CPER LCHN
Langues, Connaissances et Humanités Numériques

Bilan 2015

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.*

L'année 2015a principalement été consacrée à la mise en place du projet. Il convient en effet de noter que les aides régionales et FEDER 2015 pour ce projet n'ont été notifiées que le 9 novembre 2015, les conventions correspondantes signées en décembre 2015 pour une mise en place effective des crédits, compte tenu du changement de logiciel de gestion du CNRS, qu'en mars 2016.

Néanmoins le projet s'est mis en place conformément à ce que nous avions proposé lors de la soumission du projet à savoir :

- Structuration en quatre axes complémentaires : ingénierie des langues, Extraction et structuration de connaissances, Humanités Numériques et E-éducation.
- Soutien spécifique à des actions de recherche suite à un appel à proposition largement ouvert.
- Mise en place de plateformes d'expérimentation en support de chacun des axes structurants du projet.

A. Structuration de quatre axes complémentaires

Ingénierie des langues :

Animateurs : Christophe Cerisera (LORIA) & Alain Polguère (ATILF)
Laboratoires impliqués : ATILF, INIST, LORIA

L'axe Ingénierie des Langues du CPER LCHN a pour objectif de soutenir et de structurer la recherche en linguistique et en traitement automatique des langues et de la parole en favorisant le développement de nouvelles collaborations multidisciplinaires entre chercheurs et en mettant en place des moyens matériels conséquents qui permettent de progresser significativement dans les axes de recherche en cours et de stimuler et d'initier de nouvelles thématiques de recherche qu'il n'aurait pas été possible de mener à bien autrement.

Pour ce faire, un des outils les plus efficaces est le soutien à des projets de recherche ciblés qui répondent précisément aux attentes de l'axe en évitant le risque de disperser les fonds sur de nombreuses thématiques individuelles. La politique scientifique de l'axe IL s'est donc structurée d'une part autour de cet appel à projets, avec pour objectif de sélectionner les propositions de recherche les plus ambitieuses, et d'autre part, autour de la mise en place d'une plateforme matérielle apportant de nouvelles ressources de calcul GPU adaptées aux traitements des très grandes masses de données (Big Data) langagières qui constituent aujourd'hui le défi principal à relever en ingénierie des langues, et surtout adaptées aux méthodes scientifiques les plus récentes et les plus efficaces qui permettent de traiter efficacement ces données.

L'animation scientifique de l'axe IL a donc été balisée par les jalons suivants:

- Organisation d'une réunion d'information et de démarrage de l'axe IL le 19 novembre 2015.
- 2 décembre 2015: appel à proposition de projets. Cet appel à été coordonné au niveau du CPER LCHN et non de l'axe IL, afin de favoriser la proposition de projets scientifiques multidisciplinaires interaxes. De fait, 2 dossiers parmi les 7 propositions impliquées dans l'axe IL émargeaient également à un autre axe du CPER LCHN.
- 2 mars 2016: Tenue d'un comité de sélection des projets retenus. 11 propositions ont été remontées, parmi lesquelles 7 dossiers émargeaient à l'axe IL. Sur ces 7 propositions, 5 ont été retenues et soutenues par l'axe IL. Suite à cette réunion, les projets retenus ont pu débuter. Il a également été possible de définir, en fonction des objectifs scientifiques mis en avant dans les projets retenus, un profil de poste pour un ingénieur en soutien à ces projets. 7 mois d'ingénieur ont été affectés ainsi à l'axe IL. De même, il a été possible d'affiner alors les besoins et les spécifications de la plateforme de calcul GPU.
- 18 mars: Diffusion du profil d'ingénieur en informatique de 7 mois via les différents réseaux, notamment la plateforme d'embauche du CNRS.

Au final, cinq projets de recherche ciblés participant à l'axe IL ont été lancés en mars 2016, parmi lesquels trois projets avec une forte composante informatique pour le TAL:

- MGBChallenge, porté par Irina Illina du LORIA
- ProsodCorpus, porté par Denis Jouvet du LORIA
- UniMETA, porté par Jean-Charles Lamirel du LORIA

et deux projets avec une dominante linguistique:

- Démonette 1.3, porté par Fiammetta Namer, chercheuse à l'ATILF
- ITL-DI-Oeil, porté par Michel Musiol de l'ATILF

Les objectifs et progrès de chacun de ces projets sont résumés dans la suite de ce rapport après le bilan général des 4 axes.

Au-delà de la mise en place des projets scientifiques ciblés, le deuxième volet important de l'axe IL a concerné la définition des spécifications et l'achat de la plateforme de calcul GPU. Cette plateforme devant soutenir la recherche à moyen et long terme de l'axe IL du CPER LCHN, nous avons décidé de la définir bien entendu en fonction des besoins exprimés par les lauréats des cinq projets ciblés de recherche de l'axe IL, mais aussi en élargissant la concertation à un public plus vaste de chercheurs potentiellement intéressés par cette plateforme et susceptibles d'intervenir au cours des années suivantes du CPER LCHN. Nous avons pour cela constitué un groupe de travail ouvert à tout chercheur intéressé par les méthodes d'apprentissage profond appliquées à nos domaines de recherche.

Ce groupe a été créé en décembre 2015, et a rapidement attiré de nombreux chercheurs, essentiellement au LORIA, mais également à l'extérieur du LORIA, pour atteindre aujourd'hui 84 membres. Une mailing-list a été créée, ainsi qu'un site web (http://deeploria.gforge.inria.fr) sur lequel nous pouvons centraliser les articles, logiciels, liens les plus intéressants pour nous. Ce groupe est animé par des réunions plénières, environ une toutes les 6 semaines, où nous avons pu réaliser un tutoriel sur l'apprentissage profond, mais aussi exposer les travaux récents des membres du groupe sur cette thématique. Nous avons également profité des 40 ans du LORIA pour assister à deux séminaires de personnalités de renom du domaine: Yann Lecun, directeur de la recherche chez Facebook, professeur au collège de France et l'un des fondateurs de l'apprentissage profond, et Sander Dielemann, chercheur à Google DeepMind, qui a participé à la conception du logiciel AlphaGo exploitant l'apprentissage profond.

Ce groupe de travail est en lien étroit et direct avec les objectifs scientifiques de l'axe IL du CPER LCHN, car les méthodes d'apprentissage profond sont aujourd'hui les approches de loin les plus performantes dans quasiment tous les domaines du traitement des langues et des connaissances, elles sont dédiés au traitement de grandes masses de données et ne peuvent être réalisées que grâce à des serveurs de calcul GPU du type de ceux constituant la plateforme IL. La mise en place de cette plateforme a suivi les jalons suivants:

- 25 janvier 2016: consultation des membres du groupe de travail sur l'apprentissage profond pour définir au mieux les caractéristiques de la plateforme de calcul GPU à commander. La consultation a notamment porté sur le type de cartes GPU et le nombre de cartes GPU par serveur. Une nette préférence pour des cartes de type Nvidia Titan X a émergé, ainsi que pour une configuration de deux cartes par serveur.
- Fin février 2016: Multiples contacts avec d'une part les responsables du cluster Grid5000, et d'autre part, avec les moyens informatiques du LORIA pour étudier les diverses possibilités d'hébergement des machines GPU, afin qu'elles soient accessibles à l'ensemble des partenaires du CPER. Au final, deux solutions possibles ont été retenues: soit une intégration dans Grid5000, soit une mise à disposition dans la section DMZ du LORIA.
- Fin mars 2016: Après consultation des différents partenaires du CPER et une simulation de budget dédié à la maintenance des serveurs, il apparaît que les moyens humains affectés au CPER IL sont insuffisants pour assurer la maintenance, et la solution de l'intégration dans Grid5000 est donc choisie.
- Avril 2016: Nouvelle prise de contact avec les commerciaux de DELL pour affiner la configuration souhaitée et garantir les serveurs sur le plan technique; les cartes Titan X se révèlent incompatibles lors des simulations de configuration, et nous optons donc pour une solution de remplacement avec des cartes professionnelles Tesla K40.
- Début Mai 2016: La validation technique des serveurs est terminée, nouvelles réunions avec les responsables de Grid5000 pour définir les configurations réseaux requises, et avec les moyens informatiques du LORIA pour définir précisément les conditions d'hébergement des serveurs. Après plusieurs aller-retour par email et avoir envisagé plusieurs solutions alternatives, nous convergeons finalement vers un hébergement en salle B056 du LORIA, ce qui permet de définir précisément les armoires et matériels réseaux nécessaires.
- 18 mai: réception de 3 devis fournis par DELL et Axians; ces devis sont transmis aux responsables du CPER fin mai pour validation budgétaire, et les commandes sont passées dans la foulée.
- 17 juin: arrivée des premiers composants au LORIA. L'installation physique est prévue fin juin, et l'installation logicielle suivra.

Extraction et structuration de connaissances

Animateurs : Yannick Toussaint (LORIA) & Laurent Schmitt (INIST)

Laboratoires impliqués : ATILF, INIST, LORIA

Dans une société submergée par la profusion d'informations, capitaliser les connaissances est devenu un besoin omniprésent qui se décline de différentes façons selon qu'il s'agisse d'experts confrontés à un problème pointu ou d'un système automatique d'aide à la décision. Il est donc crucial de pouvoir disposer, maintenir et diffuser des

terminologies et référentiels d'acteurs à jour par domaine, de pouvoir détecter des évolutions et des ruptures dans des domaines scientifiques ou technologiques, d'être capable de synthétiser sous forme de connaissances des informations dispersées dans plusieurs milliers de textes ou, au contraire, d'être capable de trouver les nouvelles connaissances parfois très spécifiques qui permettront d'améliorer la réponse d'un système à base de connaissances. Notre particularité en Lorraine est de tirer profit de nos compétences en IST (INIST), en linguistique (ATILF) et en informatique (LORIA) pour répondre à ces questions pluridisciplinaires, de pouvoir ainsi associer traitement de grands volumes de textes, analyse de données ouvertes (*Linked Open Data, Big Data*) et intégration de connaissances existantes (*Open Ontologies*).

Le traitement de l'information se fait généralement par des boucles successives, partant d'une information faiblement structurée et en l'enrichissant progressivement pour la coder dans des structures complexes qu'il est alors possible de fouiller ou de classifier. Après validation par des experts, les modèles construits à partir des données produisent des connaissances réinjectables dans d'autres systèmes. Pour répondre à la diversité des problèmes posés et à la variabilité liée aux domaines d'expertise, nous ferons appel à différents types d'approches :

L'enrichissement de l'information ou des données initiales fait appel à des mécanismes d'annotation par des métadonnées, par des informations d'ordre linguistique (morphologique, syntaxique ou sémantique), par des informations contextuelles (cocitations, bibliographie), ou encore par des connaissances existantes et relatives au domaine sur lequel elles portent (terminologies, *linked open data*…). Ces annotations peuvent être produites à partir de ressources spécifiques (thésaurus), être le résultat d'outils comme des analyseurs syntaxiques (Leopar, *Stanford parser)*, de méthodes d'apprentissage (*Active Learning*) ou d'outils de fouille de données. Les enrichissements relatifs au domaine supposent le repérage des entités de base dans le domaine considéré par des méthodes d'extraction terminologique, de choix de candidats-termes et de désambiguïsation entre les usages terminologiques et non-terminologiques. Ces enrichissements nécessitent également la détection d'entités nommées de type varié (dates, lieux, adresses internet, personnes, laboratoires, institutions, projets, etc.). Les productions scientifiques possèdent des informations particulières qui permettent de les organiser en réseaux que ce soit par leurs citations respectives ou par les rédactions entre coauteurs. Traiter, structurer et normaliser ces informations permettront d'accroître le potentiel d'exploitation des connaissances enfouies dans les textes. Les méthodes et outils développés nécessitent la définition et la structuration de référentiels de différents types (structures, acteurs, entités nommées, terminologies). En particulier, l'évolution des ressources terminologiques de type langage documentaire, disponible à l'INIST, vers des ressources de type thésaurus voire ontologies passe par un important travail de normalisation et de structuration préalables nécessaires à l'exposition et à la réutilisation de ces ressources. Compte tenu de la diversité et de la volumétrie des ressources concernées, la problématique principale posée par ce chantier est la définition de procédures et/ou le développement d'outils d'aide à la structuration des corpus terminologiques pouvant avoir des origines multiples qui devrait pouvoir être diffusés sur la plateforme ORTOLANG.

Les approches numériques pour la représentation et la classification de textes (textométrie, clustering supervisé ou non) sont très utilisées dans le domaine de l'IST ou de la recherche d'information, mais également en sciences humaines (linguistique, sociologie) par exemple, pour caractériser des types de textes ou d'usages. Les méthodes de classification dynamiques sont issues de travaux sur le *datastream*, mais elles sont peu adaptées et peu fiables pour identifier les tendances fines. Le développement de nouvelles méthodes incrémentales de détection de nouveauté et de méthodes diachroniques qui s'adaptent à l'analyse précise de l'évolution des sujets dans de grands corpus textuels dynamiques est aujourd'hui un problème ouvert et représente de fait un challenge très important : le challenge est d'assurer leur efficacité sur des documents complets d'origine et de structures diverses (données bibliographiques, articles scientifiques, brevets…).

L'extraction de connaissances et leur structuration permet de capitaliser et synthétiser des informations ou connaissances dispersées dans différents documents, comme des articles scientifiques du domaine médical décrivant des patients atteints de certaines maladies. Une information est essentiellement utilisée par des processus de recherche d'information ou par des outils de classification pour en faciliter l'accès ou l'analyse. Une connaissance doit avant tout servir à raisonner, que ce soit un raisonnement humain ou un raisonnement fait par une machine. Construire des connaissances suppose de mettre au point des outils de fouille de données comme l'extraction de motifs (motifs fréquents, rares, motifs séquentiels) ou encore de motifs de graphes permettant de traduire une partie de la complexité des phénomènes observés. Ces motifs correspondent à des éléments du domaine et possèdent des propriétés. De plus, ils sont en relation avec d'autres motifs et ces relations correspondent à des relations entre objets du domaine. Des méthodes de

classifications formelles comme l'Analyse Formelle de Concepts peuvent alors créer un ensemble de concepts, structurés par un ordre partiel, point de départ pour la construction d'une ontologie. Enfin, pour être utilisés par des agents humains ou logiciels, en particulier dans le cadre du web sémantique, ces motifs doivent être représentés à l'aide de langages de représentation des connaissances adaptés comme, par exemple, les langages RDFS ou OWL-DL, ce dernier s'appuyant sur les logiques de descriptions et les modes de raisonnement associés.

Les systèmes à bases de connaissances exploitent les connaissances et mettent en œuvre des processus de raisonnement. Ces connaissances peuvent être injectées dans un processus de fouille de donnée pour le guider. Mais, il peut également s'agir de systèmes déductifs ou inductifs, de système de raisonnement à partir de cas, par exemple pour résoudre des problèmes d'aide à la décision. De plus en plus d'outils exploitent aujourd'hui des connaissances, qu'elles soient acquises spécifiquement à partir de données ou qu'elles proviennent de données du web sémantique ("linked open data").

Notre compétence lorraine autour de l'extraction et la structuration de connaissances que ce soit pour la valorisation de données ou de textes s'appuie également sur les travaux réalisés dans les autres axes de ce thème. Elle est réellement originale au niveau national et international.

Notre objectif est ici de développer une plateforme permettant de valoriser les travaux théoriques et appliqués menés dans les équipes de l'ATILF, de l'INIST et du LORIA. Cette plateforme doit aussi prolonger les travaux menés dans le cadre de plusieurs projets nationaux (ISTEX, BSN) et des ANR TermiTH, KolFlow, Hybride… Elle rendra ainsi accessible à des acteurs académiques ou industriels des outils d'enrichissement de textes, d'analyse de l'information, de visualisation et de structuration de connaissances paramétrables et adaptables à de nouveaux problèmes et à de nouveaux domaines d'application. Cependant, comme nous l'avons souligné, le processus de construction de connaissances est constitué de passes successives partant de données brutes et les enrichissant progressivement. Cela signifie donc que la plateforme va devoir mettre en avant, d'une part, des outils simples – sortes de briques élémentaires – pouvant être testés, en ligne, sur différentes données académiques, économiques ou industrielles, et, d'autre part des scénarios, potentiellement plus complexes, illustrant un cadre plus spécifique d'application.

Suite à l'appel à propositions de soutien d'actions de recherche lancé au cours de cette année, un projet relevant de cet axe et de l'axe « ingénierie des langues » a été retenu pour un financement spécifique (cf. bilan de ces projets à la suite du bilan général des quatre axes structurants du projet LCHN), il s'agit du projet UniMATA d'Extraction Unifié de Métadonnées par Auto-organisation à partir du plein texte.

Humanités Numériques

Animateurs : Christophe Benzitoun (ATILF) & Luc Massou (CREM)
Laboratoires impliqués : ATILF, CREM, CRULH, LCOMS, LIS, LORIA, LPHS, 2L2S)

Les « Humanités Numériques s'intéressent : (i) à la constitution, l'édition et le stockage contrôlé de données numérisées (ii) à la représentation d'informations liées à ces données (iii) à leur diffusion et leur valorisation via des interfaces d'interrogation adaptées aux données, mais aussi aux informations recherchées (iv) aux effets de la dissémination des technologies numériques via des outils grand public, générant des usages renouvelant les formes de socialisation et les pratiques sociales, et donnant un accès inédit à des masses de données publiques.

Le dernier CPER « Hommes et Société », au travers en particulier du projet « Langues textes et document », a permis à des équipes SHS lorraines de se positionner en ces domaines, en s'appuyant, entre autres, sur la plateforme du CNRTL aujourd'hui intégré dans l'Equipex ORTOLANG. Nous poursuivons cet effort en structurant une plateforme matérielle et logicielle de gestion de données et d'expérimentation recherche intégrant : serveurs de données sécurisés, sauvegarde sur supports externes des données, redondance de sécurité au travers d'un point de sauvegarde physiquement séparé de la salle des serveurs de données, logiciels pour la gestion de la plateforme (licence site ou logiciel spécialisé), scanner à plat de haute qualité, mise à niveau de l'infrastructure réseau et salle d'expérimentation pour l'observation des usages.

Une telle plateforme permettra de développer quatre problématiques incontournables en ce domaine :

(a) La numérisation des données. Elle s'appuiera sur deux outils que le dernier CPER a mis en place : une plateforme de numérisation de documents anciens (numérisation verticale) et une chaîne d'OCRisation de documents écrits. Ces deux outils continueront d'être utilisés dans le présent CPER et ne nécessiteront pas de développements importants en dehors d'adaptations ponctuelles en fonction des nouveaux types de données à traiter. Sur le plan des données orales dans leur forme transcrite, anonymisée et librement disponible (site

du CNRTL et d'ORTOLANG), la base « Traitement de Corpus Oraux en Français » (TCOF), héritage du dernier CPER, continuera de se développer en collaboration avec le projet ANR ORFEO auquel nous participons.

(b) La représentation des informations ajoutées aux données par les chercheurs qui les étudient. Ces informations se subdivisent en deux grandes problématiques : (i) les métadonnées qui décrivent des documents, des corpus ou des bases de données dans leur ensemble (époque, disponibilité, etc.) et (ii) les annotations qui sont des informations ajoutées à l'intérieur des données elles-mêmes (étiquette grammaticale, catégorie sémantique, etc.). Les métadonnées vont guider la manière dont les données numérisées peuvent être interrogées en sous-ensembles contrôlés et pertinents pour le monde de la recherche, de la culture, de l'industrie ou le grand public. Les annotations sur les données, quant à elles, favorisent, d'une part, l'accumulation des savoirs (historiques, patrimoniaux, culturels, sociologiques, linguistiques, etc.) au fur et à mesure des recherches menées et, d'autre part, le dialogue et la coopération entre savoir-faire issus de différentes disciplines sur des objets partagés (données et leur représentation).

(c) Les systèmes d'interrogation et de diffusion des données numérisées qui seront produits dans le cadre du présent CPER auront à charge d'une part d'aider leurs utilisateurs à constituer un sous-ensemble pertinent de données à partir des métadonnées d'une ou de plusieurs bases de données, et d'autre part, de fournir des outils de visualisation à la fois simples et efficaces des données enrichies par leurs annotations.

(d) L'observation des usages numériques. En se fondant sur le constat que les technologies numériques de l'information et de la communication deviennent progressivement des organisateurs incontournables de nos actions quotidiennes, il s'agira d'étudier les bouleversements sociaux et sociétaux induits, la mise au jour des impacts culturels de ces nouveaux supports, et la compréhension de cet univers de pratiques fait d'innovations techniques et d'appropriations sociales permanentes. Plus précisément nous proposons d'étudier leurs logiques d'évolution dans la durée du CPER grâce à une enquête pluri-méthodologique (entretiens, observations, questionnaires) et pluridisciplinaire (sciences de l'information et de la communication, sociologie, informatique, sciences du langage) sur plusieurs terrains innovants relevant de l'ingénierie de la connaissance : les TICE, le webjournalisme, les jeux numériques et les publications scientifiques en ligne. Avec là aussi des développements informatiques sur la visualisation dynamique des données d'usage.

Ces différentes problématiques sont développées au sein des laboratoires participants au travers de plusieurs projets tels :

Projet « Mémoires lorraines » porté par le l'ATILF qui a pour objectif de constituer et valoriser une ressource patrimoniale numérique de textes liés à la mémoire de la Lorraine, interrogeable par un moteur de recherche rassemblant un corpus d'écrits personnels (journaux, correspondances et autobiographies), écrits par des Lorrains ou par des personnes ayant séjourné en Lorraine. Un intérêt particulier sera porté aux textes rédigés dans des circonstances historiques remarquables, notamment la Grande Guerre et la Seconde Guerre mondiale. Dans les limites du droit d'auteur, ce corpus sera diffusé, sous forme d'extraits ou de textes intégraux, via le CNRTL et ORTOLANG, la base de données Frantext et la BNR. Ce projet souhaite notamment conserver la mémoire de la Lorraine et la mettre à disposition dans différentes institutions patrimoniales (musée, bibliothèques, écoles) ; participer à sa valorisation nationale, en faisant de ce corpus un objet d'études pour des historiens, littéraires, linguistes.

Projet AMPLorr autour de la prosographie et de la définition des espaces sociopolitiques lorrains au Moyen Âge. Ce projet, s'appuyant sur une coopération entre le CRULH et l'ATILF, mettra progressivement à disposition un large corpus constitué par les actes originaux des ducs lorrains depuis Ferry III (1251-1303) jusqu'à Charles II (1390-1431). Une première version de ces corpus est disponible à l'adresse : https://www.ortolang.fr/market/corpora/amplor. Ce corpus après prétraitement informatique, permet un enrichissement des bases de connaissances lexicales et textuelles et plus largement, pour les historiens, offre la possibilité d'analyses sérielles et thématiques (étude du gouvernement du prince dans la longue durée du (ou des) règne(s), établissement de ses itinéraires, de ses réseaux de pouvoir, études des transferts culturels, gestion de son domaine, etc.), mais aussi comparatives avec les régions et États environnants (en particulier la Grande Région).

Projet ALIENTO (Analyse Linguistique Interculturelle d'ÉNoncés sapientiels et de leur Transmission de l'Orient à l'occident et de l'occident à l'orient). Ce projet, centré sur une coopération entre le LIS et l'ATILF et mené en coopération avec l'INALCO structurée au sein d'un projet ANR, a pour objectif la construction d'une base de données sur les sources et l'étude de la transmission, de la circulation et de la postérité des énoncés sapientiels (proverbes) de la péninsule Ibérique (IXe-XVe siècle) entre les trois cultures espagnole, judéo-

espagnole et maghrébine contemporaines. Il vise donc à calculer les concordances partielles ou totales des textes, leurs connexions proches et éloignées afin de réévaluer les relations intertextuelles, en confrontant une grande quantité d'unités et en croisant des textes écrits dans des langues différentes. Un projet ERC PIECES est en cours de soumission auprès de l'ERC.

Projet HP-Papers (Henri Poincaré Papers), porté par le LPHS, a pour objet est de poursuivre et d'analyser les travaux de H. Poincaré, scientifique et philosophe, au travers de la publication et de l'édition critique de ses manuscrits et de ses travaux non encore publiés. Il répond au souhait des membres des communautés mathématiciennes et physiciennes de mettre en ligne une édition électronique des Œuvres de Poincaré, qui reste une ressource essentielle pour la recherche dans ces domaines. Nous possédons un outil performant pour réaliser une telle édition : LaTeXML, qui intègre la possibilité de marquage sémantique, et permet la recherche des formules mathématiques. Cette nouvelle édition des Œuvres de Poincaré sera annotée et indexée à l'édition en ligne de la correspondance et des cahiers de recherche de Poincaré, en cours depuis 1994 et contribuera à la valorisation d'un patrimoine intellectuel et matériel hors pair.

Projet TCOF (https://www.ortolang.fr/market/corpora/tcof), porté par l'ATILF, d'interface d'interrogation de données orales alignées texte-son et annotées automatiquement par des analyses grammaticales (collaboration avec le projet ANR ORFEO). L'objectif est ici de faire de la Lorraine une des régions incontournables dans le domaine de l'exploitation du français parlé. Le développement et l'enrichissement de ressources linguistiques orales sont actuellement des secteurs stratégiques pour mieux comprendre les mécanismes de fonctionnement et d'acquisition du langage et pour développer des applications liées au TAL.

Projet Re-Typographe autour de la représentation de typographies anciennes dont l'objectif est d'étudier l'apport que peuvent fournir les techniques d'analyse d'images et d'approximation de formes graphiques à la reconstruction des polices et des informations typographiques issues de la Renaissance. Ce projet vise à combiner l'expertise en analyse typographique de l'ANRT (Atelier National de Recherche Typographique) et celle en analyse d'images numériques du LORIA et consistera, à étudier et développer des algorithmes d'analyse graphique les plus pertinents pour permettre les différentes mesures nécessaires à la rétroconversion et la reconstruction de modèles typographiques anciennes.

Outil Collaboratif d'Analyse des processus d'Intégration, proposé par le 2L2S qui vise à interroger des bases de données prosopographiques des individus permettant d'une part les reconstitutions de carrière et d'autre part d'analyser la diversité et la mixité sociale. Les sources sont multiples en termes de forme et de localisation spatiale et l'outil à élaborer devra en permettre un partage et une complétude. L'outil sera mis à disposition des chercheurs en SHS, mais également utilisables par les entreprises et institutions désireuses de reconstituer les carrières éclatées de leurs employés. Une des applications finalisées en partenariat avec Arcelor Mittal sera de faciliter la reconstitution du parcours professionnel d'un individu au moment où il fait valoir ses droits à la retraite.

Projet d'Observatoire des usages numériques s'inscrivant dans la continuité de plusieurs opérations de l'axe 3 de la MSHL (Obsweb, TecMeus, SumTec) et du programme ANR Info-RSN (2014-2016) sur la circulation et le partage des informations sur les réseaux socio-numériques (fruit d'un partenariat entre le LCOMS et le CREM). Cette action vise à mettre en place un Observatoire des Usages numériques, grâce à des coopérations renforcées entre le CREM et des laboratoires d'informatique de l'université de Lorraine. La logique de l'observatoire consiste à considérer que des pratiques sociales et des usages sont en cours d'invention et que ces médiations sociotechniques méritent d'être étudiées au fil du temps pour en comprendre les logiques, et ce grâce à une approche d'enquête pluriméthodologique (entretiens, observation, questionnaires) et pluridisciplinaire (big data, modèles décisionnels avancés, interaction humain-machine). Cet observatoire ciblera son étude principalement sur des terrains pouvant toucher aux TICE, au web-journalisme, aux jeux numériques et aux publications scientifiques en ligne. Plusieurs dimensions seront prises en compte dans ces analyses : i) les conditions de production et de diffusion des informations et des connaissances, ii) les attitudes et comportements des publics d'usagers, iii) les mécanismes d'intercompréhension ou de blocages communicationnels, iv) le poids des facteurs technologiques dans les médiations sociotechniques.

Suite à l'appel à propositions de soutien d'actions de recherche lancé au cours de cette année, deux projets relevant de cet axe ont été retenus pour des financements spécifiques (cf. bilan de ces projets à la suite du bilan général des quatre axes structurants du projet LCHN) :

- le projet HP Papers « Web sémantique et outils de recherches »

- et le projet Site e-éducation paléographie médiévale et documents lorrains (PALEOLOR), commun avec l'axe E-éducation

E-Éducation

Animateurs : Samuel Nowakowski (LORIA), Sébastien Genvo (CREM)
Laboratoires impliqués : LORIA, CREM, ATILF, LCOMS, PErSEUs, LISEC

Le précédent CPER a permis l'émergence d'un thème nouveau, l'e-Éducation sur lequel se sont déjà positionnés plusieurs laboratoires issus de perspectives disciplinaires variées (informatique, sciences de l'éducation, sciences du langage, sciences de l'information et de la communication, psychologie, etc.). Ce thème se situe à l'interface des sciences du numérique et de l'ingénierie des connaissances, qui sont tous deux considérés comme des enjeux majeurs pour le territoire lorrain, que ce soit à travers le contrat de site ou les programmes opérationnels de la Région. Les laboratoires lorrains partenaires du projet ont acquis une visibilité nationale et internationale dans trois secteurs permettant d'investir ce domaine de façon innovante : l'apprentissage des langues (français et langues étrangères), les jeux vidéo (jeux expressifs, jeux sérieux, phénomènes de ludicisation), la modélisation numérique et l'analyse des identités et des usages (étudiants, enseignants, grand public, etc.) dans les environnements numériques d'apprentissage. Les enjeux de ces différents axes sont multiples face aux défis que l'université doit relever en termes d'accueil des publics, plus particulièrement au sujet du renouvellement de la pédagogie liée à l'évolution du profil des étudiants (massification, prise en charge des publics fragiles, etc.). Nous poursuivons les efforts initiés au précédent CPER en nous focalisant sur une réflexion méthodologique et technique autour des impacts du numérique dans l'éducation. L'accent sera mis sur les trois aspects suivants, qui seront interconnectés à de nombreux égards tant sur le plan du contenu qu'à travers une gouvernance commune :

L'apprentissage des langues, avec pour acteurs principaux le LORIA et l'ATILF, afin de développer des environnements favorisant les apprentissages, en s'appuyant sur les savoir-faire en matière de traitements du texte (génération automatique d'exercices et système de dialogue pour les apprenants), de la parole et de production de plateformes et de ressources pédagogiques numériques adaptées. L'action proposée entre dans le cadre des aides à l'apprentissage de la grammaire et de l'oral d'une langue. Ses principaux objectifs sont l'élaboration d'une formation à la pratique de la langue (système de dialogue pour les apprenants), de la grammaire (manuel électronique) ainsi qu'à l'analyse et à la visualisation de la parole, principalement destinée aux enseignants de langue, mais également à tout professionnel dont le domaine couvre les sons de la parole, ainsi que l'évaluation de l'impact de certaines modifications de la parole (comme le ralentissement sélectif) qui seront testées en situation réelle d'apprentissage dans un laboratoire de langue. Ce projet rassemblera les équipes Synalp et Parole (Multispeech) du LORIA, et les enseignants de langues de l'école des Mines de Nancy. Des collaborations sont envisagées également avec le CHU de Nancy (CLAP). Sur le plan international, ce projet s'inscrit dans la ligne des projets sur l'apprentissage des langues de l'équipe Parole (Multispeech) comme l'ANR franco-allemande IFCASL dirigée par Jurgen Trouvain (Université de la Sarre, Sarrebruck), le projet INTERREG IVA Allegro (Metz, Nancy, Saarbrücken). Des négociations sont en cours pour une collaboration avec ITEC – Interactive Technologies, KY Leuven (Belgique).

Les jeux vidéo afin d'interroger les facteurs favorisant l'engagement et la compréhension de profils variés d'apprenants dans un dispositif numérique notamment, à travers la narration et les mécanismes ludiques. En d'autres termes, il s'agit de mettre à profit les technologies vidéoludiques pour améliorer la qualité des interactions homme-machine, en vue d'impliquer au mieux l'apprenant et/ou d'adapter le dispositif selon son profil. À partir d'analyse de contenus et d'usages, l'objectif est de comprendre les spécificités de ce médium en tant que forme d'expression et de représentation de sorte à rendre ces systèmes adaptatifs et facilitateurs d'acquisition et d'apprentissage. Les études menées permettront de questionner les types d'apprentissages (formels ou informels) et de connaissances véhiculés par les jeux vidéo - qu'ils soient ou non destinés à l'éducation. Un corpus diversifié de différents genres vidéoludiques, issus de différentes périodes, sera à ce titre analysé En fonction de ces résultats, un prototypage de logiciels expérimentaux sera mené, permettant de tester par le développement les hypothèses de recherche (à travers des jeux vidéo expressifs et des « *newsgames* » notamment). Les compétences mobilisées dans cet axe mettront à profit les travaux menés au CREM sur les spécificités ludiques et narratives des jeux vidéo, les travaux développés dans l'équipe MAIA (LORIA) en termes de planification, ceux de l'équipe KIWI (LORIA) concernant la modélisation utilisateur et des interactions (compréhension et intégration des facteurs humains dans les modèles d'apprentissage automatique) et ceux de l'équipe SYNALP (intégration de moteurs de dialogue et de générateur d'exercices). Ces réflexions d'ensemble sur le médium seront enrichies par l'étude d'un terrain plus spécifique, qui s'articule avec l'axe précédent, celui des jeux sérieux orientés vers l'apprentissage des langues. En effet, au sein des

projets européens Allegro et EmoSpeech, l'équipe Synalp (LORIA) a acquis une expertise forte dans l'interfaçage entre jeux sérieux, traitement de la langue naturelle et enseignement des langues assistés par ordinateur. L'objectif est d'étendre la couverture pédagogique des maquettes déjà développées, d'utiliser les traces web des apprenants pour développer des systèmes qui s'adaptent automatiquement au niveau de l'utilisateur, et de développer des moteurs de dialogue facilitant l'apprentissage.

La modélisation numérique et l'analyse des identités et des pratiques dans les environnements numériques d'apprentissage (étudiants, enseignants, grand public, etc.) : s'appuyant sur l'analyse des pratiques des enseignants, des étudiants et de tout utilisateur en situation d'apprentissage (notamment les publics vulnérables ou fragiles), cet axe vise à proposer des modèles numériques pour mieux comprendre l'expression de compétences dans les environnements numériques, mais également à analyser en quoi ces environnements transforment ou font évoluer les usages, les pratiques et les identités des enseignants et des apprenants. La notion d'identité numérique dans nos travaux est entendue comme un ensemble de « graines » semées çà et là au gré des usages numériques. L'observation de ces « graines », traces, données, nous amène à revoir les principes de l'« être ensemble » autour des nouvelles logiques d'échange, de partage et d'exposition de soi dans les environnements numériques d'apprentissage. Elle nous conduit également à reconsidérer la question de la polymorphie des identités consciemment construites et affichées par l'individu, selon ses besoins et ses communautés d'intérêts et les identités inconscientes reconstituées par des logiciels et non maîtrisées. Un des versants du projet conduira à analyser comment ces traces peuvent être ré-exploitées pour être mises au service de pratiques réflexives et comment elles s'inscrivent dans des logiques sociales, professionnelles et/ou institutionnelles. Observer, collecter et analyser cette identité numérique impliquera de mener des investigations à partir des méthodes et outils développés dans le cadre de la plateforme d'expérimentation en humanités numériques. Un des terrains qui sera particulièrement investi concerne la constitution d'un modèle de référence de compétences harmonisé appris automatiquement à partir de plusieurs sources de données, d'un volume très conséquent : des données d'expérience-terrain (cv, lettres de motivation, offres d'emploi, etc.) et des référentiels de compétences d'entreprises de plusieurs domaines, et des référentiels standards (RNCP par exemple). Les défis scientifiques seront liés à la définition automatique d'un espace de représentation des compétences qui permettra d'alimenter des travaux réflexifs personnels et au sein de collectifs à travers la constitution d'outils numériques (recommandation, profilage, etc.) qui accompagneront la montée en compétences. Les acteurs impliqués ici sont le LORIA (équipe KIWI et SYNALP), le CREM, le LISEC, le LCOMS, PErSEUs et nombre d'actions menées se font également au sein de la MSH Lorraine (en particulier dans l'axe 3 avec les projets ADN et SUMTEC).

Suite à l'appel à propositions de soutien d'actions de recherche lancé au cours de cette année, deux projets relevant de cet axe ont été retenus pour des financements spécifiques (cf. bilan de ces projets à la suite du bilan général des quatre axes structurants du projet LCHN) :

- le projet Observatoire des usages numériques (OUN)
- et le projet de site d'e-éducation en paléographie médiévale et documents lorrains (PALEOLOR), commun avec l'axe humanités Numériques

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

Par ailleurs nous avons lancé fin 2015 un appel 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. Après double évaluation par le comité de pilotage du projet 8 actions ont été retenues pour financement sur les crédits 2015 :

DEMONETTE (ATILF)

L'objectif du projet est d'enrichir la couverture de la base Démonette, base lexicale morphologique du français organisée en réseau dérivationnel (https://www.ortolang.fr/#/market/lexicons/demonette), 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

de manière à offrir, via la plateforme qui accueillera les résultats du programme du CPER LCHN, et plus particulièrement ceux de l'action Ingénierie des langues, une ressource lexicale de grande taille, annotée au moyen d'informations fiables pour le français et dont chaque entrée code une relation morphologique entre noms, adjectifs et verbes du français.

L'évolution de la base prévoit l'intégration et l'adaptation du contenu de trois ressources : le lexique Lexeur (Hathout & Fabre, 2002), la base MORDAN, cf. https://apps.atilf.fr/mordan/ (Koehl, 2012) et l'ensemble des relations dérivationnelles extraites du lexique GLAWI (Nabil Hathout, Franck Sajous and Basilio Calderone, 2014).

Lors de chaque intégration, nous veillons à produire une représentation unifiée (qui caractérise actuellement la version 1.2) des ressources lexicales incorporées, pourtant différentes tant en contenu qu'en objectifs, et à compléter la couverture de ces ressources par le calcul des valeurs morphosémantiques contenues implicitement dans les données d'origine. De la sorte, ces ressources trouvent de nouveaux emplois, et servent à tisser, au sein de la même structure d'accueil qu'est Démonette, un réseau de plus en plus complexe.

Pour garantir une stabilité des résultats à long terme, la conception des programmes de migration doit anticiper deux types de problèmes : l'extension régulière de l'architecture de la base sans en compromettre la structure préexistante (avec une partie obligatoire, une partie « importante » (codage des types morphosémantiques et des définitions) et une partie optionnelle (codage de la représentation phonologique), et l'ajustement aux besoins descriptifs de l'ensemble des valeurs qui peuvent être affectées à un trait donné, dont le domaine de définition doit être prévu d'avance.

Actions menées depuis janvier 2016, et prévisions d'ici la fin 2016

- La mise au format Démonette de la base Lexeur est réalisée (enrichissement par les formes du féminin, identification des opérations morphologiques, génération des types et des gloses sémantiques). Ce travail de migration a été présenté lors des conférences IMM16 ("Computational methods for descriptive and theoretical morphology") et LREC 2016 (Hathout & Namer 2016).
- Suite à l'incorporation de Lexeur, prévue pour la fin 2016, Démonette, dans sa version 1.3 décrira 35.764 lexèmes, comportera 171.750 entrées dont 108.888 couples (M1,M2) différents (il faut garder en tête qu'un même couple peut avoir été importé à partir de différentes ressources. Cette redondance n'est pas gênante, dans la mesure où l'origine de chaque information est annotée dans la base).
- Le manuel d'utilisation en français de la base a été mis à jour, et rend compte des nouvelles caractéristiques de celle-ci. Sa traduction en anglais sera effectuée au moment de la diffusion de la version 1.3 de Démonette.
- Une vacataire vient d'être recrutée pour valider et si besoin corriger certains contenus de Démonette.

Les tâches qu'elle a à accomplir sont les suivantes :

1) À partir d'exemples attestés, vérifier la pertinence et la validité des étiquettes sémantiques des noms de procès (@ACT et @RES, cf Figure 1) de Démonette
2) Il s'agira par exemple de valider le fait que *lavage* peut désigner une activité @ACT ("Pendant la durée du lavage, nous recommandons de ne pas ouvrir le hublot de la machine à laver") plus difficilement un résultat @RES ("?En voilà de beaux lavages"); qu'à l'inverse, *admiration* n'est jamais interprétable comme une activité.
3) Vérifier dans quels cas les noms d'agent masculin (@AGM, cf. Figure 1) et féminins (@AGF) servent aussi (voire exclusivement) à dénoter des instruments : c'est le cas d'*interrupteur*, pas de *danseur*
4) Regarder si la définition attribuée à Mot1, dans chaque entrée (cf. Figure 1), correspond à un sens attesté.

Les résultats de ce travail sont attendus pour fin août, et serviront à modifier, si besoin est, les règles de prédiction des propriétés sémantiques des mots construits. Nous nous attendons en effet à modifier le programme d'intégration pour tenir compte des corrections apportées aux résultats bruts donnés à valider.

HP-PAPERS (LHSP)

Dans la volonté de produire une édition critique et numérique des manuscrits, de l'œuvre éditée et de la correspondance d'Henri Poincaré, HP–Papers a pour ambition de regroupersur sa plateforme numérique l'ensemble de la correspondance de ce scientifique majeur de son époque et de 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.

Les travaux portent sur une annotation sémantique de la correspondance de Poincaré, le déploiement d'un outil de recherches exactes et la création d'un outil de recherches approchées.

Forte de l'expérience de SemanticHPST, l'équipe constituée envisage de mettre en place une plateforme, un triple store et des outils de recherches. Pour ce faire, nous nous appuierons sur les ontologies construites par SemanticHPST et par SyMoGIH (Système Modulaire deGestion de l'Information Historique) développé par le pôle numérique du laboratoire d'histoire de Lyon, le LARHRA.

Même si l'année 2016 est loin d'être achevée, il est possible de pointer quelques résultats et avancées dans le cadre d'HP-Papers. Signalons tout d'abord que l'équipe s'est renforcée du côté informatique. Dorénavant, Emmanuel Nauer (du Loria) et Nicolas Lasolle (stagiaire Télécom Nancy) participent à ce projet.

Initialement, le calendrier était le suivant :

- 2016 : Mise en place de l'ontologie associée à la correspondance de Henri Poincaré, déploiement d'un triple store pour l'accès aux triplets liés à cette correspondance ;
- 2017 : Déploiement d'un système de recherche exacte et développement d'un outil de recherche approchée s'appuyant sur des règles de transformation de requêtes SPARQL.

Pour des raisons pratiques, il a été décidé d'inverser ce calendrier et donc dans un premier temps de se concentrer sur les recherches exacte et approchée. Le stage en cours de Nicolas Lasolle (étudiant en 2e année de l'école d'ingénieur Telecom Nancy) porte sur cette problématique. Plus précisément, les tâches sur lesquelles Nicolas Lasolle

travaille durant son stage sont les suivantes :

- Mise en place du moteur de recherche exacte de Corese / KGRAM;
- Définition d'une classe Java pour les règles de transformation de requêtes SPARQL : applicabilité, application, sérialisation XML;
- Outil de recherche approchée par parcours de l'espace des requêtes structuré par un ensemble de telles règles ;
- Test sur un ensemble d'une dizaine de règles déjà spécifié.

Par ailleurs, il est nécessaire afin de contrôler l'efficacité de ces règles de les tester sur une ontologie construite autour de la correspondance d'Henri Poincaré. Cette ontologie n'est que temporaire et largement lacunaire. L'objectif est de travailler sur la base d'une ontologie de plus haut niveau. Nous optons pour celle de SyMoGIH, quitte à la compléter pour être plus en adéquation avec le corpus particulier qu'est une correspondance largement scientifique et à faire évoluer SyMoGIH. Une collaboration avec les chercheurs à l'origine de SyMoGIH est en cours.

Au cours de l'année 2016 cette action a conduit aux actions de valorisation suivante :

Date	description	Lieu
2 et 3 juin 2016	Participation à la journée d'étude "SyMoGIH" (OB, JL et EN)	LARHA, Université de Lyon
30 mai 2016	Comité scientifique (OB) du 2e workshop 'Semantic Web for Scientific Heritage (SW4SH)'	Anissaras, Crète
26-28 Août 2016	Participation au colloque de Digital|HPS	Norman, Oklahoma University

ITL-DI-Œil (ATILF, LORIA, LHSP)

L'opération "interrelation troubles du langage, discours et processus oculomoteurs" s'appuie sur un programme de recherche visant l'élaboration d'une ressource de corpora (enregistrements audio et visio-moteurs d'interactions verbales en face à face) destinée à la description des troubles du langage et du discours, dont des personnes souffrant de divers syndromes psychiatriques, en particulier, sont susceptibles. Ces troubles peuvent affecter par exemple la prosodie, le lexique, la syntaxe, les relations inter-actes au niveau intra-intervention, au niveau de l'échange ou encore au niveau de la transaction dialogique. Il s'agit aussi de décrire les vraisemblables relations entre troubles du langage et comportement oculomoteur (saccades et stratégies de fixation) tel qu'elles s'expriment dans un contexte « naturel » d'interaction verbale. Nous envisageons de créer des outils informatisés d'appréhension de ces artefacts langagiers, visio-moteurs et discontinuités discursives. L'objectif scientifique de cette opération est double : d'une part modéliser la relation entre compétence langagière et comportement oculomoteur à l'épreuve des modèles et théories du discours, d'autre part expliciter la relation entre langage et processus cognitifs de type 'inférence' ou 'raisonnement' ainsi que certains déterminants de la relation entre langage et pensée (modélisation formelle de type sémantique). Au plan clinique, cette opération est susceptible d'apporter des outils de confirmation diagnostique précis dont manquent les professionnels de la

psychopathologie, voire de créer des outils d'anticipation diagnostique pour les populations à risque ou encore de prise en charge psychothérapeutique.
Le recueil de données a lieu dans un hôpital psychiatrique à Aix-en-Provence.
Nous avons mis au point un double-système d'enregistrement simultané du comportement visuo-moteur et du comportement verbal en combinant deux eye-trackers. Les interactions cliniques enregistrées associent une personne porteuse d'un syndrome psychiatrique (schizophrénie, paranoïa) à un psychologue. Des groupes contrôles seront constitués en remplaçant les patients par des sujets typiques appariés. Chaque sujet expérimental subira la passation d'une batterie de tests neuropsychologiques.
L'analyse des résultats permettra d'expliciter à court terme l'hypothèse de type méthodologique selon laquelle les mouvements involontaires de contrôle de l'attention (saccades), en tant que marques d'étonnement ou d'hésitation par exemple, renseignent quant à la manière dont les éléments de langage, les composants de la Forme Logique et de la Force Illocutoire des propositions et actes de langage contribuent à la cohésion et à la cohérence du discours étant donné l'accomplissement des stratégies d'adaptation que les interlocuteurs exercent l'un à l'encontre de l'autre au cours de la progression de l'interaction.
Enfin les marqueurs de la variation du comportement attentionnel des interlocuteurs que sont les saccades, pourront être bientôt combinés à l'analyse du déplacement du regard au cours de l'interaction (fixations) notamment dans le cadre de l'accomplissement de l'intercompréhension, puis ramenés à l'analyse des discontinuités discursives et autres spécificités langagières des interlocuteurs.

Cette année de lancement de l'opération a été consacrée à la préparation et à la mise en place du programme de recherche, avec notre partenaire hospitalier et clinique d'une part, en l'installation du matériel et au commencement du recueil de données d'autre part.
Nos collaborateurs hospitaliers et cliniques sont Philippe Combes et Monique D'Armor (Médecin-Psychiatres CHS Montperrin / Aix-en-Provence), Sophie Bathélémy (Psychologue clinicienne, CHS Montperrin /Aix-en-Provence) et Guy Gimenez (Professeur de psychopathologie clinique, Aix-Marseille Université).
Une fois, le programme de recherche établi, nous avons déposé les demandes d'autorisation préalables au lancement du recueil de données. Toutes sont satisfaites à ce jour, à savoir :

- Accord du Comité des Personnes Sud Mediterranée 1 (CPP), Numéro d'identification : 2015-A01457-42
- Accord de l'Agence Nationale de Sécurité du Médicament et des Produits de Santé (ANSM) pour une autorisation d'essai clinique ne portant pas sur un produit de santé
- Accord de la Commission National Informatique et Libertés (CNIL) : Déclaration de conformité à une méthodologie de référence (numéro de déclaration 1909238 v 0)

La méthodologie du recueil de données a été élaborée et engagée dès le 1er semestre 2016 ; chaque sujet appréhendé (témoin ou personne souffrant de troubles mentaux) est soumis :

- à une phase initiale d'approche de son profil neuropsychologique comportant des tests neuropsychologiques WAIS-IV, TMT et CVLT, assurant aussi l'évaluation des fonctions exécutives ;
- à un entretien semi-dirigé enregistré (audio + vidéo) ;
- à un entretien réalisé avec un système de double eye-tracking intégrant les parcours visuels (positionnement du regard et fréquence des saccades oculaires, durées de fixations, dilatation de la pupille.

Le groupe expérimental comprend quatre sous-groupes de 15 patients (n = 60), définis comme suit :

- un groupe de patients paranoïaques présentant un état délirant (DSM-5 : « Troubles délirants ») en Soins Libres (SL), âgés de 18 et plus, hospitalisés sur le C.H.Montperrin ou en ambulatoire.
- un groupe de patients paranoïaques présentant un état délirant (DSM-5 : « Troubles délirants ») en Soins Sans Consentement (S.S.C.), âgés de 18 et plus et suivis sur le C.H.Montperrin en Programme de Soins.
- un groupe de patients schizophrènes présentant un état délirant (DSM-5 : « Schizophrénie ») en SL, âgés de 18 et plus et hospitalisés sur le C.H.Montperrin ou en ambulatoire.
- un groupe de patients schizophrènes présentant un état délirant (DSM-5 : « Schizophrénie ») en SSC, âgés de 18 et plus et suivis sur le C.H.Montperrin en Programme de Soins.

Le groupe témoin composé de sujets typiques comprendra 20 sujets appariés. Il sera augmenté si nécessaire en fonction des caractéristiques spécifiques du groupe expérimental.
Réalisations : Les deux eye-trackers dont dispose le laboratoire Atilf ont été installés au CHS Montperrin ; Marie-Hélène Pierre (qui participe au recueil de données) a formé la psychologue locale à l'utilisation du matériel en avril 2016.
Recueil de protocoles : 6 protocoles complets auront été recueillis entre le 1er juin et la mi-juillet 2016.
-déplacement de Marie-Hélène Pierre du 10 au 20 juillet pour aide au recueil de données

-déplacement de Michel Musiol du 21 au 24 juillet pour animation d'un séminaire de recherche avec l'équipe (l'hébergement des Nancéiens est assuré par le l'hôpital).

MGBChallenge (LORIA)

L'équipe Multispeech du Loria travaille depuis de nombreuses années sur la transcription des émissions de radio ou de télévision. Des campagnes d'évaluation sont organisées pour mesurer les performances de systèmes complets de transcription et pour comparer les méthodes et les modèles utilisés. Ces campagnes confrontent la recherche à de nouveaux problèmes issus de données réelles et permettent d'explorer des idées novatrices.

Notre objectif est de participer à une campagne d'évaluation internationale : le challenge MGB-2 (*Multi-Genre Broadcast 2*) [4]. Ce challenge repose sur des enregistrements de télévision en anglais de la *British Broadcasting Corporation* (BBC).

La participation à ce challenge nous permettra d'obtenir des ressources extrêmement importantes :

- 1600 heures de fichiers audio et les sous-titres correspondants,
- un corpus de texte de 640 millions de mots de sous-titres d'émissions de télévision,
- un système de transcription de base (fourni par les organisateurs) au niveau de l'état de l'art.

Elle nous permettra aussi d'exploiter pleinement le travail de recherche que nous avons effectué sur la transcription depuis quelques années, de catalyser la mise en commun des expertises des différents membres de Multispeech, et de comparer nos résultats de recherche à ceux des systèmes concurrents d'autres institutions internationales. De plus, la participation à cette campagne va augmenter encore plus la visibilité internationale du laboratoire et de la région.

Ce projet nécessite un très grand effort de calcul et d'expérimentation, car il requiert la mise en œuvre de réseaux de neurones profonds et d'une très grande quantité de données d'apprentissage. L'utilisation de la plateforme de calcul *Ingénierie des langues,* et tout particulièrement les machines avec GPU, sera indispensable.

Afin de figurer en bonne position dans ce challenge, il est nécessaire d'avoir un système « état de l'art » basé sur les réseaux de neurones profonds (Deep Learning, *Deep Neural Networks* ou DNNs). Les DNNs peuvent être utilisés à différents niveaux : traitement du signal, modélisation acoustique et modélisation linguistique. Ils seront largement utilisés dans notre système destiné au challenge MGB. En particulier, le système de base fourni par les organisateurs exploite les DNNs pour la modélisation acoustique.

Ce challenge pose de nombreux défis scientifiques.

- Certaines émissions comportent beaucoup de parole spontanée (débats, séries télévisées…). La parole spontanée entraîne des modifications de la prononciation, par exemple des réductions (/il y a/=> /ja/), des hésitations (euh), des reprises et des répétitions. Ces particularités ne sont généralement pas retranscrites au niveau des sous-titres.
- Une autre difficulté provient du fond sonore **:** musique de fond, bruit de rue lors d'une émission en live, applaudissements, rires…
- Il y a beaucoup de locuteurs, avec des accents différents ou des locuteurs non natifs.
- La quantité de données d'apprentissage est très importante : 1600 heures soit de l'ordre de 580 millions de vecteurs de paramètres.
- Les sous-titres sont « approximatifs », c'est-à-dire que ce ne sont pas des transcriptions exactes de ce qui a été dit. En plus de l'absence des disfluences (parole spontanée), les sous-titres sont contraints par la vitesse de lecture des sous-titres à l'écran et donc sont souvent raccourcis.

OUN (CREM PERSEUS, CEREFIGE)

Ce projet porte plus particulièrement sur le domaine du sport électronique de sorte à questionner les types d'apprentissages (formels ou informels) et de connaissances véhiculés par les jeux vidéo, et les techniques de persuasion mises en œuvre afin de favoriser l'engagement de l'utilisateur. Il vise à investir un premier terrain d'observation en 2016 pour initier la création d'un observatoire des usages numériques, ceci grâce à des coopérations pluridisciplinaires renforcées entre trois laboratoires de l'université de Lorraine, à savoir le CREM (sciences de l'information et de la communication), PERSEUS (ergonomie et sciences cognitives) et le CEREFIGE (sciences de gestion). Dès à présent, les membres du projet ont fait une recension bibliographique liée à la narration au sein des jeux vidéo, le projet scientifique de l'axe se proposant d'interroger l'engagement de l'utilisateur dans un dispositif numérique (narration, mécanismes ludiques) et la compréhension du joueur des

connaissances véhiculées. Des analyses d'usages liées au jeu League of legends ont débuté, afin de mettre en évidence les facteurs favorisant l'engagement comportemental des joueurs au sein d'un jeu dont le modèle économique (Free to play) dépend directement de la participation des joueurs. Il s'agit notamment de mettre en avant la façon dont les connaissances transmises par le jeu incitent à l'acte d'achat (technologie persuasive). Ces travaux permettent également de répondre au projet de l'axe en Humanités numériques en ce qu'il s'agit de comprendre l'évolution des pratiques du numérique à partir d'une plateforme d'expérimentation dédiée, le matériel acquis dans ce cadre permettant d'initier la constitution de cette plateforme concernant l'équipement en réalité virtuelle. L'acquisition d'équipement spécifique aux pratiques vidéoludiques, notamment une station PC de jeu, permettra ultérieurement d'approfondir conjointement l'analyse des logiques immersives liées à l'exemplification technologique (particulièrement ce qui a trait au rôle du photoréalisme dans l'engagement utilisateur).

PALEOLOR (CRULH Atelier Diplomatique)

Ce projet permet de répondre à une autre partie de l'axe liée aux environnements favorisant les apprentissages, en s'appuyant sur les savoir-faire en matière d'informatisation de l'écrit et des images et de production de plateformes et de ressources pédagogiques numériques adaptées. Engagés dans des projets successifs (PRINCILOR, AMPLorr, puis TRANSSCRIPT) visant à mettre en ligne les actes médiévaux des ducs de Lorraine à l'usage de la communauté scientifique, les membres du projet PALEOLOR se proposent de faire contribuer les réalisations présentes et futures au contenu de l'enseignement supérieur, et plus largement au rayonnement d'une « science pour tous ».

Il s'agit concrètement de répondre à un double objectif :

1. Favoriser un apprentissage actif de la paléographie ;
2. Faire des documents (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 :

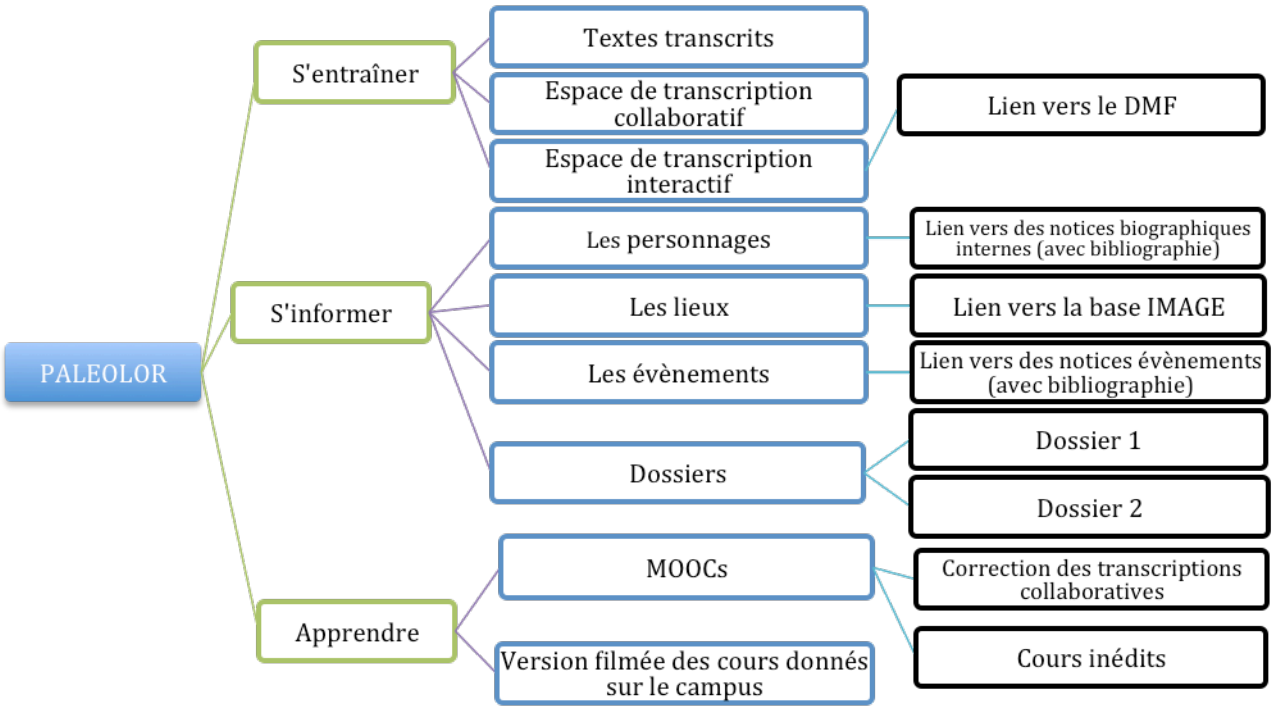
Apprendre la paléographie : exercices de paléographie interactifs (choix des exercices par niveau de difficulté, période) ;

Découvrir la Lorraine médiévale : les documents les plus riches du corpus TRANSSCRIPT, à partir desquels les étudiants pourront découvrir des lieux et personnages lorrains ; un espace qui s'enrichira progressivement des productions d'étudiants.

Conformément au calendrier prévisionnel, ces quelques mois ont été consacrés à une réflexion théorique sur les fonctionnalités et le plan du site web.

La première partie du site (« S'entraîner »), à destination des étudiants ou d'historiens amateurs, présentera des exercices de paléographie sous différentes formes :

- « classique », i.e. la photographie d'un acte et sa transcription ;
- « collaborative » : un espace type « pad » offre la possibilité à plusieurs utilisateurs (accès sur inscription) de transcrire conjointement un même texte. L'exercice aurait une durée limitée dans le temps, avec mise en ligne d'un nouveau document toutes les quinzaines par exemple ; le corrigé serait proposé lors d'un MOOC (cf ci-dessous), puis déposé dans la partie « textes transcrits » ;
- « interactive » : au survol de la photographie du texte, des bulles apparaissent pour proposer la transcription et des liens vers les notices biographiques, les évènements historiques, le DMF < http://www.atilf.fr/dmf/> ou la base IMAGE du CERPA < http://cerpa.univ-lorraine.fr >.



Une deuxième partie (« S'informer »), plus généraliste, présentera des données historiques sur la Lorraine sans avoir à passer par le texte et/ou sa transcription. Elle citera les documents présents sur le site avec des liens pouvant y renvoyer. Les dossiers seront des travaux d'étudiants présentant un acte, sa transcription et son analyse.

Enfin, le site abritera une partie « Apprendre ». L'objectif est de mettre en place une plateforme de cours en ligne accessible sur inscription. Certains cours pourront être en différé (filmés sur le campus puis diffusés en ligne), d'autres programmés à des dates et horaires précis (MOOCs). Certains MOOCs pourront être consacrés à la correction des transcriptions collaboratives, d'autres proposeront des cours inédits.

Prosodcorpus (LORIA, ATILF)

La prosodie porte des informations linguistiques et extralinguistiques essentielles pour, entre autres, structurer le message vocal, transmettre l'état émotionnel du locuteur, préciser une emphase, etc. Les paramètres prosodiques standards sont l'énergie et la durée ainsi que la fréquence fondamentale des sons. De nombreux algorithmes existent pour calculer la fréquence fondamentale du signal sonore. En ce qui concerne la durée des sons, elle s'obtient à partir de la segmentation du signal vocal en phonèmes, qui elle-même résulte généralement d'un alignement automatique de la parole sur le texte correspondant (la segmentation manuelle de gros corpus étant beaucoup trop coûteuse à réaliser). La segmentation de corpus de parole en unités linguistiques (mots, phonèmes, groupes prosodiques…) sert également pour la synthèse vocale, pour l'indexation de données vocales, pour la recherche d'extraits par un concordancier, pour des études linguistiques sur de gros corpus de parole (par exemple, étude des particules de discours, études statistiques sur les prononciations des mots…), sans oublier l'élaboration de diagnostics en apprentissage de langues.

La segmentation automatique du signal de parole en mots et en phonèmes ainsi que le calcul de la fréquence fondamentale fonctionne plutôt bien sur des signaux de parole de bonne qualité, et avec des transcriptions correctes (pour ce qui concerne la segmentation automatique). Cependant, les performances se dégradent lorsque les signaux de parole sont altérés (par exemple signal faible, mauvais rapport signal à bruit, paroles superposées…), et surtout il n'existe aucune mesure précisant la fiabilité des paramètres calculés (tant pour la segmentation que pour les estimations de la fréquence fondamentale).

L'objectif du projet porte donc sur l'amélioration de la précision et de la robustesse du calcul des paramètres prosodiques ; la détermination d'un indice de qualité (mesure de confiance) associé ; l'application des traitements améliorés sur plusieurs corpus de parole et l'exploitation des informations prosodiques résultantes et des mesures de confiance associées dans des études linguistiques. Cela conduit aux quatre tâches du projet qui se déroulent en parallèle :

- T1 – Amélioration du calcul des paramètres prosodiques
- T2 – Détermination d'un indice de qualité des paramètres prosodiques
- T3 – Application des traitements améliorés sur plusieurs corpus de parole

- T4 – Etudes linguistiques exploitant ces paramètres prosodiques et les indices de qualité associésBilan 2016 – à mi-année

T1 – Amélioration du calcul des paramètres prosodiques

Des travaux ont débuté cette année pour appliquer des approches à base de réseaux de neurones pour déterminer la frontière entre phonèmes. Une architecture à base de MLP (Multi Layer Perceptron) et une architecture à base de réseaux récurrents LSTM (Long Short Term Memory) ont été définies. Ces deux modèles ont été appris sur un corpus audio étiqueté manuellement (KIEL). Nous sommes en train de les évaluer et de déterminer la meilleure paramétrisation acoustique.

T2 – Détermination d'un indice de qualité des paramètres prosodiques

Des approches à base de réseaux de neurones ont également été abordées pour la mesure de fiabilité des segmentations phonétiques et des résultats de calcul de la fréquence fondamentale.

Un stage de master a permis d'étudier plus précisément la dégradation des performances de plusieurs algorithmes de calcul de la fréquence fondamentale sur des données de parole bruitées. La deuxième partie du stage a porté sur l'élaboration d'un classifieur à base de réseaux de neurones (MLP et LSTM) pour estimer la probabilité que la fréquence fondamentale soit correcte. Dans la version courante, le classifieur exploite des coefficients acoustiques provenant du signal (ex. coefficients cepstraux) ainsi que des résultats intermédiaires des algorithmes de calcul de la fréquence fondamentale.

T3 – Application des traitements améliorés sur plusieurs corpus de parole

Cette tâche n'est pas applicable pour l'instant. Elle ne pourra être mise en œuvre qu'après validation des approches sur les mesures de confiance, ou bien lorsque des techniques améliorées de calcul des paramètres prosodiques seront disponibles.

T4 – Etudes linguistiques exploitant ces paramètres prosodiques et les indices de qualité associées

Concernant les études linguistiques, les premiers travaux mis en œuvre exploitent les segmentations en mots et en phonèmes élaborées dans le cadre du projet ORFEO. A partir de ces données, les paramètres prosodiques ont été calculés sur plusieurs milliers de fichiers, correspondant à près de quatre millions de mots. Les occurrences de mots et d'expressions pouvant être utilisées en tant que particules de discours ont été détectées. Une petite partie de ces occurrences est en cours d'annotation (fonction particule de discours ou pas…) grâce au soutien du CPER sur cette opération. Les fréquences d'occurrences de ces mots et expressions varient notablement d'un corpus à autre, ce qui méritera une analyse plus fine en fonction du type de parole des corpus.

UniMETA (LORIA, INIST)

L'objectif du projet de recherche UniMETA est 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.

L'avancement du projet est conforme aux objectifs décrits dans le plan de travail. Les 5 premiers mois du projet ont été consacrés à la formalisation de la démarche d'indexation automatique basée sur la maximisation d'étiquetage, en partant de différentes méthodes d'indexation de surface basées, qui sur des extracteurs terminologiques, qui sur des analyseurs syntaxiques ad hoc. Les méthodes basées sur des extracteurs sélectionnés sont issues d'une analyse approfondie de l'état de l'art menée dans cette première phase de projet.

Les résultats produits ont été ensuite exploités de manière coordonnée avec la structure des textes en se basant sur les blocs logiques de premier niveau, de manière à extraire les traits saillants dans chaque bloc en s'appuyant sur le principe de la sélection de variables. La méthode de sélection de variables utilisée est une méthode originale basée sur la métrique de maximisation des traits, que nous développée dans l'équipe Synalp. Cette méthode a déjà fait ses preuves pour la classification des données textuelles. Elle permet également de caractériser les traits extraits par des valeurs de contrastes dépendant de leur contexte de manifestation.

L'analyse parallèle des performances d'indexation du couplage de la méthode proposée avec les extracteurs terminologiques et du fonctionnement de ces extracteurs ou de ces analyseurs seuls a été menée dans deux contextes différents, qui celui du clustering diachronique, qui celui du résumé automatique. L'étude liée au résumé automatique a fait l'objet d'un stage de fin d'études de Master de 5 mois d'Hazem Al Sied, financée avec les fonds attribués au projet pour l'année 2016.

L'étude des performances de la méthode dans le cadre du clustering a été conduite de manière coordonnée avec nos propres travaux de recherche et avec ceux menés dans le cadre du projet ISTEX-R. Elle a consisté à mesurer la capacité des différentes méthodes d'indexation susmentionnées à fournir un étiquetage de résultats de clustering

permettant de maximiser le nombre de correspondances temporelles entre des clusters relatifs à différentes périodes de temps. Cette première analyse s'est appuyée sur les données de référence du projet ISTEX-R, à savoir un ensemble d'environ 10000 publications scientifiques issues de collections hétérogènes et se rapportant au thème du vieillissement. Elle a permis de mettre en évidence les performances supérieures obtenues par le couplage des analyseurs ou des extracteurs avec la technique de sélection et de contraste de traits que nous avons proposée. Cette première étude a fait l'objet et d'un exposé invité dans une conférence internationale [2], ainsi que d'une publication dans une conférence internationale de rang A [3].

L'étude menée par Hazem Al Sied se rapportant au résumé automatique exploite les mêmes techniques d'indexation. Elle est encore actuellement en cours

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

Par ailleurs nous avons défini une première étape de mise en place des plateformes « Ingénierie des langues » et « Humanités numérique » (cf. ci-dessous point sur l'acquisition et la mise en œuvre des équipements).

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.

Au terme de la première année de ce projet, nous avons choisi de n'indiquer 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

Publications

Boulton, Alex., Integrating corpus tools and techniques in ESP courses
ASp, La revue du GERAS, revue.org (en ligne) / Bordeaux : GERAS (imprimé), 2016, 69, pp.111-135

Boulton, Alex Data driven learning in digital contexts
The Encyclopedia of Language and Education (3rd ed.), vol. 9: Language, Education and Technology, 2016

Bruneau, Olivier, Serge Garlatti, Muriel Guedj, Sylvain Laubé, et Jean Lieber (2015). 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

Hathout, Nabil and Namer, Fiammetta (2016). 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) ». Calzolari , N., Choukri , K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard , B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis , S. Portorož, Slovenia, European Language Resources Association (ELRA) : 1084-1091.

Marcon, Mario. Annotation morphosyntaxique et lemmatisation automatiques des parémies: défis et enjeux
Editions Universitaires de Lorraine - Presses Universitaires de Nancy. ALIENTO, 7, 2016

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

Communications à des conférences

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

Blanchard, Jérôme, Pestel, Cyril, Petitjean, Etioenne et Pierrel, Jean-Marieb L'Equipex ORTOLANG de mutualisation de ressources linguistiques écrites, orales et multimodales.
CMLF , 2016, Tours, France. Congrès Mondial de Linguistique appliquée

Hathout, Nabil and Namer, Fiammetta (2016). Modeling Meaning-Form Discrepancy in Word Formation within ParaDis, a Four Level Paradigm-based Modular Framework. « AnaMorphoSys ». Stump, G. and Walther, G. Lyon, ASLAN, CNRS.

Hathout, Nabil and Namer, Fiammetta (2016). Paradigms in word-formation : new perspectives on data description and modeling. « Workshop : SLE 2016 ». Naples.

Hathout, Nabil and Namer, Fiammetta (2016). A Multi-level Paradigm-based Model of Competition in Word Formation. « 17th International Morphology Meeting ». Rainer, F., Gardani, F. and Peters, E. Vienna.

Hathout, Nabil and Namer, Fiammetta (2016). Enriching the Démonette morpho-semantic network : computational and linguistic issues. « International Morphological Meeting 17 : Workshop on Computational methods for descriptive and theoretical morphology ».

Kister, Laurence , Marcon, Mario, Jacquey, Evelyne et Barreaux, Sabine Indices lexico-syntaxiques pour la reconnaissance des termes de la forme N_Adj dans les textes intégraux, Toth15, Jun 2015, Chambéry, France

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.

Lamirel, J.-C., Dugué N., Cucac P., New efficient clustering quality indexes, Proceedings of IJCNN 2016, Vancouver, BC, Canada, July 2016.

Mise à disposition de ressources

Démonette, version 2 : https ://www.ortolang.fr/market/lexicons/demonette