslexico.atilf.fr). Un ingénieur du CPER a mis en place une plateforme OJS pour la gestion des notices et les liens avec les rédacteurs et relecteurs.

Projet « Systèmes Lexicaux » :

- Développement et mise à disposition du site web https://lexical-systems.atilf.fr/ présentant les Systèmes Lexicaux (https://lexical-systems.atilf.fr/).

Projet « TCOF » :

- TCOF est un corpus oral d'enfants et d'adultes destiné à l'analyse des interactions (https://www.ortolang.fr/market/corpora/tcof). Une refonte de la plateforme de dépôt des corpus (transcription, fichiers audio/vidéo, métadonnées) a été réalisée. De plus, dans ce cadre, un stage a été effectué autour de l'annotation automatique du corpus en POS.

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 ; le 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.

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 « CoReA2D » :

- 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. 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 « ItsyBisty » :

- L'outil ItsyBisty est un éditeur de graphes lexicaux (décrit en détail ci-dessous dans la rubrique *Bilan du travail des ingénieurs*).

5. Plateforme « E-éducation »

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

L'expressive Gamelab est une plateforme d'analyse de contenus et d'usages de jeux vidéo, de dispositifs ludiques ouverte à des chercheurs, doctorants, postdoctorants et étudiants de Master et Licence (http://www.expressivegame.com). Depuis, septembre 2019, le site de présentation de la plateforme a fait l'objet :

- d'une refonte intégrale de l'interface du site (reformulation des catégories, mise en page, ajout de nombreux éléments rendant le site dynamique, notamment à travers la mise en avant des actualités en page d'accueil) ;
- de l'amélioration du moteur de recherche intégré, qui ne pouvait pas précédemment réaliser des recherches dans le contenu des pages, mais uniquement dans leurs titres ;

- de la restructuration de la base de données du fonds d'archive (jeux, magazines, ouvrages scientifiques, etc.) mis à disposition à l'Expressive Gamelab (voir ci-dessous).

Valorisation du projet Expressive Gamelab

Sur la période concernée, l'Expressive Gamelab a accueilli le projet Goblinz Story, financé par l'intermédiaire d'un Appel à Manifestation d'Intérêt de la Région Grand Est « Aide aux projets collaboratifs de Recherche & Développement (R&D) et d'Innovation » (partenaires : studio de développement Goblinz Studio / Centre de recherche sur les médiations). Il a pour l'instant fait l'objet de trois conférences avec actes (le projet s'est terminé en février 2020). L'Expressive Gamelab mène également des projets de recherche–création, dont un a abouti sur la période concernée à travers la création du jeu expressif *Lie in my heart* (disponible depuis octobre 2019 sur la plateforme steam). Ce projet a trouvé de nombreux échos dans la presse nationale (entre autres journaux, *Le Monde*, *Libération*, *Les Inrockuptibles*, un article dans *The Conversation*) et internationale (*Eurogamer*), a abouti à plusieurs conférences et présentations pour le grand public (Salon IndieCade à la BNF- Paris, conférence à la Paris Games Week, présentation au Luxembourg Gaming Xperience) et a fait l'objet d'une communication en colloque, plusieurs conférences et publications d'articles à l'international étant (notamment en février 2020 un cycle de conférences à l'UQAM, Montréal, et une publication d'un ouvrage collectif en espagnol sur la notion de jeu expressif, coordonné par des chercheurs de l'université Alcalá de Henares, Madrid). Enfin, des liens se sont renforcés avec les formations du département Information–Communication de l'UFR SHS-Metz, notamment le parcours de Master Conception de dispositifs ludiques, en mettant en œuvre au sein de la plateforme la réalisation de certaines analyses menées dans les enseignements. Le site du Gamelab permet de mettre en visibilité l'ensemble de ces initiatives.

- **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 sont en cours de test, 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.

Dans le cadre de ce projet, les Tobii Glasses (lunettes permettant l'enregistrement de données en *eye-tracking*) sont utilisées en classe pour les observations et analyses d'usage des premiers prototypes d'interfaces tangibles augmentées.

- **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, déposées 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 ;
  - 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.

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

- **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 commun 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).

En 2019, il a effectué la mise en place effective sous la forme d'une application Web. L'application propose aux utilisateurs d'annoter automatiquement leurs textes en syntaxe de dépendances, en utilisant un ou plusieurs analyseurs existants, puis de valider ces annotations manuellement. Elle offre aussi la possibilité d'apprendre de nouveaux modèles d'annotation syntaxique à partir de corpus déjà annotés.

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.

- **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.

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

- **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 ;
- 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 ;
- implémentation du protocole ActivityPub, recommandé par le W3C pour les réseaux sociaux distribués. L'aboutissement de ces développements doit permettre à Needle d'intégrer le Fediverse au moment de sa publication sous licence libre, favorisant ainsi sa dissémination.

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 18 000 vues), qui ont permis d'attirer plus 1 100 bêta-testeurs supplémentaires. Needle a été présenté lors d'un séminaire du CEREFIGE, de la journée de lancement du projet Impact LUE OLKi en mars 2019, ainsi que 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.

Enfin, Needle est au cœur du propos d'un chapitre d'ouvrage à paraître sous la direction de Pascal Robert.

Plusieurs centaines de bêta-testeurs ont utilisé la plateforme Needle et révélé son potentiel à être consolidée et porté sur mobile. Ces développements s'appuieront sur une refonte de l'interface et sur l'implémentation de mécaniques d'engagements pour retenir et impliquer les utilisateurs. Des algorithmes d'IA alimenteront un moteur de recommandation hybride et assisteront la modération des contributions des utilisateurs. Le projet pourrait débuter à l'été 2020 dans la perspective de la création d'une entreprise à l'automne 2021.

- **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 qui est opérationnel ; ce moteur permet la recherche sur le titre d'une notice ou la recherche plein texte ;
- 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/). Le moissonnage par isidore.science est effectif depuis septembre 2019.

Valorisation du projet Publictionnaire

Le dictionnaire comporte désormais plus de 300 notices et a fait l'objet d'une communication en novembre 2018 lors d'une journée d'étude sur Paris. Il est référencé par le site Eduscol qui signale des notices pertinentes pour l'éducation aux médias et à l'information, le français et les langues. Des lettres thématiques mentionnant les notices pertinentes sont régulièrement diffusées.

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.

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.

De 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, F. Namer a été oratrice invitée à Prague en septembre 2019 pour la conférence DeriMo.
- À 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. 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.