CPER LCHN
Langues, Connaissances et Humanités Numériques

———————

Bilan 2016—2107

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.*

A. Organisation du projet

La structuration du projet en quatre axes est restée stable :

- Ingénierie des langues
- Ingénierie des connaissances
- Humanités numériques
- E-Éducation

Nous avons lancé fin 2015, puis fin 2016 deux appels à propositions en vue de soutenir des opérations scientifiques souhaitant exploiter les plateformes en cours de mise en place et nécessitant une aide financière en budget de fonctionnement ou en équipements matériels spécifiques. Le premier appel a permis de financer 8 actions, le second a prolongé 6 actions et soutenu deux nouvelles actions. Nous donnons ci-dessous les bilans de 7 des 8 projets soutenus lors de ce deuxième appel.
Le dernier projet soutenu (MultEmod) n'a pas encore reçu le matériel spécifique commandé dans le cadre de ce CPER suite à des problèmes administratifs pour établir la commande. Le projet MulEmod sera décrit dans la section suivante.

Dans le cadre de ce CPER, nous disposons également de ressources via le Feder pour recruter des ingénieurs en CDD. Jusqu'à aujourd'hui nous avons eu des difficultés à recruter des ingénieurs sur ce dispositif. C'est pourquoi nous avons demandé au Ferer une prolongation pour l'utilisation des crédits 2016 au-delà du 31/12/2017 (par courrier de la déléguée régioanle de CNRS le 1er juin 2017). Nous sommes dans l'attente d'une réponse.

Jusqu'à aujourd'hui nous avons donc recruté :

- Sur crédits 2015, Ismaël Bada (9 mois de septembre 2016 à mai 2017) : il a travaillé à l'installation des serveurs GPU commandés sur les crédits CPER (environ 2 mois), sur le projet ProsodCorpus (environ 3 mois) et sur le projet MGB (environ 4 mois).
- Sur crédits 2016, Ismaël Bada (7 mois de juin 2017 à décembre 2017) : il continue son travail sur les projets ProsodCorpus et MBG, il va également intervenir dans le projet HPPapers
- Sur crédits 2016, Simon Méoni (6 mois de juillet 2017 à décembre 2017) : il travaille sur les deux projets CoReA2D et Démonette.

Nous sommes en train de recruter deux autres ingénieurs qui seront en priorité chargé de faire du développement web. C'est en effet la demande la plus fréquente de la part des chercheurs du domaine. D'une part, nous avons besoin de site et de services ouvert vers l'extérieur pour valoriser et pour rendre plus visibles les recherches, les outils et les ressources développés dans nos laboratoires. D'autre part, nous avons besoin d'outils internes en ligne pour créer, maintenir, mettre à jour les données sur lesquelles on travaille.

B. Soutien spécifique à des actions de recherche suite à un appel à proposition largement ouvert

DEMONETTE (ATILF, porté par Fiammetta Namer)

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. Schématiquement, chaque entrée fournit des renseignements d'ordre sémantique, catégoriel, et morphologique sur Mot1 et Mot2, relativement l'un à l'autre.

Mot_1	Mot_2	Cat_1	Cat_2	Suf	Typ_1	Typ_2	Definition Mot_1	Relation
agriculteur	agriculture	Ncms	Ncfs	eur	@AGF	@ACT	agent masculin de agriculture	indirect
agriculture	agriculteur	Ncfs	Ncms		@ACT	@AGM	action pratiquée par agriculteur	indirect
agression	agresser	Ncfs	Vmn----	ion	@RES	@	résultat de agresser	descendant
agresser	agression	Vmn----	Ncfs		@	@ACT	réaliser le agression	ascendant

Figure 1 : échantillon de Démonette

La **première phase** du projet Démonette-1.3 (financement 2016-2017) était d'enrichir la couverture de la base Démonette (https://www.ortolang.fr/market/lexicons/demonette), au moyen des ressources issues du lexique Lexeur, où chaque entrée comporte un nom masculin d'agent, e.g. laveur, ainsi que le ou les unités

lexicales morphologiquement apparentées et désignant le procès réalisé par cet agent, e.g. lavage, lavement, laver.
Ces migrations ont donné lieu au calcul automatisé des propriétés dérivationnelles illustrées dans l'échantillon de la Figure 1 et ont été complétées par des entrées mettant en jeu le correspondant féminin du nom d'agent masculin, généré automatiquement (lav**euse** pour laveur, mais compétit**rice** pour compétiteur).

Dans la **phase actuelle** d'évolution de Démonette, nous poursuivons le peuplement de la base en y intégrant le contenu du Lexique dérivationnel Mordan (https://sites.google.com/site/koehlaurore/these). Cependant, en premier lieu, notre objectif est de compléter le contenu actuel par de nouvelles informations, morpho-phonologiques, inférables à partir des caractéristiques des relations constituant chacune des entrées de la base, telle que celle-ci résulte de la phase écoulée.

MGB (LORIA, porté par Irina Illina)

Les ***modèles de langage*** (ML) jouent un rôle clé dans les systèmes actuels de reconnaissance automatique de la parole. Grace à ces modèles, les phrases reconnues respectent l'enchainement correct des mots. Dans les systèmes état de l'art, le modèle de langage est une combinaison de modèles *n*-gram et de modèles fondés sur les **réseaux de neurones**. Ces deux modèles sont combinés car ils sont complémentaires. Les ML sont appris sur des corpus de textes très variés. Mais, le contenu d'un document audio est généralement fortement influencé par le domaine, ce qui peut inclure le thème, le genre (débat, documentaire, etc.) et le style d'élocution. Pour augmenter les performances du ML dans un domaine spécifique, il est nécessaire d'adapter le modèle à ce domaine. Le corpus MGB est très grand, multi-genres, multi-domaines et couvrant toute la gamme des émissions de télé. *Il est indispensable d'adapter un modèle de langage au domaine d'un document audio à reconnaître.*
Dans le cadre du stage de Master 2 de Anna Currey, nous avons développé quelques méthodes d'adaptation des *modèles n-grams* (cf. l'article de SLT 2016 ci-dessous). Les méthodes proposées permettent d'ajouter des n-grams correspondant aux nouveaux mots non vus pendant l'apprentissage du modèle de langage. Les méthodologies d'adaptation *d'un modèle de langage neuronal* peuvent être classifiées en deux grandes catégories : *modification des entrées du modèle* ou a*daptation des paramètres internes du modèle*. Dans le cas de la modification des entrées du modèle, des paramètres auxiliaires sont ajoutés à l'entrée du réseau de neurones. Ces paramètres auxiliaires peuvent être, par exemple, des informations thématiques. Cependant, l'introduction de ces paramètres auxiliaires nécessite le re- apprentissage complet de tout le modèle neuronal. L'adaptation des paramètres internes du modèle consiste à ajouter des couches complémentaires au réseau de neurones. L'apprentissage des poids de ces nouvelles couches utilise un petit corpus spécifique de données d'adaptation. Un avantage de cette méthode est que le ré-apprentissage complet n'est pas nécessaire. Nous avons proposé une méthode d'adaptation de RNN (*Recurrent Neural Network*) et actuellement nous effectuons l'évaluation de cette méthode sur les données réelles.
Concernant la **sélection de données pour l'apprentissage** d'un réseau de neurones profond, nous proposons d'utiliser un réseau de neurones profond (DNN) pour classifier les segments audio en deux catégories : sous-titres qui correspondent à l'audio et sous-titres qui ne correspondent pas. Actuellement nous analysons quelles informations, acoustiques et linguistiques, sont pertinentes pour la tâche de sélection et serviront d'entrées au DNN. Ce travail s'effectue dans le cadre du stage de Master 2 de Juan Karsten (mars-aout 2017, 5 mois). Voir la section suivante pour plus de détails sur le travail effectué.
Les **modèles acoustiques** basés sur les DNN nécessitent une grande quantité de données

d'apprentissage. Des techniques d'augmentation de données telles que l'ajout de bruit, de réverbération ou la modification de la vitesse d'élocution sont souvent utilisées pour augmenter la taille du jeu de données et la performance de reconnaissance. Le choix des techniques d'augmentation et des paramètres associés a été traité de façon heuristique jusqu'à présent. Nous avons proposé un algorithme pour pondérer automatiquement des données perturbées en utilisant une variété de techniques et/ou de paramètres d'augmentation. Les poids sont appris de manière discriminative de façon à minimiser le taux d'erreur par trame en utilisant l'algorithme standard de descente de gradient d'une manière itérative. Les expériences réalisées sur le corpus CHiME-3 indiquent une amélioration relative du taux d'erreur sur les mots (WER) de 15% (cf. l'article d'ICASSP 2017 ci-dessous). Fait intéressant, la distribution de l'ensemble d'apprentissage obtenu après pondération diffère significativement de celle de l'ensemble de test. Ces résultats restent à généraliser au corpus MGB.
Comme prévu, l'ingénieur Ismael Bada a été affecté à ce projet pour 2 mois. Il a mis en place un **système d'apprentissage multi-GPU**, car l'apprentissage en utilisant 700 heures de parole sur 1 seul GPU peut prendre plus de 2 semaines.

HP-PAPERS (LHSP, porté par Olivier Bruneau)

1. Introduction et présentation des objectifs généraux de l'opération

Dans la volonté de produire une édition critique et numérique des manuscrits, de l'œuvre éditée et de la correspondance d'Henri Poincaré, il a été décidé qu'HP–Papers qui regroupe, pour l'instant, sur une plateforme numérique l'ensemble de la correspondance de ce scientifique majeur de son époque devait se doter d'outils liés au web sémantique afin d'aider les chercheurs en histoire et philosophie des sciences à produire de nouvelles recherches sur ce savant.

Forte de l'expérience de SemanticHPST (http://www.msh-lorraine.fr/index.php?id=671), l'équipe constituée envisage de mettre en place une plateforme, d'un triple store et des outils de recherches. Pour ce faire, nous nous appuierons sur les ontologies construites par SemanticHPST et par SyMoGIH (http://symogih.org/) (Système Modulaire de Gestion de l'Information Historique).

Lors de la première année, avec l'aide de Nicolas Lasolle, élève ingénieur en 2e année de Telecom Nancy, le projet a surtout avancé sur la définition de diverses règles de transformation de requêtes SPARQL et sur leur implémentation. Pour cela, une syntaxe a été établie. Cette dernière pourra être réutilisée lors de l'implémentation d'autres règles de transformation que nous serons amenées à créer. Divers algorithmes d'application et de vérification ont été mis en place afin d'établir une mise en œuvre au sein d'un prototype. Tout ceci s'appuie d'une part d'une ontologie (partielle pour l'instant), d'une base RDFS et sur la librairie Corese.

2. Insertion dans le programme scientifique du CPER LCHN

Les besoins en terme de plateforme portent avant tout sur deux aspects :

- une machine virtuelle avec un peu d'espace disque (de l'ordre 2 à 3 To) nécessaire pour installer nos différents outils liés au web sémantique (Protegé, SPARQL End Point, etc.) et opérer nos essais ;
- un soutien en ingénierie de quelques mois afin de mettre en place ces outils et permettre d'avoir un regard extérieur à notre projet.

3. Objectifs détaillés, contexte scientifique, problématique, méthodologie, diffé- rentes tâches de recherché

Les objectifs sont les mêmes qu'au début du projet. Néanmoins, le calendrier a été boule- versé. Par conséquent, lors des années 2015 et 2016, nous avons mis l'accent davantage sur la mise en œuvre d'un prototype mettant en pratique les idées relatives aux règles de transformation de requêtes SPARQL. Une IHM java (figure ci-dessous) s'appuyant sur le moteur CORESE a été développée. Elle est opérante sur un petit corpus.

PALEOLOR (CRULH Atelier Diplomatique, porté par Christelle Loubet)

Résumé des objectifs du projet (10 à 15 lignes) :

Les trois programmes scientifiques successifs menés au sein de l'Atelier diplomatique depuis 2012 (PRINCILOR, AMPLorr, puis le projet ANR-FNR TRANSSCRIPT) lui permettent aujourd'hui de mettre à disposition des étudiants et du grand public un corpus de plusieurs centaines d'actes (plus de 500 à ce jour), en le dotant d'outils permettant sa lecture et sa compréhension dans le cadre d'un projet d'e-éducation, dont l'objectif est double :

1. Favoriser un apprentissage actif de la paléographie (science qui traite des écritures anciennes, de leurs origines et de leurs modifications au cours des temps et plus particulièrement de leur déchiffrement).
2. Faire des documents du corpus TRANSSCRIPT des points d'appui pour une meilleure connaissance de l'histoire de la Lorraine médiévale

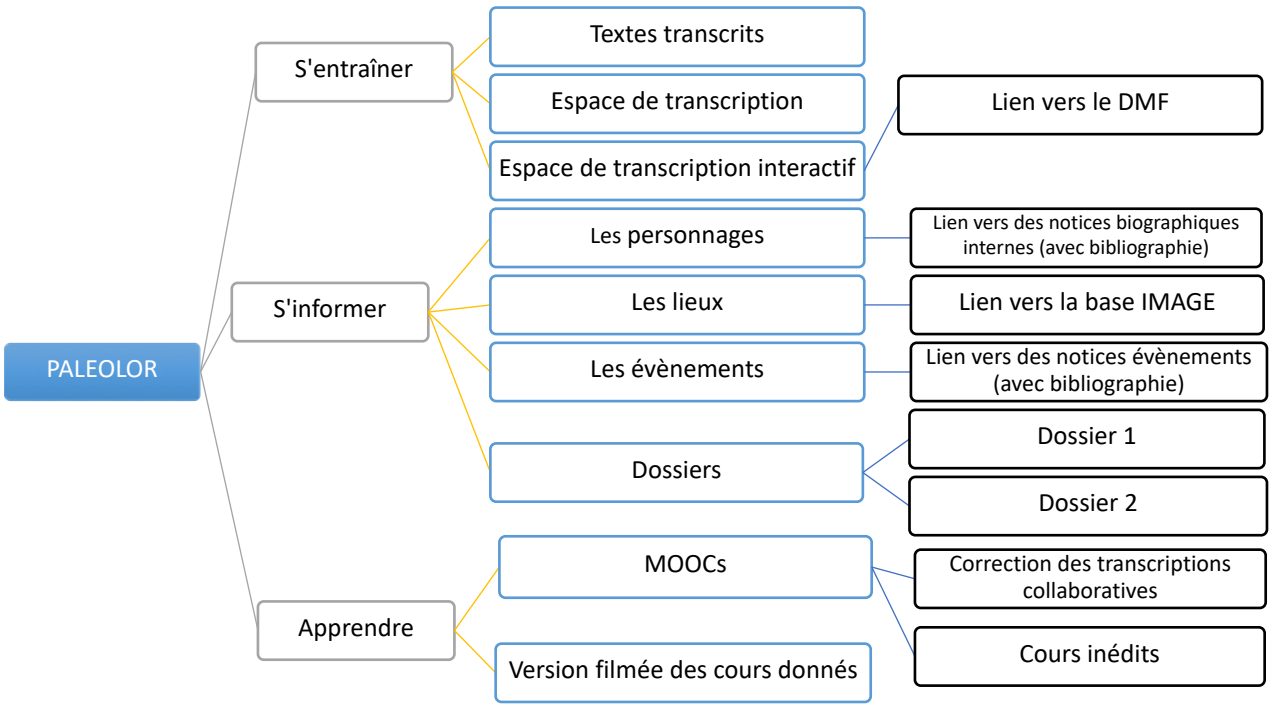
Le site devra proposer plusieurs espaces aux étudiants :

- Apprendre la paléographie : exercices de paléographie interactifs (choix des exercices par niveau de difficulté, période) ; un espace qui s'enrichira progressivement des productions d'étudiants
- Découvrir la Lorraine médiévale : grâce aux documents les plus riches du corpus TRANSSCRIPT, à partir desquels les étudiants pourront découvrir des lieux et personnages lorrains ; grâce au partenariat avec la base IMAGE

Résultats obtenus en 2016-2017 (max 1 page) :

Une enquête menée auprès des étudiants de L3 et de Master en paléographie médiévale sur le site de Nancy a tout d'abord permis de mieux cerner leurs besoins et leurs attentes en matière d'e-education (cours de paléographie et exercices en ligne, aides à l'interprétation des documents, informations sur l'histoire régionale facilitant la contextualisation des documents, etc.).

Un fois les besoins identifiés, nous avons pu élaborer un premier plan de site semblant répondre aux attentes des étudiants, tout en permettant aussi une ouverture au « grand public » :



Nous avons aussi discuté avec M. Charles Kraemer, ancien responsable scientifique de la base IMAGE de l'Hiscant-Ma, des possibilités d'interaction entre cette base de données et notre projet de site internet. Ces discussions sont bien entendu amenées à se poursuivre avec son successeur, M. Cédric Moulis.

Enfin, nous avons associé une première cohorte d'étudiants au projet en les faisant travailler tout un semestre sur les dossiers documentaires devant composer l'ossature du site (cf. effets observables).

Effets observables :

- ### Relations avec le milieu universitaire lorrain, national et international

Les étudiants de master ont travaillé au 1er semestre de l'année universitaire 2016-2017 sur des dossiers documentaires destinés à être ensuite mis en ligne dans le cadre du projet. Il s'agit à ce jour d'une quinzaine de documents transcrits et commentés qui pourront être mis en ligne dès le site internet créé. Il s'agit en fait de continuer à faire progresser les étudiants en paléographie tout en les associant à un projet scientifique et en les familiarisant à l'enjeu du numérique en SHS.

- ### Effets de levier sur les appels d'offres régionaux, nationaux et internationaux

Il existe un intérêt patrimonial non négligeable qui pourrait intéresser la région Grand Est, notamment grâce au parcours de découverte de la Lorraine médiévale. Ce parcours pourrait d'ailleurs être étendu aux deux autres espaces (Alsace et Champagne) recouverts par la région en partenariat avec les Universités de Strasbourg et de Reims.

- ### Relations avec le milieu industriel lorrain, national et international

- ### Effet de levier sur la recherche partenariale et les perspectives applicatives

À terme, le site (doublé d'une application ?) pourrait devenir un support d'apprentissage à part entière, utilisé en cours de paléographie. Les étudiants pourraient ainsi être impliqués dès la L3 dans l'incrémentation du site tout en profitant de sa dimension informative. Une application pédagogique serait ainsi le développement de l'autonomie et

de la dimension collaborative dans l'apprentissage de la paléographie.

Prosodcorpus (LORIA, ATILF, porté par Denis Jouvet)

Résumé des objectifs du projet

La prosodie porte des informations para-lexicales essentielles pour structurer le message vocal, transmettre l'état émotionnel du locuteur, préciser une emphase, etc. Les paramètres prosodiques sont l'énergie et la durée des sons, et la fréquence fondamentale. La durée des sons s'obtient à partir de la segmentation du signal vocal en phonèmes, qui résulte généralement d'un alignement automatique de la parole sur le texte correspondant. La segmentation de corpus de parole en mots et en phonèmes sert également pour d'autres applications comme la synthèse vocale, l'indexation de données vocales, …, sans oublier l'élaboration de diagnostics en apprentissage de langues.

La segmentation automatique en mots et en phonèmes et le calcul de la fréquence fondamentale, fonctionnent plutôt bien sur des signaux de parole de bonne qualité. Cependant les performances se dégradent lorsque les signaux de parole sont corrompus (signal faible, présence de bruit, paroles superposées, …), et surtout il n'existe aucune mesure précisant la fiabilité des paramètres calculés.

L'objectif du projet (pluriannuel) porte sur l'amélioration du calcul des paramètres prosodiques ; la détermination d'un indice de qualité (mesure de confiance) associé ; et l'exploitation de ces informations dans des études linguistiques. Cela conduira à des outils plus précis, plus robustes et plus fiables pour la segmentation de parole et le calcul des paramètres prosodiques.

Bilan des années 2015 et 2016

Avancement du projet de recherche / Description des résultats obtenus

Les travaux de la première année (2016) ont porté sur trois aspects : la segmentation phonétique, le calcul de la fréquence fondamentale, et l'annotation et l'étude prosodiques de particules de discours.

Segmentation phonétique.

En vue d'obtenir des frontières phonétiques plus précises deux approches fondées sur les réseaux de neurones profonds (DNN) ont été développées.

La première approche utilise un réseau LSTM (Long Short-Term Memory) pour déterminer des frontières phonétiques uniquement à partir du signal de parole paramétré.

L'autre approche consiste à évaluer les frontières issues d'un système d'alignement parole-texte pour, soit leur attribuer une mesure de confiance, soit déterminer de meilleures frontières. L'un des modèles étudiés combine LSTM et MLP (Multilayer Perceptron) et utilise un apprentissage non supervisé ; il a été évalué sur les frontières de mots dans le cadre du projet ORFEO [Serrière et al., 2016].

Fréquence fondamentale.

Un stage de master a permis de débuter une étude sur les mesures de confiance associées à l'estimation de la fréquence fondamentale (F0). Trois algorithmes de calcul de F0 ont été considérés, et plusieurs modèles à base de réseaux de neurones ont été étudiés pour estimer une mesure de confiance sur chaque valeur de F0. Les meilleures performances ont été obtenues avec un réseau LSTM [Deng et al., 2017].

Par la suite une dizaine d'algorithmes de calcul de la fréquence fondamentale ont été considérés. L'analyse des performances de détection du F0 a été menée sur un corpus de référence artificiellement bruité à différents niveaux de rapport signal-à-bruit, et sur un corpus de parole multi conditions (incluant des données bruitées). La robustesse au bruit varie notablement d'un algorithme à l'autre, et la principale cause d'erreur résulte d'une mauvaise décision voisé / non-voisé.

Particules de discours.

Les alignements parole-texte des différents corpus ORFEO (TCOF, OFROM, VALIBEL, …) ont été utilisés pour extraire des segments de parole contenant des lexèmes (tels que « bon », « quoi », « alors », « donc », « enfin », « quand même », …) qui peuvent être utilisées ou non en tant que particules discursives. L'objectif étant d'étudier les corrélations entre la réalisation prosodique de ces énoncés et la fonction discursive (ou non) de ces mots. Pour chaque mot, plusieurs centaines d'occurrences ont été extraites, avec un contexte suffisamment large (une quinzaine de mots avant et après)

pour pouvoir estimer le rôle sémantico-pragmatique des lexèmes étudiés, ainsi que les paramètres prosodiques correspondant à ces segments.

Pour trois des lexèmes étudiés, un millier d'occurrences ont été annotées manuellement pour distinguer l'emploi particule de discours ou non, et pour chaque emploi comme particule, une étiquette sémantico-pragmatique (par ex. conclusif, interruption, expressivité, etc.) a été ajoutée. Une courte liste d'étiquettes sémantico-pragmatiques a été définie spécifiquement pour chaque lexème.

Une étude statistique a été menée sur les données correspondant au lexème « bon » [Lee et al., 2016]. L'analyse des autres données annotées s'est poursuivie en 2017.

Utilisation des plates-formes de recherche mises en place dans le cadre du CPER.

Les clusters de calcul ont été utilisés, entre autres, pour :

- Le calcul de divers paramètres prosodiques, et la préparation des corpus (extraction de segments de parole) en vue des annotations des particules
- Le calcul et l'évaluation du pitch avec une quinzaine d'algorithmes de détection de pitch.

De plus, les traitements impliquant des réseaux de neurones ont bénéficié de la disponibilité de GPU sur certains clusters de calcul.

UniMETA (LORIA, INIST, porté par Jean-Charles Lamirel)

Résumé des objectifs du projet

L'objectif du projet de recherche UniMETA était celui d'étudier l'exploitation des processus d'auto-organisation en combinaison avec les techniques de traitement automatique des langues pour la génération automatique de métadonnées de contenu (indexation, résumé, réseaux sémantiques réduits associés aux textes, ...), jusqu'à leur classification non supervisée, ceci dans le contexte des données documentaires.

Le projet UniMETA-phase 2 vise à élargir le champ de ces études en menant une analyse plus approfondie des interactions et des unifications entre les méthodes de traitement linguistique de surface, les méthodes statistiques, telles que les méthodes de sélection de variables, et, les méthodes de clustering ou les méthodes équivalentes, telles que les méthodes Bayésiennes de mélange et les méthodes neuronales d'apprentissage profond (deep learning) pour la génération automatique des métadonnées. C'est une approche originale qui pourra s'appuyer sur l'étude et l'adaptation de composants nouveaux, comme ceux basés sur la maximisation des traits dans le but de pouvoir d'obtenir la plus grande flexibilité. Les contraintes des unifications qui seront étudiées sont celles de pouvoir opérer sur des données très volumineuses à une échelle proche du temps réel et de manière incrémentale, mais également d'être suffisamment polyvalentes pour couvrir à la fois le cadre de l'indexation automatique, celui du résumé automatique et celui de la génération de représentations sémantiques condensées du contenu des textes.

Les résultats obtenus lors du challenge CL-SCIsumm 2016, qui sont les meilleurs parmi ceux les approches proposées par les équipes internationales ayant participé au projet, ont été publiés dans un journal international. Ils permettent d'envisager une visibilité large au sein de la communauté internationale de traitement automatique des langues et, de manière plus globale, dans celle de l'apprentissage automatique.

Le travail réalisé a participé au succès de la candidature de Nicolas DUGUE pour les concours MCF. Le co-endrament du stage de Master d'Hazem AL ZIED a en effet produit suffisamment de résultats, dont des résultats de portée internationale, pour jouer sur la crédibilité de cette candidature.

Le travail réalisé a également corrélativement participé au succès de la candidature d'Hazem AL ZIED sur un contrat de thèse de 3 ans à l'ATILF.

L'expérience acquise en résumé automatique et en synthèse de contenu permettent de renforcer la visibilité du laboratoire comme partenaire privilégié dans les appels d'offres nationaux et internationaux lies au traitement automatique des grandes collections de documents. Elles permettent aussi de renforcer la visibilité de la région Lorraine en tant que partenaire privilégié pour les applications en traitement automatique des langues à grande échelle.

Résultats obtenus

Nous avons basé nos expérimentations en 2016 sur plusieurs corpus de référence, dont le corpus CLSciSumm. Ces corpus nous ont permis de bénéficier de résultats de référence associés à des approches classiques ou innovantes, ce qui nous a également permet d'établir, comme prévu dans le plan de travail, un «état de l'art» et de proposer une comparaison avec notre méthode.

Nous avons ensuite travaillé sur l'exploitation de notre méthode pour générer des résumés de communauté dans le cadre du challenge CL-SciSumm. Le challenge a impliqué la participation d'une vingtaine d'équipes internationales. Nous avons remporté ce challenge en fournissant par ailleurs le seul système fonctionnant de manière totalement non supervisée, donc applicable à grande échelle. Le rapport des résultats est fourni dans la référence [3]. L'avancée scientifique et expérimentale du projet est donc plus que conforme au plan.

Ce travail de recherche et d'expérimentation a fait l'objet du stage de Master de Hazem AL ZIED [2], en co-encadrement avec Nicolas DUGUE, post-doctorant dans l'équipe SYNALP.

CoReA2D (ATILF, porté par Évelyne Jacquey)

Réalisations

Le bilan de la première réunion a été de lancer le WP0 du projet, à savoir une exploration non seulement théorique (bibliographique) mais aussi active de l'existant (GateTeamware, WebAnno, Glozz et TXM) via la réalisation de mini-campagnes d'annotation.

Campagne test (mai - juin 2017) : outils en ligne (GateTeamware et WebAnno)

Objectif : Focus sur les fonctionnalités de gestion de campagne.

Outils

Abandon de GateTeamware du fait d'une absence de maintenance de cette extension de Gate et du fait de difficultés d'installation non solubles faute d'interlocuteur.

Installation de WebAnno sur une machine (ARGES) avec accès externe possible (https). L'installation a fait l'objet de l'utilisation de la philosophie des Dockers.

Corpus

9 résumés d'articles extraits du corpus TermITH sur Ortolang dans 3 disciplines de TermITH, l'archéologie, la chimie et la linguistique.

Couches d'annotation testées en annotation manuelle sur texte nu

Anotation morpho-syntaxique, annotation en occurrences de candidats et évaluation de la valeur terminologique de chaque occurrence annotée, projection du lexique transdisciplinaire du projet TermITH et projection du lexique TLF'Phraseo

L'annotation manuelle s'appuie sur des enrichissements déjà calculés dans le cadre de TermITH sur les textes. L'adaptation qui a été réalisée ici a consisté en une interface de visualisation.

Rôles

3 annotateurs (L. Kister, S. Meoni, S. Ollinger) et 1 curateur (E. Jacquey), 2 administrateurs (E. Jacquey, S. Meoni)

Déroulement

Préparation : mai 2017, rédaction de guides d'utilisation de WebAnno et de l'interface de visualisation, guides de campagne par couche d'annotation

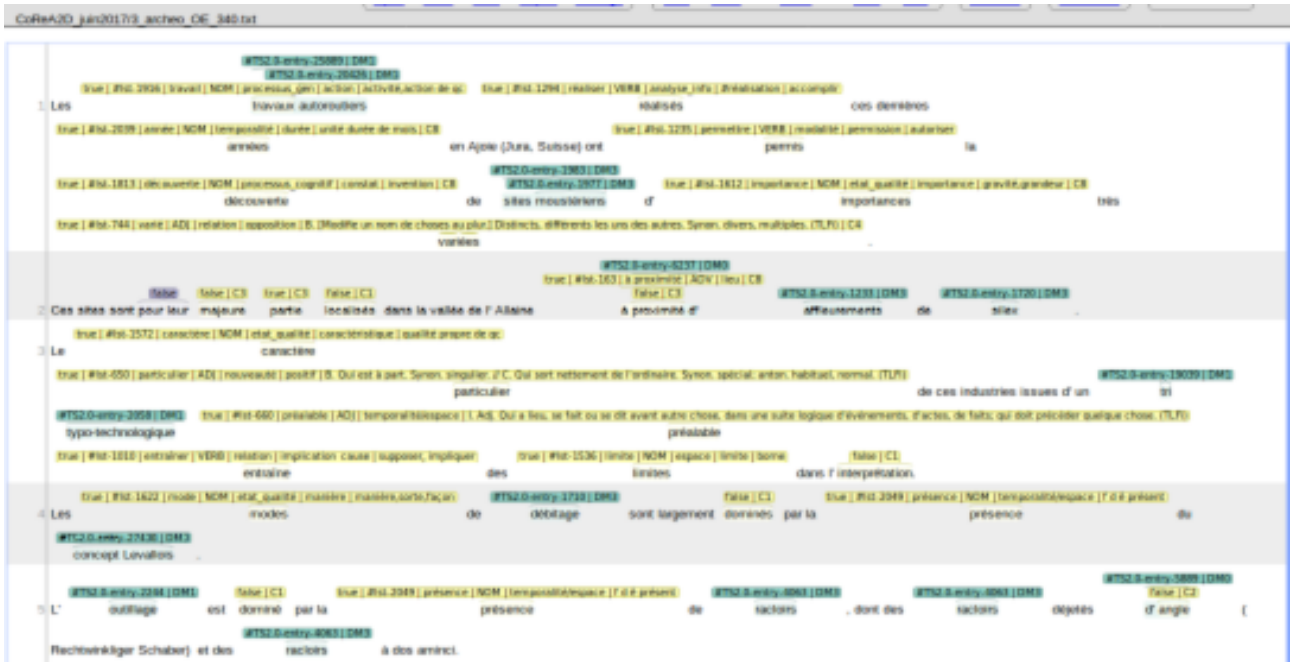
Annotation et curation : 1ère quinzaine de juin 2017

Bilan

Le bilan de cette campagne sur WebAnno est balancé.

Administration : interface très souple et très simple à maîtriser pour importer les textes à annoter et encoder les caractéristiques d'une nouvelle campagne, des bugs de ci de là cependant.

Annotation : interface dont l'ergonomie laisse à désirer à tous points de vue (couleurs dignes de l'époque Windows95, agencement des informations et retaillage des fenêtres impossibles) ; visualisation des annotations en cours en empilement et décalage spatial des caractères pouvant rendre totalement illisible le texte en cours d'annotation.



Curation : même difficulté que pour l'interface d'annotation ; accords inter-annotateurs toujours présentés deux à deux, l'accord global devrait être calculé indépendamment de l'outil.

	evelyne	laurence	sandrine	simon
evelyne	-	0.58	0.72	0.54
laurence	136/210	-	0.46	0.53
sandrine	117/138	119/210	-	0.37
simon	133/140	137/210	116/140	-

Annotation morph-syntaxique : abandon

Occurrences terminologiques : accords (Fleiss) 2 à 2 de 0,37 à 0,72

Occurrences de phrasèmes de langue générale (TLF'Phraseo) : accord parfait sur la qualité de la reconnaissance d'un phrasème (Kappa à 1,00) et accord de bon à parfait sur le choix du 'bon phrasème' en fonction de son sens tel que décrit dans TLF'Phraseo (Fleiss de 0,86 à 1,00).

Occurrences du lexique transdisciplinaire : accord faible à bon sur la qualité de la reconnaissance de l'unité lexicale transdisciplinaire (Kappa de 0,38 à 0,70) et accord moyen à bon sur le choix du lexème trandisciplinaire parmi les possiblers (Fleiss de 0,66 à 0,78).

C. Mise en place de plateformes d'expérimentation en support de chacun des axes structurants du projet

Équipements sur crédits 2015

• **Plateforme « Humanités Numériques » (69 816 €)**

Compte tenu des investissements effectués à l'ATILF lors du dernier CPER, l'accent a été mis sur la consolidation de cette plateforme en termes de système de stockage et de sauvegarde pour un montant de 65 476 €, accompagné de deux serveurs de moyenne puissance (PowerEdge R730 Server) pour un total de 4 340 €, soit un investissement total de 69 816 €.

Le système de stockage et de sauvegarde (ProDeploy Dell Storage SC Disk Series 200/220), d'une capacité de 2 To a été dimensionné pour pouvoir répondre à l'ensemble des demandes de gestion de corpus de cet axe « humanités Numériques ».

Quant aux deux serveurs de moyenne puissance (PowerEdge R730 Server), ils devraient permettre, du moins au début de ce projet, de répondre aux demandes de création de machines virtuelles d'expérimentation pour les actions scientifiques du projet qui en feraient la demande. Ces matériels ont été commandé sur le marché CNRS et sont opérationnels.

• **Plateforme « Ingénierie des langues » (59 502 €)**

Pour cette plateforme « Ingénierie des Langues » nous avons décidé l'achat de 6 serveurs de calcul GPU. Chacun de ces 6 serveurs est équipé de deux cartes GPUs Nvidia Tesla K40, qui sont des cartes GPU professionnelles destinées à effectuer du calcul intensif et qui sont indispensables pour les recherches dans le domaine émergent du deep-learning. Chaque carte GPU Nvidia Tesla K40 coûte environ 2 000 € ; nous avons donc acheté deux cartes GPU par serveur. Chaque serveur coûte donc 8 259 €. Le montant total d'investissement pour cette plateforme à 59 502 € (d'autres équipements sont nécessaires : rack, cartes réseau, …). Le matériel est installé au LORIA et est fonctionnel depuis octobre 2016.

• **Autres investissements d'équipement (14 500 €)**

Des équipements spécifiques et des postes de travail ont été financés en soutien aux projets de recherche évoqués plus loin (partie fonctionnement).

Équipements sur crédits 2016

• **Plateforme « E-éducation » (43 170 €)**

L'équipement de cette plateforme consiste d'une part en du matériel de recueil de données utilisateurs qui sera utilisé dans les différents projets de l'axe E-éducation. Il s'agit de différents types de matériel d'eye-tracker (Tobii glasses et Tobii bar X2-60, 38 170 €) et l'un labo mobile (22 000 €) qui permet de réaliser le travail d'enquête auprès de utilisateurs en dehors du laboratoire et donc dans de meilleures conditions. Un équipement de type GameLab (5 000 €) est également prévu pour les projets concernant les serious games.

• **Compléments plateforme « Ingénierie des langues » (36 004 €)**

L'usage des clusters de GPU se développe rapidement et les GPU achetés sur les crédits 2015 sont très largement utilisés. Pour répondre aux demandes croissantes des chercheurs, il faut continuer à élargir l'offre en termes de matériel GPU. Deux nouveaux serveurs sont en commandes ainsi que du matériel complémentaires (switch réseau, cablages, …).

• **Compléments plateforme « Humanités Numériques » (22 690 €)**

Pour cette plateforme, un nouveau serveur Dell (4 810 €) est prévu ainsi que l'achat de licences pour le logiciel VMWare qui permettra d'exploiter au mieux la plateforme et de répondre efficacement aux besoins des chercheurs en ce qui concerne la création de machines virtuelles et l'accès à des espaces de stockage pour les corpus.

• **Autres investissements d'équipement (31 000 €)**

Un système d'acquisition 3D performant est prévu pour équiper un articulographe présent au LORIA (17 000 €). Ce matériel va permettre le recueil de données de bien meilleure qualité pour les recherches utilisant cet articulographe comme le projet de tête parlante expressive.

D'autres équipements spécifiques et des postes de travail seront financés en soutien aux projets de recherche évoqués ci-dessous.

Nous n'indiquons ci-dessous que les publications des actions de recherche ayant reçu un soutien financier direct du projet LCHN.
A ces productions il convient d'ajouter les publications dans les thématiques du projet de l'ensemble des laboratoires impliqués dans le projet : LORIA (UMR 7503), ATILF (UMR 7118), INIST (UPS 76), LHSP-Archives Henri-Poincaré (UMR 7117), CREM (EA 3476), CRUHL (EA 3945), LIS (EA 7305), 2L2S (EA 3478), LISEC (EA 2310), PErSEUs (EA 7312), LCOMS (EA 7306),

Productions scientifiques liées directement aux soutiens reçus dans le cadre du projet LCHN

2015

Hathout, N. and F. Namer. *La base lexicale morphologique du français Démonette1.1*. Nancy - Toulouse, https://www.ortolang.fr/market/lexicons/demonette.

Bruneau, Olivier, Serge Garlatti, Muriel Guedj, Sylvain Laubé, et Jean Lieber. *SemanticHPST : Applying Semantic Web Principles and Technologies to the History and Philosophy of Science and Technology*, dans Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron- Zucker, et Antoine Zimmermann (Éds.), The Semantic Web : ESWC 2015 Satellite Events, vol. 9341 de Lecture Notes in Computer Science : Springer International Publishing, p. 416–427

2016

Hathout, N. and F. Namer. *Enriching the Démonette morpho-semantic network: computational and linguistic issues*. International Morphological Meeting 17: Workshop on Computational methods for descriptive and theoretical morphology, Vienna.

Hathout, N. and F. Namer. *A Multi-level Paradigm-based Model of Competition in Word Formation.* 17th International Morphology Meeting, Vienna.

Hathout, N. and F. Namer. *Paradigms in word-formation: new perspectives on data description and modeling.* Workshop : SLE 2016, Naples.

Hathout, N. and F. Namer. *Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French.* Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, European Language Resources Association (ELRA).

Hathout, N. and F. Namer. *Modeling Meaning-Form Discrepancy in Word Formation within ParaDis, a Four Level Paradigm-based Modular Framework AnaMorphoSys : Analyzing Morphological Systems*, 20-22 juin 2016, (lien : http://www.univ-lyon2.fr/culture-savoirs/podcasts/anamorphosys-analyzing-morphological-systems-704334.kjsp?RH=podcasts), Lyon, ASLAN, CNRS.

Anna Currey, Irina Illina, Dominique Fohr. *Dynamic adjustment of language models for automatic speech recognition using word similarity*. IEEE Workshop on Spoken Language Technology (SLT 2016), Dec 2016, San Diego, CA, United States.

Bartkova, Katarina , Bastien, Alice and Dargnat, *How to be a discourse particle ?* Speech Prosody 8, Boston University, May 31-June 3, 2016.

L. Lee, K. Bartkova & M. Dargnat. *Particules discursives et indices prosodiques en français*. Journées franco-suisses, Neuchâtel, octobre 2016.

G. Serrière, C. Cerisara, D. Fohr & O. Mella. *Weakly-supervised text-to-speech alignment confidence measure*. Proc. COLING'2016, 26th International Conference on Computational Linguistics, décembre 2016, Osaka, Japan. <hal-01378355>.

Hazem AL ZIED, Nicolas DUGUE, Jean-Charles LAMIREL, *Automatic summarization of scientific publications using a feature selection approach*, International Journal on Digital Libraries, Special Issue on CL-SCISumm 2016 Challenge – JCDL 20166 Conference

Lamirel J.-C., *Performing and visualizing temporal analysis of large text data issued for open sources : past and future methods*, 12th IEEE International Conference : Beyond Databases, Architectures and Structures (BDAS'2016), Krakow, Poland, May 2016.

Jean-Charles Lamirel, Nicolas Dugué, Pascal Cuxac: *New efficient clustering quality indexes*. Procedings of IJCNN 2016, pp 3649–3657.

Padroni, S., Demily, C., Franck, N., Bocéréan, C., Hoffmann, C., Musiol, M. *Ajustement comportemental et mouvements de saccades oculaires dans la schizophrénie*. L'évolution psychiatrique 81 (2016) 365–379.

2017

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

Hathout, N., Lignon, S. et Namer F. *Morphophonologie et paradigmes dans la base dérivationnelle Démonette*. ISMO - First International Symposium of Morphology - Lille, France, December, 13-15th 2017 Université de Lille3 - Villeneuve d'Ascq.

Sunit Sivasankaran, Emmanuel Vincent, Irina Illina. *Discriminative importance weighting of augmented training data for acoustic model training*. 42th International Conference on Acoustics, *Speech, and Signal Processing (ICASSP 2017), Mar 2017, New Orleans, United States.*

Sheikh, I., Illina, I., Fohr, D., Linarès, G. *Improved Neural Bag-of-Words Model to Retrieve Out-of-Vocabulary Words in Speech Recognition*. Dans Proceedings of Interspeech, 2016.

B. Deng, D. Jouvet, Y. Laprie, I. Steiner & A. Sini. *Towards confidence measures on fundamental frequency estimations*. Proc. ICASSP'2017, 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, USA, mars 2017.

K. Bartkova, M. Dargnat, D. Jouvet & L. Lee. *Annotations de particules de discours en français sur une large variété de corpus*. Atelier ACor4French - Les corpus annotés du français ; TALN'2017, Traitement Automatique des Langues Naturelles, Orléans, juin 2017.

D. Jouvet & Y. Laprie. *Performance analysis of several pitch detection algorithms on simulated and real noisy data*. In EUSIPCO'2017, 25th European Signal Processing Conference, Kos Island, Greece, Août 2017.

D. Jouvet, K. Bartkova, M. Dargnat & L. Lee. *Analysis and automatic classification of some discourse particles on a large set of French spoken corpora*. In SLSP'2017, 5th International Conference on Statistical Language and Speech Processing, Le Mans, Octobre 2017.

O. Bruneau, E. Gaillard, N. Lasolle, J. Lieber, E. Nauer, J. Reynaud, *A SPARQL Query Transformation Rule Language — Application to Retrieval and Adaptation in Case-Based Reasoning.* In : Aha D., Lieber J. (eds) Case-Based Reasoning Research and Development. ICCBR 2017. Lecture Notes in Computer Science, vol 10339. Springer, Cham, p. 76-91.

Jean-Charles Lamirel, Nicolas Dugué, *A new promising approach for clustering quality evaluation*, Int J Applied Intelligence

Nicolas Dugué, Jean-Charles Lamirel, *Une métrique de sélection de variables appliquée à la centralité et à la détection des rôles communautaires*, Long paper, Proceedings of EGC 2017, Grenoble, France.

Jean-Charles Lamirel, *Feature maximization metric and its application to large data*, 5th CMM Pucón Symposium: Data Science for Frontier Astronomy, Biology and Medicine, Puerto Varas, Chile (2017).