Vague D
Campagne d'évaluation 2017 – 2018

Unité de recherche

Dossier d'autoévaluation

Informations générales

Nom de l'unité : Equipe de Recherche Textes, Informatique, Multilinguisme
Acronyme : ERTIM

Champ de recherche de rattachement : Sciences du langage

Nom du directeur pour le contrat en cours : Mathieu Valette
Nom du directeur pour le contrat à venir : Mathieu Valette

Type de demande :

Renouvellement à l'identique

Établissements et organismes de rattachement :

Liste des établissements et organismes tutelles de l'unité de recherche **pour le contrat en cours** et **pour le prochain contrat** (tutelles).

Contrat en cours :
- Inalco

Prochain contrat :
- Inalco

Choix de l'évaluation interdisciplinaire de l'unité de recherche ou de l'équipe interne :

Oui □ Non x

DOSSIER D'AUTOÉVALUATION

1. Présentation de l'unité

Introduction

L'ERTIM est une Equipe d'Accueil créée en 2005 par la fusion de deux centres de recherches de l'Inalco : le Centre de Recherche en Ingénierie Multilingue (CRIM) et le Centre d'Etudes et de Recherche en Traitement Automatique des Langues (CERTAL). Les champs disciplinaires de l'ERTIM sont le Traitement Automatique des Langues et l'ingénierie multilingue. L'ERTIM est sise à la Maison de la Recherche de l'Inalco, 2 rue de Lille, Paris 7e.

L'ERTIM est une petite équipe composée de 6 permanents mais un nombre important d'impermanents (doctorants, contractuels, stagiaires etc.) participe à son dynamisme. Sur la période 2012-2017, 48 personnes ont été membres de l'équipe.

Le conseil de l'équipe est composé de 4 permanents (1 IGR, 1 PRAG, 1 MCF, 1 PR) et de 2 représentants des doctorants, auxquels s'ajoutent le directeur de l'équipe et le directeur adjoint.

Tableau des effectifs et moyens de l'unité

L'équipe comporte peu de permanents (1 PU, 2 MCF, 1 PRAG, ½ PAST, 1 IR, ½ assistante) mais les contractuels constituent entre 2/3 et 4/5 des effectifs, suivant les périodes (doctorants, post-docs, recrutements sur projet, stagiaires, etc.)

L'équipe fonctionne dans une large mesure sur le modèle des laboratoires de sciences, c'est-à-dire un travail collaboratif avec une présence permanente de beaucoup de ses membres. Les doctorants qui ne sont pas salariés par ailleurs sont présents presque quotidiennement, les doctorants sous Conventions Industrielles de Formation par la Recherche (contrats CIFRE) viennent ½ journée à 2 jours par semaine suivant la convention établie et les besoins qu'ils expriment.

Le budget de l'équipe, qui comprend la dotation du Conseil Scientifique, les contrats CIFRE et divers contrats de recherche, suffit à couvrir l'ensemble des frais de mission et d'équipement des personnels, doctorants, contractuels ou permanents. Dans les faits, il n'est pratiquement jamais nécessaire d'arbitrer les dépenses.

Politique scientifique

Le projet scientifique s'articule autour des thèmes suivants :

- la recherche en linguistique et sémantique des textes pour les applications en Traitement Automatique des Langues (TAL) ;
- le développement de méthodologie pour l'ingénierie des textes et des documents numériques multilingues ;
- la production de ressources multilingues (documentaires, lexicales, terminologiques, textuelles, didactiques).

Les travaux de l'ERTIM se caractérisent par :

- Une recherche axée sur **la résolution de problèmes scientifiques posés par les applications**. Les recherches de l'équipe ont pour objectif de répondre à des problèmes applicatifs concrets (fouille de textes, extraction d'information, apprentissage des langues) L'ERTIM s'efforce d'adopter une position critique par rapport aux modèles théoriques dominants et propose des alternatives fondées sur le primat donné à la praxis et au sémiotique, par opposition à l'ontologie et à la référence.
- Une **attention continue à la demande sociale**, caractérisée notamment par des projets en (i) ingénierie des connaissances médicales (fouille de textes en fœtopathologie pour la valorisation de la recherche médicale : ANR ACCORDYS (2012-2017), en (ii) veille sanitaire (par exemple ANR DontDoIt détection des comportement suicidaires sur les réseaux sociaux, soumis deux fois par le LIRMM, non retenu). À cette attention portée à la demande sociale, s'ajoute naturellement une attention portée à la demande économique, notamment par le biais de thèses CIFRE (7 thèses CIFRE sur la période 2012-2017, 4 soutenues, 3 en cours).

- Un intérêt marqué pour les **nouveaux usages et les nouvelles pratiques du texte** (Internet, réseaux sociaux, etc.) où la linguistique des textes a tout intérêt à construire ses problématiques (e-réputation, extraction d'information sur des forums de discussion, identification des autorités, aide à la compréhension des textes numériques, etc.). La linguistique pratiquée à l'ERTIM est une linguistique des textes considérés comme des objets culturels et confinant à une sémiotique des cultures. Témoins des cultures, les textes sont étudiés dans toutes leurs dimensions (production, interprétation, traduction, transformation, etc.).

- Le **Master TAL** dont l'ERTIM est Equipe d'Accueil, met chaque année sur le marché du travail une quinzaine de diplômés qualifiés dans les domaines du traitement de corpus multilingue et de ses applications (fouille de textes, agrégation de contenus, analyse de sentiments, élaboration de terminologie, veille, lexicologie). Au total, et depuis sa création (comme DESS dans les années 80), ce sont plus de 500 ingénieurs linguistes qui ont été mis sur le marché du travail, principalement en Ile-de-France mais aussi aux Etats-Unis, au Canada, en Europe. Ils constituent un réseau d'alumini précieux.

- La **valorisation ingénierique des langues enseignées par l'Inalco**. La quasi-totalité des doctorants de l'ERTIM sont issus de la formation TAL et travaillent sur des langues du domaine Inalco. Ils sont ainsi à même de répondre aux demandes de l'industrie concernant la maîtrise informatique de ces langues (problématiques du multilinguisme, des écritures et des encodages, étiqueteurs, etc.). Outre un intérêt ingénierique, le bénéfice scientifique est de pouvoir confronter les modèles de la linguistique des textes aux langues variées maîtrisées par les étudiants.

2. Produits de la recherche et activités de recherche

Bilan scientifique

L'ERTIM n'est pas divisée en sous-équipes mais adopte une structuration en 4 axes stratégiques auxquels participent ses différents membres.

Axe 1 : Sémantique de corpus pour les applications TAL

Cet axe a pour contexte général l'articulation de la théorisation en linguistique et du TAL comme technologie. Il vise notamment à évaluer en quoi les propositions théoriques des linguistiques du texte (et notamment de la sémantique textuelle, F. Rastier, 2001, *Arts et sciences du texte*, Paris, PUF) peuvent participer à la résolution de problèmes posés par les applications du TAL et de l'ingénierie linguistique.

Le TAL offrirait aujourd'hui de sérieux débouchés sociaux et économiques à la linguistique si le rôle de celle-ci ne s'y réduisait pas considérablement depuis quelques années. Il s'apparente en effet, de plus en plus, à un rôle de sous-traitance dans le TAL applicatif. Les applications en fouille de textes, par exemple, ont peu recours à la théorisation linguistique et encore moins aux théories du texte (linguistique textuelle, analyse du discours, sémantique interprétative). Le TAL actuel transforme peu à peu les linguistes en *annotateurs experts* et écarte tacitement leurs propositions théoriques du domaine. En élaborant de puissantes méthodes d'extraction des connaissances et de classification de textes, c'est l'informatique qui, aujourd'hui, est confronté à la complexité des textes, à la textualité. Les linguistes de corpus sensibles au texte et soumis à son principe de réalité, sont sans doute plus à même de résister à ce désengagement de la discipline dans le TAL, à condition toutefois qu'ils prennent la mesure des enjeux applicatifs du TAL et de son exigence en termes d'évaluation. C'est dans ce cadre-là que nous entendons, dans l'axe 1 de l'équipe, proposer des solutions de traitement automatique s'appuyant sur une théorisation linguistique du texte. Il s'agit notamment d'élaborer des méthodologies de traitement de corpus pour des applications en recherche d'information, classification de documents et fouille de textes.

Travaux correspondants

- ANR AAP générique TALAD (AGORA, Cergy-Pontoise, coordinatrice, ANR), 2017-2021 : Analyse et traitement automatique de discours.

- Thèse de Kévin Deturck (CIFRE Viseo, Grenoble) : « Détection de personnes influentes dans une sélection de médias sociaux » (2017-2020), dir. M. Valette, F. Segond, D. Nouvel (co-encadrement).

- Thèse d'Amélie Martin (CIFRE SNCF, Paris) : « Méthodologie d'analyse textuelle semi-automatisée des discours voyageurs pour la qualification des déplacements multimodaux » (2015-2018), dir. F. Segond, M. Valette.

- Thèse de Qinran Dang (bourse China Scholarship Council) : « Analyse textuelle de corpus de discours écologiques relatifs au wu mai (brouillard de pollution) en Chine au moyen de méthodes de fouilles de textes » (2015-2019), dir. M. Valette, co-encadrement N. Turenne (LISIS Paris-Est)

- Thèses de LiYun Yan (sans financement) : Fouille d'opinion en chinois, dir. M. Valette, co-encadrement C. Grouin (LIMSI, Orsay).

- Thèse d'Océane Ho Dinh (contrat doctoral actuellement sans emploi) « Méthodes et outils pour le traitement automatique du vietnamien. Application en Humanités Numériques : fouille comportementale sur le web social » (2011-2017), dir. M. Valette.

- Thèse d'Aurélien Lauf (CIFRE actuellement ingénieur R&D chez AMI Software, Montpellier) « Evolution du « buzz » sur internet : identification, analyse, modélisation et représentation dans un contexte multilingue » (2010-2013), dir. M. Valette.

- Thèse de Jugurtha Aït Hamlat (CIFRE actuellement ingénieur R&D chez SemantiWeb, Paris) « Analyse et structuration des récits d'expérience issus du web en français, anglais et arabe, en vue de leur classification thématique » (2010-), dir. M. Valette, D. Nouvel (co-encadrement).

- Thèse d'Egle Eensoo (ingénieur R&D chez XIKO, Paris) « Caractérisation sémantique de textes pour la recherche d'information multilingue – recherche d'une méthodologie » (2008-abandon), dir. M. Valette.

Autres

- Soumission infructueuse H2020 DontDoIt (LIRMM Montpellier coordinateur), 2015 : projet de veille sanitaire (prévention du suicide).

- Soumission infructueuse ANR DoNoDoIt (LIRMM Montpellier coordinateur), 2015, 2016 : projet de veille sanitaire (prévention du suicide), version française du H2020 *supra*.

Axe 2 : Acquisition de connaissances

Cet axe a pour objet l'élaboration de méthodes et leur mise en œuvre pour l'acquisition et le traitement de corpus multilingues et multi-écritures (méthodes et outils), la reconnaissance et l'extraction d'informations linguistiques (structuration de lexiques et de terminologies/ontologies, etc.), la détection d'informations ciblées (« entités » d'intérêt) dans des textes de spécialité. La détection d'informations ciblées dans des textes de spécialité demande de recenser la terminologie en usage dans le domaine concerné dans la ou les langues concernées. L'ERTIM, qui a été force de proposition en terminologie textuelle (D. Bourigault D. et M. Slodzian, 2000, « Pour une terminologie textuelle », *Terminologies Nouvelles*, n° 19 : 29-32), s'appuie pour cela à la fois sur l'analyse des termes utilisés dans les corpus du domaine et sur un recensement des bases terminologiques pertinentes existantes. La dimension multilingue implique de plus la prise en compte des spécificités des langues concernées, en particulier en termes d'écriture (caractères) et de segmentation (délimitation des mots graphiques). Ces méthodes pour collecter des ressources terminologiques et les appliquer à l'extraction d'informations ciblées ont été mises au point.

Travaux correspondants

- Projet ANR CONTINT ACCORDYS (2012-2017) (INSERM, Paris, coordinateur) : fouille de textes en fœtopathologie pour la valorisation de la recherche médicale.

- Thèse de Bénédicte Diot-Parvaz (contrat doctoral) « Élaboration d'une terminologie localisée pour l'aide à l'accès au droit et l'intégration des migrants hindiphones, ourdophones et pendjabiphones » (2016-2019), dir. A. Montaut (SeDyL, CNRS-Inalco) et M. Valette.

- Thèse de Zhen Wang (CIFRE actuellement ingénieure R&D chez GEOL Semantics, Paris) « Extraction en langue chinoise d'actions spatio-temporalisées réalisées par des personnes ou des organismes » (2010-2016) (langue traitée : chinois), dir. P. Zweigenbaum.

- Thèse de Pierre Marchal (contrat doctoral actuellement ingénieur R&D chez SAP, Boston) « Extraction de lexiques bilingues français/japonais à partir de corpus parallèles et comparables » (2010-2015) (en codirection INALCO – Waseda University), codir. Th. Poibeau, Y. Lepage.

- Thèse de Gaël Patin (CIFRE, actuellement directeur R&D XIKO, Paris) « Extraction interactive et non supervisée de lexique en chinois contemporain appliquée à la constitution de ressources linguistiques dans un domaine spécialisé » (2007-2013), dir. P. Zweigenbaum.

- Thèse de Mani Ezzat (CIFRE, actuellement ingénieur R&D chez Dassault Systèmes, Paris) « Passage de données non structurées à des données structurées : acquisition de relations entre entités nommées à partir de corpus » (2008-2013), dir. Th. Poibeau.

- Application en ligne pour l'acquisition de schémas prédicatifs verbaux en japonais à partir d'un corpus non étiqueté de textes journalistiques : http://marchal.er-tim.fr/ikf – prototype présentée en déc. 2015, documentation et diffusion en préparation), réalisée par P. Marchal.

Autres

- Soumission infructueuse ANR TOP « Trust Opinions » (ERTIM coordinateur, montage en collaboration étroite avec VISEO), 2015 : projet de fouille de données web (identification d'experts)

Axe 3 : Technologies éducatives et apprentissage des langues

Cet axe vise la recherche en technologies éducatives et la conception de méthodes et d'outils d'apprentissage des langues fondés sur la création de ressources intégrant des techniques de corpus et de TAL. Ayant réalisé plusieurs applications pour l'apprentissage en ligne (ALPCU, GAELL, A. Chalvin, E. Eensoo et F. Stuck, 2013, « Mining a parallel corpus for automatic generation of Estonian grammar exercises », *eLex proceedings*,), l'ERTIM a orienté ses

recherches sur l'articulation entre la lecture en langue seconde (L2) et l'acquisition incidente de compétences linguistiques. Elle prend ainsi en compte d'une part les pratiques contemporaines telles que la lecture électronique en situation nomade ou l'utilisation des médias sociaux et d'autre part, les développements récents de l'informatique (robustesse des outils du TAL dans certaines langues, stockage sur serveur distant) et la disponibilité d'abondantes ressources textuelles numérisées. Ces travaux se focalisent en particulier sur les mesures de la lisibilité et la mise en place de dispositifs de guidage interprétatif pour la lecture.

Travaux correspondants

- Thèse de Jennifer Lewis-Wong (sans financement) : « Fréquence lexicale et lisibilité de textes en L2 : une étude comparative de textes birmans et anglais » (2015-2018), codir. S. S. Hnin Tun (LACITO, CNRS) et M. Valette.
- Thèse de Satenik Mkhitaryan (contractuelle IGE ERTIM) : « Lisibilité de textes et recherche automatique de contenus pédagogiques : le cas du hindi » (2014- actuellement en césure), dir. M.Valette, encadrement : F. Stuck.
- Thèse de Nadia Makouar (ATER Rouen, actuellement ingénieure ICiMa, Amiens) « Ressources textuelles pour la langue arabe et méthodologie de constitution de corpus avec la sémantique textuelle : application au e-learning » (2009-2014), codir. D. El-Khattab (Université Hassan II Mohammedia), M. Valette.
- Générateur d'exercices automatique GAELL, un générateur d'exercices grammaticaux et lexicaux d'estonien, en collaboration avec la Section d'Estonien de l'INALCO : http://www.er-tim.fr/estonie/html/index.php (mise en ligne déc. 2014).
- Cours/méthodes en ligne d'estonien, bulgare, slovaque, slovène (suivant la méthodologie du projet Socrates Lingua ALPCU), http://www.er-tim.fr/~stuck/coursAC/index.html / /indexSK.html etc. (2015-2016)
- Maquette DEJA LU : maquette pour une application d'aide à la lecture en L2 : http://www.er-tim.fr/~dejalu/_maquette/html/dejaLu.html (login/mdp « visiteur ») (mise en ligne sept . 2017)
- Projet USPC SPECIALIST « Specialised Languages for academic Institutions' Speech and Text » (CLILLAC ARP coordinateur), (2015-2018) : didactique de la traduction en langues de spécialité (projet lauréat mais sous doté en raison de la non-reconduction de l'Idex USPC)

Autres

- Soumission infructueuse USPC alREADy « Réalisation d'une preuve de concept pour un dispositif d'aide à la compréhension de textes en L2 : mise en œuvre sur l'anglais en secteur LANSAD et formation continue » (ERTIM coordinateur), 2012.
- Soumission infructueuse ANR DEJA LU « Dispositif en ligne d'aide à la lecture en langue seconde » (ERTIM coordinateur), 2013.

Axe transversal : corpus et multilinguisme

La plupart des travaux correspondants à cet axe sont ventilés dans les autres axes, mais on mentionnera ici ceux qui apparaissent spécifiques : enjeux théoriques et pratiques du TAL multilingues et mise en œuvre de techniques de TAL pour le traitement de corpus de langues peu dotées.

Cet axe constitue l'une des principales originalités de l'ERTIM dans le paysage scientifique : l'équipe bénéficie de sa position stratégique au sein de l'Inalco et s'adosse à la formation du Master Ingénierie Linguistique qui lui permet de drainer des étudiants de diverses origines, internationales et/ou apprenants de langue orientale de l'Inalco. Les thèmes abordés sont les enjeux théoriques et pratiques des corpus multilingues (parallèle et comparable), la problématique du multilinguisme dans le traitement automatique du document numérique et la prise en compte technique des spécificités associées (écritures, encodages). Plusieurs réalisations de l'axe transversal sont liées à des recherches menées dans d'autres axes.

Travaux correspondants

- Plateforme USPC MultiTAL « Plateforme de documentation et d'expertise des outils et ressources pour le traitement automatique des langues orientales et des langues peu dotées » (ERTIM coordinateur) (2015-) (1 post-docs, 2 IGE, 11 stages) http://multital.Inalco.fr
- Projet USPC APRECIADO « Analyse et spatialisation des Perceptions et Représentations sociales des Changements environnementaux en Afrique De l'Ouest sahélo-soudanienne » (2013-2016) PLEIADE Paris Nord, coordinateur.
- Projet Inalco MANTAL « Analyse morphosyntaxique du bambara à partir d'un corpus partiellement désambiguisé et de techniques d'apprentissage automatique » (2014-2017), en collaboration avec le LLACAN, 5 stages.

- Thèse d'Asma Zamiti (contrat doctoral) : « Conceptions d'outils pour le traitement automatique du tunisien (arabizi) : application à l'analyse des réseaux sociaux en Tunisie » (2015-2018), dir. M. Valette.

- Collaboration avec le projet ANR ALIENTO (CERMOM) : co-encadrement d'un stage « Fouille de motifs et translittération de l'arabe »

Autres

- Projet USPC TAL-SHS « Outils de Traitement Automatique des Langues pour les SHS » (LIPN, Paris Nord, coordinateur) (2015-2018) (projet lauréat mais sous doté en raison de la non-reconduction de l'Idex USPC).

Faits marquants

(1) La plateforme numérique MultiTAL « Plateforme de documentation et d'expertise des outils et ressources pour le traitement automatique des langues orientales et des langues peu dotées » http://multital.Inalco.fr

MultiTAL est une plateforme experte en traitement automatique des langues (TAL) focalisée sur les langues orientales et/ou peu dotées. Elle référence et documente une sélection d'outils de TAL, qui ont été testés et renseignés par l'équipe. La plateforme permet d'identifier et d'installer rapidement un outil en fonction de multiples critères (tâche, langue, système, méthode, etc.). Pour chaque outil, une documentation multilingue standardisée décrivant son installation et son exécution est générée automatiquement. La documentation suit un protocole normalisé qui est indépendant des documentations des auteurs et peut être générée en plusieurs langues.La plateforme est mise à jour régulièrement. En plus de pouvoir être consultée en ligne, une ontologie décrivant les différents outils peut être interrogée. Actuellement, la plateforme repose sur l'évaluation de 165 outils, dont 91 ont été sélectionnés pour publication. Ces outils permettent d'exécuter 142 tâches. 47 langues sont concernées.

(2) Le projet APRECIADO (Paris Nord, Paris Diderot, Inalco) : Analyse et spatialisation des Perceptions et des REprésentations des Changements environnementaux Intervenus en Afrique De l'Ouest depuis l'indépendance. APRECIADO est un projet qui a réuni des géographes (PLEIADE, Paris Nord et PRODIG, Paris Diderot, et des linguistes (ERTIM). Le projet se proposait d'étudier la perception qu'ont les acteurs (populations, gestionnaires, scientifiques, organisations non gouvernementales, etc.) des changements environnementaux intervenus en Afrique de l'Ouest depuis les indépendances (variation de la pluviosité, évolution de l'occupation et de l'utilisation du sol, déprise rurale, croissance urbaine, etc.) en analysant un corpus d'entretiens en diverses langues vernaculaires (peul, wolof, djerma). L'ERTIM a procédé à la sélection, à la transcription et à l'analyse textométrique des corpus fournis par l'équipe de géographes. Le projet a été conclu par deux manifestations scientifiques internationales :

1. Colloque international « *Changements socio-environnementaux et dynamiques rurales en Afrique de l'Ouest* », Paris, INALCO / Paris Diderot, 4 et 5 juillet 2016.
2. Atelier « *Environnement sahélo-soudanien en changement et devenir des espaces ruraux* » (6 juillet 2016, dans le cadre des 4e Rencontres des Etudes Africaines en France, Paris (INALCO), 5-7 juillet 2016.

(3) La création et l'organisation du HackaTAL, le premier Hackathon [1] français dans le domaine du TAL (http://hackatal.github.io/2016/), à l'occasion de la conférence TALN 2016. Des acteurs majeurs du secteurs comme Google et Systran y ont participé. Le HackaTAL a eu lieu dans les locaux de Google Paris. Les tâches concernaient la détection d'évènements et l'implémentation de système de gestion de dialogue.

(4) La co-organisation de deux conférences majeures dans les champs disciplinaires de l'équipe :

- Les *Journées internationales d'Analyse statistique des Données Textuelles* (JADT) qui réunissent tous les deux ans, depuis 1990, des chercheurs travaillant dans les différents domaines concernés par les traitements automatiques et statistiques de données textuelles. L'ERTIM a co-organisé avec le CLESTHIA de Paris Sorbonne Nouvelle l'édition 2014 qui s'est déroulée sur le principal de l'Inalco (65 rue des Grands Moulins).

- Les conférences conjointes TALN (Conférence sur le Traitement Automatique des Langues), JEP (Journée d'Etudes de la Parole) et RECITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues). L'ERTIM a coordonné l'édition 2016 de RECITAL et participé activement à l'organisation de TALN et JEP, les trois conférences se sont déroulées sur le principal de l'Inalco (65 rue des Grands Moulins).

[1] « hackathon désigne un événement où un groupe de développeurs volontaires se réunissent pour faire de la programmation informatique collaborative, sur plusieurs jours. C'est un processus créatif fréquemment utilisé dans le domaine de l'innovation numérique » (source : Wikipédia).

3. Organisation et vie de l'unité

Pilotage, animation, organisation de l'unité

Vie de l'équipe

L'équipe comporte peu de permanents (1 PU, 3 MCF, 1 PRAG 1 IR, ½ assistante) mais les contractuels constituent entre 2/3 et 4/5 des effectifs, suivant les périodes (doctorants, post-docs, recrutements sur projet, stagiaires, etc. auxquels s'ajoute ½ PAST de façon récurrente). Elle fonctionne dans une large mesure sur le modèle des laboratoires de sciences, c'est-à-dire un travail collaboratif avec une présence permanente de beaucoup de ses membres. Les doctorants qui ne sont pas salariés par ailleurs sont présents presque quotidiennement, les doctorants sous contrats CIFRE viennent ½ journée à 1 jour par semaine suivant la convention établie et les besoins qu'ils expriment.

Gouvernance

L'équipe a longtemps fonctionné selon une gouvernance pragmatique guidée par les projets sur appel d'offre. Malgré une organisation en axes (stabilisée depuis l'AG du 7 décembre 2011), les projets financés demeurent le vecteur de structuration essentiel. Quoique perçue comme une exigence plus administrative que scientifique dans cette petite équipe, la gouvernance par axe s'installe peu à peu : chaque projet soumis à une agence de moyens l'est dans le but de dynamiser un axe.

En parallèle aux réunions de projets ou d'encadrement, nombreuses et non dénombrées, l'équipe organise deux Assemblées Générales par an. Depuis 2015, les réunions sont le plus souvent adossées à des séminaires internes (doctorants). Début 2016, deux listes de diffusion interne ont été créées, l'une concerne l'ensemble des membres de l'équipe, l'autre les seuls doctorants. Pour des raisons statutaires, l'équipe est dotée d'un conseil composé de deux représentants doctorants élus, d'un représentant IATOSS, de deux représentants EC, d'un directeur et d'un directeur adjoint. Dans les faits, la presque totalité des permanents font donc partie de ce conseil.

L'équipe organise 4 à 6 fois par an un séminaire de recherche avec des intervenants extérieurs, ouvert à tous.

Parité

Parmi les membres permanents d'ERTIM, la parité homme/femme est respectée dans chaque catégorie. Quant à la direction de l'équipe, la tendance est à l'alternance : après une longue direction féminine, elle est aujourd'hui assurée par un homme. Côté doctorants (en cours ou diplômés), les femmes sont surreprésentées (66% contre 33%). Ceci s'explique par l'audience très largement féminine de la formation adossée à l'équipe, principal vivier du recrutement des doctorants d'ERTIM.

Protection et sécurité

L'unité de recherche suit les préconisations du site de la rue de Lille où elle est hébergée. Le site est soumis à des règles strictes de sécurité (contrôles des entrées, gardiennage, caméras) ce qui assure la protection indispensable à un fonctionnement conforme à la réglementation. Un registre de santé et de sécurité au travail est disponible à l'accueil. Y sont consignées toutes les observations et suggestions relatives à la prévention des risques et à l'amélioration des conditions de travail. Un registre Spécial de Signalement d'un Danger Grave et Imminent est en revanche réservé uniquement aux membres du personnel et disponible au secrétariat de la Présidence et du DGS. Les risques les plus importants concernent les missions effectuées sur le travail de terrain et plus particulièrement pour les chercheurs qui peuvent avoir à se déplacer dans des régions à risques. La direction générale des ressources humaines se charge de vérifier le lieu de séjour du missionnaire et consulte le site du ministère des affaires étrangères http://www.diplomatie.gouv.fr/fr/conseilsauxvoyageurs_909/index.html. Les missionnaires sont invités, pour tout départ en mission à l'étranger, à s'enregistrer sur la plate-forme Ariane du MAE

4. Analyse SWOT

Points forts

- Une position stratégique unique dans le paysage scientifique, à la fois de par son ancrage à l'INALCO (ingénierie multilingue) et par l'originalité des propositions théoriques défendues, notamment, la sémantique textuelle de F. Rastier, qui propose une alternative et nourrit un dialogue avec les approches dominantes en TAL et ingénierie des connaissances et ouvre de nouvelles perspectives vers une *sémiotique des cultures*[2], particulièrement pertinentes dans le contexte Inalco.

[2] Cf. Rastier, F. & S. Bouquet, *Une introduction aux sciences de la culture*, Paris PUF, 2002.

- Une excellente réputation, aussi bien dans le monde académique qu'industriel, qui se manifeste par de nombreux projets collaboratifs et de fréquentes sollicitations auxquelles l'ERTIM ne peut pas toujours répondre, faute de ressources humaines suffisantes (lire *infra* les points faibles) ;
- Une culture du projet se concrétisant par une activité de soumission, de participation et de coordination de projets de recherche régulière ;
- Une forte attractivité, notamment doctorale ;
- Un solide adossement à l'enseignement (Master TAL).

Progrès observables depuis le 1er janvier 2013 (en référence à l'analyse SWOT de 2012)

- Une meilleure capitalisation des connaissances et des compétences relevant des ressources humaines contractuelles et doctorales (par nature non pérennes), grâce notamment à la **plateforme MultiTAL**.
- Davantage de collaborations au sein de l'Inalco (GAELL, MANTAL, ALIENTO, etc.).

Points faibles

- **Des effectifs permanents insuffisants** ; en dépit du recrutement d'un MCF actif et publiant en remplacement d'un non publiant, les effectifs permanents de l'ERTIM apparaissent toujours insuffisants. Le recrutement d'un MCF, voire d'un PU (pour renforcer l'encadrement et permettre un renouvellement à la direction de l'équipe) serait très opportun.
- Une activité de publication encore un peu insuffisante compte tenu des standards en TAL tout du moins. Les publications d'un membre associé très actif masque ce relatif manque de productivité.
- Une stratégie de publication à parfaire (choix des supports de publication).

Possibilités liées au contexte

- L'Inalco lui-même : plus de 90 langues enseignées, une université unique dans le paysage européen, des opportunités de collaboration et de recrutement d'étudiants variées.
- La mise en œuvre de nouveaux modes de partenariats entre l'université et les industriels « conception contre développement » : l'équipe participe à l'élaboration et la conception d'outils logiciels de traitement automatique qui sont développés ensuite par l'industrie.
- L'ERTIM a entamé un dialogue avec la SOAS (School of Oriental and African Studies, Londres) laquelle ne dispose pas d'équipe de TAL, dans la perspective de mettre en place des collaborations et des échanges d'étudiants.

Risques liés au contexte

- Une perte de l'identité scientifique de l'équipe, induite par exemple par une union contrainte ou si l'équipe n'affirme pas suffisamment son identité scientifique en renforçant notamment son activité éditoriale et en améliorant encore la capitalisation de ses compétences.

Analyse

Les résultats et l'auto-évaluation de l'équipe font ressortir une bonne cohérence du projet scientifique et sa pertinence compte tenu de son environnement, au sein de l'Inalco en premier lieu, mais aussi dans un contexte plus large : Comue USPC, réseau industriel, partenariats nationaux et internationaux. Le seul véritable élément de faiblesse de l'équipe tient à son sous-effectif et à ses conséquences : l'équipe permanente est réduite, le temps de recherche de beaucoup de ses membres est mis en concurrence avec des tâches d'administration (forte implication dans les instances de l'Inalco) et d'animation pédagogique, notamment dans le contexte d'une formation dont l'offre est très complète (L2/L3, M1/M2, D). En conséquence, l'activité de publication peut sembler un peu insuffisante au regard de l'activité réelle.

La stratégie de l'équipe mise en œuvre pour pallier ces faiblesses sur l'exercice 2014-2018 comprend les points suivants :

A. Le renforcement de l'activité éditoriale. Elle constitue un des défis majeurs d'une équipe dont le projet scientifique garde toute sa légitimité, *a fortiori* à un moment où les regroupements effectués font peser un risque de normalisation sur les positions scientifiques originales. Les nombreuses collaborations avec de grands laboratoires (LIMSI, INSERM, PLEIADE, PRODIG) sont la preuve que les compétences de l'ERTIM sont prisées, même indépendamment de son expertise multilingue. Pour améliorer l'activité éditoriale de l'équipe, plusieurs solutions sont actuellement mises en œuvre (liste de diffusion, présentation des travaux en interne et veille sur les supports de publications).

B. La valorisation systématique des productions de l'équipe (logiciels, ressources, corpus) en particulier celles des contractuels et des stagiaires. La plateforme MultiTAL constitue à cet égard une avancée majeure pour l'équipe, qui a été également saluée par la communauté.

C. Le site Web de l'équipe (http://www.er-tim.fr), constitue un outil de diffusion et de structuration des travaux plus souple à maintenir que celui de l'Inalco (http://www.inalco.fr/ertim). Sa mise à jour est actuellement satisfaisante. L'ouverture pendant l'exercice 2012-2017 de comptes LinkedIn et Facebook complète cette stratégie de communication à l'intention de l'industrie principalement.

D. Mais cela ne remplacera pas **une campagne de recrutement de permanents** ambitieuse, à tous les niveaux : ingénieur (corps stratégique dans une équipe à forte dominante technologique et qui n'en comprend qu'un en permanent), EC (une demande conjointe de MCF avec la formation Sciences du langage a été faite deux années consécutives sans résultat). Le rapprochement de l'ERTIM avec une autre équipe parisienne, bien que non crucial, a été étudiée mais aucun projet n'est apparu suffisamment mûr pour le moment.

5. Projet scientifique à cinq ans

Le projet scientifique de l'équipe se situe dans la continuité de celui proposé en 2012 sur deux versants (*les humanités numériques* et *le TAL pour les applications*), et ajoute une nouvelle dimension, *l'outillage des langues peu dotées*.

L'outillage des langues peu dotées

La description des langues et les typologies linguistiques existantes ne recoupent pas celles requises pour leur outillage. Par exemple, des langues non apparentées peuvent nécessiter des outils similaires ou certains outils communs du point de vue de leur traitement automatique : l'estonien (langue finno-ougrienne, fennique) peu outillé est une langue agglutinante et compositionnelle au même titre que l'allemand (indo-européenne, germanique) nécessitant un analyseur morphologique reconnaissant les affixes et les composants internes des mots ; le chinois (sino-tibétaine) et l'anglais (indo-européenne, germanique) sont toutes deux des langues isolantes qui requièrent un regroupement des mots graphiques en unités de sens complexes. Par ailleurs, de multiples problèmes bloquants sont souvent méconnus ou sous-estimés dans le traitement informatique des langues (encodage, sens de l'écriture, ligature, segmentation, etc.).

Notre hypothèse est qu'une connaissance approfondie et exhaustive des outils de TAL selon un référentiel commun, permet non seulement l'adaptation et la spécification d'outils pour le traitement des langues moins dotées (ex. l'adaptation d'un étiqueteur morphosyntaxique statistique existant à une nouvelle langue) mais aussi des innovations en matière de méthodes et d'algorithmiques utilisables sur toutes les langues a priori (ex. les modèles statistiques de traitement du corpus). Forte de cette expérience, qui nous a permis, à titre exploratoire, d'adapter un outil d'annotation morphosyntaxique pour le traitement du tibétain et d'autres travaux similaires (projet MANTAL pour le traitement automatique du bambara, thèse d'Asma Zamiti pour le traitement automatique du tunisien arabizi), l'ERTIM a pour ambition d'acquérir et de capitaliser un savoir-faire lui permettant de créer de nouveaux outils d'annotation.

Dans cette perspective, la plateforme MultiTAL développée par l'ERTIM, plateforme experte en matière de traitement technologique des langues, a été mise à la disposition de la communauté en janvier 2017. Elle propose des connaissances actualisées et une expertise critique relatives aux outils de traitement automatique des langues (programmes, composants logiciels, etc.) disponibles, en particulier (i) lorsqu'elles relèvent du domaine Inalco ; (ii) lorsqu'elles sont peu dotées.

Humanités numériques

Les humanités numériques — les premiers pas des humanités numériques relevaient d'ambitions à la fois patrimoniales, éditoriales et documentaires. Il s'agissait de collecter, numériser et organiser de documents pour les rendre interrogeables (indexation, navigation). Les projets ont ensuite porté sur la normalisation des bases textuelles avec l'établissement de formats d'échange et de normes d'encodage, lesquels ont facilité des travaux d'annotations philologiques permettant de complexifier les outils d'interrogation.

L'utilisation, l'adaptation et la création d'outils et de méthodologies de linguistique de corpus adaptées à l'herméneutique des textes constituent l'enjeu actuel des humanités numériques. Il s'agit désormais de développer des méthodes d'aide à l'interprétation des textes, s'inspirant à la fois de la philologie et de l'herméneutique traditionnelle et des méthodes TAL en fouille de textes. En 2011, l'ERTIM a bénéficié d'un contrat doctoral fléché afin de favoriser l'émergence de cette problématique au sein de l'Inalco. Le projet était consacré à la réalisation de méthodologies de fouille sémantique de corpus comparables multilingues et multi-écritures (comportements sanitaires de la jeunesse vietnamienne tels qu'ils s'expriment sur les forums de discussion - thèse d'Océane Ho Dinh). Cette

recherche a constitué la première pierre d'une approche des humanités numériques prenant en compte les transformations de l'écrit et anticipant sur les futurs matériaux textuels avec lesquels les chercheurs en sciences humaines et sociales ont à travailler désormais : les corpus sont constitués à partir de données issues du Web afin de capter au mieux les tendances culturelles et linguistiques. A la suite de cette thèse, d'autre travaux ont été lancés (thèse de Qinran Dang). Le projet « humanités numériques » de l'équipe s'appuie sur son savoir-faire et son expertise en matière d'analyse des documents du web social. La participation de l'équipe à des projets interdisciplinaires comme ALIENTO (avec des philologues) et APRECIADO (avec des géographes) atteste de cette dynamique. Un membre de l'équipe, par ailleurs, est en train de mettre en place la revue *Analyses et méthodes formelles pour les humanités numériques* (http://www.openscience.fr/Analyses-et-methodes-formelles-pour-les-humanites-numeriques).

TAL pour les applications

Didactique des langues et les nouveaux usages nomades — Les usages liées aux TIC invitent à multiplier les opportunités d'apprentissage et de formation, en tenant compte des variétés des situations : collaborative ou individuelle, sédentaire ou nomade, en contexte formel ou non formel. Prenant acte des limites des outils TICE actuels et forte de son expérience acquise au cours des projets SOCRATES LINGUA (notamment ALPCU, 2003-2007), et des différentes applications réalisées ensuite pour l'Inalco (GAELL, mise en ligne des méthodes ALPCU, maquette DEJA LU, thèses de S. Mkhitaryan et J. Wong sur la lisibilité), l'ERTIM étudie actuellement, avec l'appui de la SATT IdF Innov la réalisation de méthodes dont la vocation est l'acquisition de compétences partielles en réception de l'écrit. Il est beaucoup plus facile pour les apprenants de se confronter à une grande variété de textes en langue seconde (L2) via Internet, plutôt que d'interagir oralement avec des locuteurs de la L2. S'initier à une langue et à une culture par la lecture est une pratique millénaire facilitée et renouvelée aujourd'hui grâce à la banalisation de la compétence informationnelle (littératie) et l'accessibilité des ressources documentaires sur Internet. Il est possible de concevoir une méthode d'aide à la lecture transposable à d'autres langues, en adaptant des outils de TAL et de linguistique de corpus éprouvés, destinés initialement à d'autres champs applicatifs (linguistique descriptive, ingénierie des connaissances, etc.).

L'ingénierie des connaissances et des informations subjectives — Elle constitue un domaine d'investigation historique par l'ERTIM depuis sa création. L'enjeu est de concevoir des méthodes de détection et d'extraction, pour construire des bases d'information à partir de données textuelles multilingues en agrégeant des informations hétérogènes et linguistiquement enrichies. Beaucoup des recherches en cours par l'équipe entrent de plain-pied dans ce cadre, en proposant plusieurs travaux d'extraction d'informations linguistiques. Le projet ANR ACCORDYS (2012-2017) a constitué une instanciation récente de cette recherche. La thèse de Bénédicte Diot-Parvaz sur la constitution d'une terminologie juridique multilingue français-hindi-ourdou participe également de la pérennisation de cette recherche. L'ERTIM approfondit aussi ses recherches menées en extraction d'informations subjectives. Les thèses d'Egle Eensoo, Amélie Martin et Liyun Yan y participent.