agence d'évaluation de la recherche et de l'enseignement supérieur

Vague D :
campagne d'évaluation 2012 - 2013

EA 2520 ERTIM
Équipe de Recherche Textes, Informatique, Multilinguisme

2.1. Projet scientifique de l'unité

1. Présentation de l'unité

a. Historique

L'ERTIM est une Équipe d'Accueil née en 2005 de la fusion de deux centres de recherches de l'INALCO : le Centre de Recherche en Ingénierie Multilingue (CRIM, créé en 1986 par Monique Slodzian) et le Centre d'Études et de Recherche en Traitement Automatique des Langues (EA CERTAL, créé en 1982 par Patrice Pognan). L'ERTIM hérita dans une large mesure de la culture scientifique du CRIM. L'équipe est adossée au département Textes, Informatique, Multilinguisme (TIM) qui offre une formation L2/L3 (Licence « Langue du Monde et Formation Appliquée », mention « Traitements Numériques Multilingues »), M1/M2 Master mention « Science du langage et langues appliquées », spécialité co-habilitée « Ingénierie linguistique » comportant 3 parcours (2 professionnels, 1 recherche) et le doctorat.

b. Caractérisation de la recherche

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 toujours eu pour objectif de répondre à des problèmes applicatifs concrets. A titre d'exemple, dans les années 1990, la terminologie textuelle élaborée par M. Slodzian et D. Bourigault visait le problème de l'inadéquation des terminologies de domaines aux usages à partir d'une réflexion théorique sur les genres textuels et un outillage technique (extraction de syntagmes à partir de corpus raisonnés) (Min. Indus. SAFIR). Dans les années 2000, l'instrumentation textométrique des concepts de la sémantique textuelle (F. Rastier) a eu pour objectif la conception de nouveaux algorithmes de filtrage de textes (EC PRINCIP et ANR C-MANTIC). Dans les années 2010, ce sont l'identification des signaux faibles et l'extraction d'information subjective (fouille d'opinion, analyse des sentiments) qui apparaissent comme des terrains de prospection et d'expérimentation porteurs pour l'équipe. 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 par des projets en (i) ingénierie des connaissances médicales (pharmacovigilance : ANR VIGITERMES ; fouille de textes en fœtopathologie pour la valorisation de la recherche médicale : ANR ACCORDYS), (ii) en veille sanitaire (prévention contre le tabagisme : ANR C-MANTIC ; comportements liés à la prévention contre le Sida au Vietnam : Contrat doctoral d'O. Ho Dinh), et en (iii) filtrage de contenu préjudiciable et illicite (racisme et xénophobie sur Internet : projet européen *Safer Internet* PRINCIP). À cette attention portée à la demande sociale, s'ajoute naturellement une attention portée à la demande économique — hier, pour la constitution de ressources

terminologiques pour l'industrie (projet SAFIR avec EDF), aujourd'hui, beaucoup par le biais de thèses CIFRE (AMI Software, ARISEM, GEOL SEMANTICS, etc.).

Par ailleurs, le **Master spécialité « Ingénierie Linguistique »** (alias PLURITAL, http://plurital.org) dont l'ERTIM est Equipe d'Accueil, met chaque année sur le marché du travail plus d'une dizaine de diplômés qualifiés en TAL et ingénierie linguistique, 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), la traduction outillée (chef de projet de traduction, localisation d'applications et sites web, traducteur), la gestion de l'information multilingue (veille, webmarketing, documentation et archivage...).

- 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.). L'équipe occupe à cet égard au sein de l'INALCO une position épistémologique originale, entre sciences du langage et sciences de la culture (i.e. les études aréales). En cela, l'héritage intellectuel de F. Rastier, membre de l'ERTIM jusqu'à son départ en retraite fin 2009, constitue une source d'inspiration très stimulante et sans cesse renouvelée. Le projet scientifique de l'ERTIM pour l'exercice 2014-2018 entend poursuivre et approfondir les réflexions menées dans le cadre du réseau Roger T. Pédauque (dont Monique Slodzian était membre) autour de la révolution technologique que constitue la dématérialisation du texte (lecture sur liseuse, tablette, etc.)

- La participation à la **valorisation du patrimoine linguistique (et culturel) de l'INALCO**, laquelle s'instancie d'une part dans l'outillage TAL des langues étudiées et enseignées à INALCO, et d'autre part, par la réalisation d'outils et de ressources didactiques pour l'apprentissage des langues (plateforme et outils d'apprentissage ou d'aide à la lecture, textes didactisés), ou le développement de méthodologies de fouille de textes pour les humanités numériques.

- La **valorisation ingénierique des langues enseignées par l'INALCO**. Le Master spécialité « Ingénierie Linguistique » draine en effet une grande variété d'étudiants, non seulement de l'INALCO et des autres universités partenaires de PLURITAL (Sorbonne Nouvelle-Paris 3, Paris Ouest Nanterre La Défense, Paris Diderot-Paris 7) mais aussi du monde entier. La quasi-totalité des doctorants de l'ERTIM sont issus de la formation et travaillent sur des langues du domaine INALCO (arabe, chinois, estonien, japonais, vietnamien). 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.

- Un réseau de collaborations industrielles et académiques très riche : INSERM, CNRS LIMSI, CNRS LINA, CNRS IRIT, CNRS ATILF, CNRS LIP6, SYLED, IOTA Bamako (Mali), Niamey (Niger), Saint-Pétersbourg (Russie), Tbilissi (Géorgie), Téhéran (Iran), Thessalonique (Grèce), Hagen, Frankfurt, Berlin, Magdebourg, Hildesheim (Allemagne), Uppsala (Suède), Gent (Belgique), Helsinki (Finlande), Naples (Italie), Dublin (Irlande), Maribor (Slovénie), Sofia (Bulgarie), Banska-Bystrica (Slovaquie), Tokyo (Japon), Penang (Malaisie), Tartu (Estonie), Bombai (Inde), etc.

c. Organigramme fonctionnel et règlement intérieur

L'équipe a longtemps fonctionné selon une gouvernance pragmatique guidée par les projets sur appel d'offre. L'organisation en axes date de 2005, année de la fusion du CRIM et du CERTAL. Elle a été remaniée fin 2008 pour la contractualisation 2010-2013 et amendée lors de l'AG du 7 décembre 2011 en prévision du futur quinquennal. Les projets financés demeurent le vecteur de structuration essentiel mais la gouvernance par axe s'installe peu à peu : depuis 2010, chaque projet soumis à une agence de moyens l'est dans le but de dynamiser un axe. Cette stratégie vise à renforcer le rôle des animateurs d'axe, rendus nécessaires depuis que l'équipe accueille des doctorants, tout en préservant la culture de projet qui a fait le succès du CRIM puis de l'ERTIM.

L'exercice 2014-2018 verra le renforcement des outils d'animation actuellement mis en place (réunion d'équipe tous les 15 jours, présentation des travaux en cours, mise en place d'un planning des publications, implication plus forte des doctorants CIFRE actuellement peu présents dans le laboratoire). Le vote du règlement intérieur est prévu à l'ordre du jour de l'AG de l'automne 2012.

2. Analyse SWOT et objectifs scientifiques de l'unité

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 fructueux avec les approches dominantes en TAL et ingénierie des connaissances et ouvre de nouvelles perspectives vers une *sémiotique des cultures*[1], particulièrement pertinentes dans le contexte INALCO.
- une excellente réputation, aussi bien dans le monde académique qu'industriel, qui se manifeste par de nombreux projets collaboratifs et de fréquentes sollicitations ;
- 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, mais qui s'observe aussi au niveau des demandes d'affiliation de type « associée » ;
- un solide adossement à l'enseignement (Master Ingénierie linguistique).

Points faibles

- des effectifs permanents insuffisants ;
- une certaine difficulté à capitaliser des connaissances et des compétences relevant des ressources humaines contractuelles et doctorales, par nature non pérennes ;
- une stratégie de publication à parfaire (choix des supports de publication) ;
- des collaborations au sein de l'INALCO encore un peu rares et sporadiques.

Possibilités liées au contexte (*opportunity*)

- 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 ;
- le LABEX Empirical Foundations of Linguistics (désormais LABEX EFL) que l'ERTIM pourrait rejoindre à brève échéance et, d'une manière générale, le PRES « Université Sorbonne Paris Cité » (désormais USPC) où des collaborations sont actuellement mises en place ; par exemple avec Paris 3 (EA SYLED/CLA^{2}T), Paris 7 (EA CLILLAC-ARP/EILA), Paris 3 (EA DILTEC), Paris 13 (LIPN).
- 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.

Risques liés au contexte (*threat*)

- 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 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 : PRES Sorbonne Paris Cité, 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) (cf. Bilan*)*. En conséquence, l'activité de publication peut sembler un peu insuffisante au regard de l'activité réelle. Cette difficulté ancienne, déjà remarquée par l'AERES en 2009, est également explicable par la culture de projets applicatifs du CRIM (1986-2005) qui était plus attaché à la valeur pratique qu'à la valeur académique, comme en témoignent les productions [105][106][108][109][113].

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

La stratégie de l'équipe mise en œuvre pour pallier ces faiblesses et qui se poursuivra 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, LINA, INSERM) sont la preuve que les compétences de l'ERTIM sont prisées, même indépendamment de son expertise multilingue (l'ANR ACCORDYS porte essentiellement sur le français et l'anglais). Pour améliorer l'activité éditoriale de l'équipe, plusieurs solutions sont actuellement mises en œuvre (en particulier le développement d'outils de partage des connaissances, intensification des séminaires internes avec présentation des travaux 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. À cet égard, une architecture serveur-client destinée à accéder facilement à des traitements contrastifs de corpus multilingues a été conçue sur cahier des charges ERTIM dans le cadre du projet C-MANTIC, et partiellement développée par un prestataire, pour accueillir les programmes développés ou adaptés par l'équipe, à des fins de réutilisabilité. Il s'agit du LPU (Linguistic Processing Unit), un outil de traitement des corpus linguistiques en vue de leur analyse de masse. Il se singularise par les aspects suivants : (i) il est optimisé pour la constitution de corpus à la volée, (ii) il donne la possibilité d'intégrer des traitements définis par les utilisateurs ou d'autres laboratoires, sans être spécialisé dans une gamme de traitements ; (iii) il prend en charge de façon automatisée l'application des traitements aux corpus : les utilisateurs n'ont pas besoin de demander qu'un traitement soit appliqué ; cela est fait par le système en fonction des modèles définis par les utilisateurs dans leurs projets. Le LPU est actuellement en phase de test et devrait être opérationnel courant 2013.

C. Le site Web de l'équipe (http://www.crim.fr), en cours de modernisation, devrait également constituer un outil de diffusion et de structuration des travaux. Il a été conçu sur une architecture CMS destiné à permettre aux chercheurs et aux intervenants du département TIM de mettre en ligne leur productions. Un groupe de travail Site Web a été mis en place en janvier 2012 pour en assurer l'évolution.

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 — François Stuck), maître de conférences (le poste de M. Fanton sera vacant en 2014), voire professeur pour renforcer l'équipe encadrante. Le rapprochement de l'ERTIM avec une autre équipe parisienne, bien que non crucial, est également à l'étude, mais aucun projet n'est suffisamment mûr pour le moment.

3. Mise en œuvre du projet

a. Propositions scientifiques : textes, sémantique et multilinguisme

Du texte numérisé au document numérique — Le passage d'une problématique du *texte numérisé* à celle du *document numérique* constitue un des enjeux de la linguistique de corpus, organon des sciences du langage. La linguistique des textes s'est longtemps consacrée à l'analyse des textes littéraires ou politiques, aux genres globalement bien décrits. Elle est désormais confrontée à une grande variété de discours et de genres nouveaux, indéterminés, polymorphes, souvent multilingues, et en permanente évolution (après avoir périmé les « pages perso », les blogs sont aujourd'hui en voie d'obsolescence). S'il s'agit souvent d'une modernisation de pratiques anciennes, ces genres sont aussi la trace de nouvelles pratiques.

Parmi les linguistiques du texte, la sémantique textuelle participe à ce débat[2]. Ayant pour objet empirique le texte et non le mot, la phrase ou l'énoncé, traditionnellement privilégiés, cette linguistique-science des textes renoue avec une tradition rhétorique et herméneutique oubliée du XXème siècle et se focalise sur l'étude de la textualité, des genres textuels, des discours et de leurs corollaires (cohésion textuelle, intertextualité, etc.). Son appareil théorique est depuis le début des années 90 adossé à la linguistique de corpus et au TAL[3]. L'instrumentation logicielle s'associe ainsi aux outils théoriques et conceptuels. Constitutive de la linguistique de corpus, elle donne lieu à ce qu'on pourrait appeler son « cercle vertueux » : d'un côté, les grandes masses de données textuelles ou documentaires nécessitent, pour être analysées et décrites, des instruments *ad hoc*, de l'autre, cette instrumentation permet de construire de nouveaux observables qui seraient demeurés invisibles autrement. L'interprétation des textes assistée par ordinateur fait certes l'objet d'une riche littérature mais elle est le plus souvent réduite à quelques aspects

[2] Lire Rastier, F. (2011) *La mesure et le grain. Sémantique de corpus*, Paris, Champion.
[3] Lire Valette, éd., F. (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée, Syntaxe & Sémantique, n°9.*

récurrents : désambiguïsation, identification des domaines et des thèmes. D'une manière générale, la question du sens est dominée par une approche lexicale, selon laquelle le sens est dans les mots et se calcule compositionnellement à partir de ceux-ci. Or, ni la problématique du texte ni celle du document numérique ne peuvent s'en satisfaire exclusivement. Elles impliquent de prendre en compte les ensembles de textes, les corpus, l'archive, l'intertexte. Le mot demeure un objet important, notamment dans les langues qui le privilégient, mais il importe de le rapporter à sa juste proportion, c'est-à-dire un fragment du texte.

La forme sémantique comme objectivation linguistique — Entre le mot et le texte, les propositions théoriques de la sémantique textuelle permettent d'étudier la structuration sémantique d'un texte, par le biais d'objectivations sémantiques, c'est-à-dire des réseaux de traits sémantiques qui en assurent la cohésion. La sémantique textuelle en détaille deux catégories : (i) l'isotopie (récurrence d'un même trait sur un empan de longueur variable, de la phrase au corpus) ; (ii), la molécule sémique (patrons stabilisés et récurrents de traits hétérogènes). À l'interface entre le lexique et le texte, les réseaux de traits permettent d'étudier de manière approfondie à la fois le lexique, le texte et leur relation cohésive. Leur intérêt descriptif et applicatif a été montré dans différents contextes[4].

Un programme de recherche pour l'exercice 2014-2018 s'articulera, par conséquent, autour de trois objectifs conjoints : (i) approfondir les connaissances actuelles sur les objectivations sémantiques connues et notamment référencées par la sémantique textuelle, (ii) en identifier de nouvelles que la théorie n'a pas su jusque là reconnaître faute d'une instrumentation adéquate et, enfin, (iii) créer de nouveaux observables sémantiques, textuels et lexicaux. À ces objectifs minimaux, nous adjoignons des considérations épistémologiques (statut gnoséologique des objectivations sémantiques)[5], descriptives (variations et régimes d'élaboration et d'interprétation des objectivations dans des corpus variés et dans différentes langues) et applicatives (exploitation pour l'identification d'objets linguistiques non lexicalisés et échappant de ce fait aux moteurs de recherche par exemple, cas des signaux faibles, de l'expression de la subjectivité, et de la néologie sémantique).

Enjeu pour le multilinguisme — De récentes recherches exploratoires sur les structures de traits sémantiques amènent à formuler l'hypothèse selon laquelle une unité lexicale est un cas particulier de forme sémantique[6]. Si le mot est privilégié dans notre tradition linguistique, c'est pour des raisons logocentriques mais étudier l'unité lexicale comme une forme sémantique permet d'envisager un continuum depuis le morphème jusqu'au texte et de décrire les mots comme des phénomènes textuels, au même titre que les thèmes, les structures actancielles, etc. En bref, le mot est aussi redevable au texte de son contenu sémantique que l'inverse. L'approfondissement des connaissances sur les structures de traits sémantiques dans une perspective transculturelle et translingue paraît aujourd'hui cruciale. On apportera à cet égard une grande importance à la problématique des corpus comparables, plus intéressante que celle des corpus parallèles parce qu'ils respectent et soulignent les variations culturelles qui s'expriment notamment par les discours et les genres textuels, mais aussi par différentes thématisations à l'oeuvre dans un texte (cf. le concept de *taxème* — petite classe sémantique correspondant à une situation pratique précise). En bref, les corpus comparables donnent à étudier non pas des équivalences traductionnelles mais des équivalences praxéologiques. Ainsi par exemple, le terme corporatisme, sera systématiquement traduit en anglais par « corporatism » dans la documentation officielle multilingue européenne ou canadienne (http://www.linguee.fr) alors qu'une identification sur corpus comparables donne à lire la traduction usuelle — mais figurant rarement dans les dictionnaires — « vested interests ».L'application en lexicographie multilingue de la lexicologie textuelle fondée sur les formes sémantiques, sera le cadre de nouvelles propositions pour la réalisation de dictionnaires dématérialisés, affranchis de l'indexation alphabétique, hypertextuels et sans contrainte de volume. Des concepts typologiques comme le *taxème* pourront notamment avoir un rôle à jouer dans l'élaboration de dictionnaires translingues à vocation didactique.

b. Nouveaux champs d'application

L'exploration de champs d'application voisins où la linguistique est peu présente est une des caractéristiques du projet scientifique de l'ERTIM. Il lui importe en effet que la linguistique adapte ses objets et ses théories aux pratiques émergentes du texte (Internet, dématérialisation) et à la diversité des langues et des usages culturels. Dans la mesure où l'équipe centre son activité scientifique sur l'analyse desdites pratiques en contexte multilingue, son projet repose sur *une prospection continue des nouveaux champs d'applications pertinents*. Après avoir exploré la

[4] Projet européen PRINCIP (2002-2004), projet ANR C-MANTIC pour ce qui concerne l'équipe. Voir aussi les actes du colloque *Documents, textes, œuvres (autour de François Rastier)*, à paraître. Par exemple, dans le cadre du projet PRINCIP, l'équipe a mis en évidence la variété des formes sémantiques associées la forme « étranger » dans les textes racistes et les textes antiracistes. Dans ceux-ci, elle est en cooccurrence avec « irrégularité » et « régularisation », tandis que dans ceux-là, elle cooccurre de façon privilégiée avec « illégalité » et « naturalisation ».

[5] Cette problématique pourrait faire l'objet d'une opération dans le labex EFL (Axe 5, coordonné par A. Nazarenko) dans laquelle l'ERTIM sera impliquée (opération à l'étude à l'heure où ces lignes sont écrites).

[6] Lire Reutenauer, C. (2012). Vers un traitement automatique de la néosémie : approche textuelle et statistique. Thèse de doctorat, Université de Lorraine ; Valette, M. (2010), « Propositions pour une lexicologie textuelle », *Les configurations du sens,* Peter Blumenthal & Salah Mejri, éd., *Zeitschrift für Französische Sprache und Literatur*, 37, Franz Steiner Verlag, éd., pp. 171-188.

terminologie, l'extraction de connaissances, le filtrage de texte, l'extraction d'information subjective, l'équipe prospecte actuellement différents domaines, s'insérant dans des problématiques élaborées en 2011, en s'appuyant sur diverses formules collaboratives. On en dessine ci-dessous les principaux contours.

L'ingénierie des connaissances — Elle constitue un domaine d'investigation privilégié par l'ERTIM depuis sa création. L'enjeu est de concevoir des méthodes 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 ou programmées par l'équipe entrent de plain-pied dans ce cadre de problématiques, en proposant plusieurs travaux d'extraction d'informations linguistiques[7]. Après les ANR VIGITERMES et C-MANTIC, le projet ANR ACCORDYS (20012-2015) constitue une instanciation récente de cette recherche. Il s'agit de mettre au point concrètement ce type de méthodes sur un cas d'usage : la construction d'une base de cas de malformations prénatales et son utilisation pour aider le diagnostic de dysmorphologie foetale et la recherche médicale (identification de gènes, recherche clinique), ce qui représente un enjeu de société crucial (le diagnostic des maladies rares est rendu difficile par la rareté et la dispersion des données médicales). Dans ce cadre, l'ERTIM collabore notamment avec des partenaires TAL et Ingénierie des Connaissances avec lesquels elle a noué une relation déjà ancienne (INSERM, LIMSI).

L'ERTIM approfondira aussi ses recherches menées actuellement en extraction d'information subjective, notamment dans la perspective d'allouer au texte, et plus particulièrement aux corpus, le statut de *systèmes d'organisation des connaissances*[8] dans la mesure où les textes planifient les connaissances : connaissances de haut niveau lorsqu'il s'agit de connaissances expertes (techniques, scientifiques) et connaissances de bas niveau lorsqu'elles n'ont pas été sanctionnées ou éditées par une autorité (web 2.0, forum de discussion). Cette planification est différentielle dans la mesure où les textes explicitent et organisent des connaissances apparentées de manières différentes[9]. A cet égard, l'ERTIM est notamment en contact avec Orphanet (portail des maladies rares et des médicaments orphelins) et Eurordis (fédération d'associations de malades et d'individus actifs dans le domaine des maladies rares) pour un projet d'extraction des signaux faibles (symptomes) dans les forums de discussions de familles de malades (en collaboration avec l'INSERM).

Autres projets collaboratifs (mais non décrits ici)

- traduction et textométrie multilingue (avec le SYLED de Paris 3 et l'EILA de Paris 7),
- DOXAI : veille stratégique (avec l'IRIT, Toulouse)

Les humanités numériques — elles sont une des mutations importantes des sciences humaines et sociales actuellement. La banalisation du support numérique et les grands chantiers de dématérialisation des textes anciens offrent de nouvelles opportunités non seulement en termes d'accès aux données, mais aussi — et surtout — en termes d'analyses renouvelées desdites données. Les premiers pas des humanités numériques relevaient d'ambitions à la fois patrimoniales, éditoriales et documentaires. Beaucoup d'initiatives consistaient en collectes de documents, numérisations et collations pour ensuite 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 et d'étiquetage (morphosyntaxique, lexical) 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 est consacré à la réalisation de méthodologies et d'outils de fouille sémantique de corpus comparables multilingues et multi-écritures. Il s'agit d'analyser les comportements sanitaires de la jeunesse vietnamienne tels qu'ils s'expriment sur les forums de discussion[10]. Cette recherche constitue la première pierre d'une approche originale 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 auront à travailler désormais : les corpus sont constitués à partir de données issues du Web (blog de presse en ligne, forum de discussion) afin de capter au mieux les tendances culturelles et linguistiques. Le projet « humanités numériques » de l'équipe s'appuie sur son savoir-faire et son

[7] Extraction de lexiques spécialisés en chinois (G. Patin, doct.), extraction d'actions spatio-temporalisées (Z. Wang, doct.), extraction de lexiques bilingues français/japonais à partir de corpus parallèles et comparables (P. Marchal, doct.), extraction et décomposition morphématique des unités polylexicales de l'allemand (M.-A. Moreaux, MCF).

[8] Lire Zacklad, M. A. Giboin, éd. (2010) *Applications à base de SOC hétérogènes. Thésaurus, ontologies, folksonomies, Document numérique* 2/2010 (Vol. 13).

[9] Lire Slodzian, M. Valette, M. (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, Actes du 12e Colloque International sur le Document Electronique. Organisé du 21 au 23 octobre 2009 à l'Université de Montréal (CIDE.12), Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

[10] Thèse d'Océane Ho Dinh (2011-2014).

expertise en matière d'analyse des documents du web social acquis notamment dans le cadre du projet ANR C-MANTIC.

Le projet, à terme, doit intéresser les aires culturelles de l'INALCO et donner lieu à des collaborations internes à l'institut. Des contacts ont lieu avec l'équipe CERMOM de l'institut (dans le cadre du projet ALIENTO INALCO-Université de Lorraine).

Didactique des langues et les nouveaux usages nomades — Les nouveaux 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 deux projets SOCRATES LINGUA qu'elle a coordonnés (ATHOS, 1997-1999 et ALPCU, 2003-2007), l'ERTIM étudie la réalisation de méthodes dont la vocation est l'acquisition de compétences partielles en réception de l'écrit, omniprésent dans nos sociétés de l'ère numérique où l'usage des moteurs de recherche est devenu quotidien et l'information accessible est multilingue. Les apprenants ont toutes les facilités de se confronter à une grande variété de textes en langue seconde (L2) *via* Internet, beaucoup plus facilement que d'interagir oralement avec des locuteurs de la L2. Ainsi, 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. En bref, en se concentrant sur la lecture de l'écrit, nous adoptons un principe d'économie cognitive pour l'apprenant[11]. L'objectif circonscrit, 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.). Dans ce cadre, les compétences à acquérir ne relèvent pas de la maîtrise de la langue à proprement parler mais de la compréhension de textes, qui peuvent être spécialisés (techniques, scientifiques) à des fins professionnelles de veille et de recherche. L'ERTIM entend prospecter les points suivants : (i) complémentarité des supports (applications fixes et nomades) ; (ii) auto-construction des corpus par l'apprenant à partir de critères de genres, de thématiques et de lisibilité (*readibility*)[12] ; (iii) combinaisons d'approches universalistes (compréhension conditionnée par des compétences extralinguistiques) et spécifiques aux langues (focalisation sur les propriétés de la langue cible) (iv) acquisition incidente de compétences lexicales ou textuelles (à l'aide d'outils de recontextualisation : concordancier, index des unités déjà lues) ; (iv) simulation du tutorat.

Malgré la prééminence de l'anglais, une part importante de la connaissance est produite dans des textes rédigés en langue nationale, que ce soit dans le domaine industriel (brevets, documentation technique, etc.) ou dans le domaine académique (en particulier dans les sciences humaines et sociales (SHS) où la cohabitation entre les langues nationales et l'anglais perdure). Pour prendre l'exemple des SHS, les linguistes allemands et français partagent et poursuivent une tradition séculaire (philologie, *Textwissenschaft*) qui se distingue singulièrement de la tradition anglo-saxonne (logicienne). Mais cette tradition continentale est peu diffusée en langue anglaise. Ainsi, dans plusieurs domaines scientifiques, l'incompréhension conduit à une acculturation scientifique majoritairement anglo-saxonne. Or, pour beaucoup de chercheurs, et compte tenu de la parenté des langues, notamment savantes, il suffirait d'une pratique de lecture assistée pour désenclaver les connaissances produites en langue nationale. Dans le contexte INALCO, le désenclavement des langues et des connaissances qu'elles véhiculent apparaît comme un défi majeur que l'ERTIM souhaite relever.

Dans cette perspective de recherche en compréhension de l'écrit, plusieurs projets sont en préparation : (i) avec Paris 3 (DILTEC) pour un projet de mesure de la lisibilité, (ii) avec Paris 3 (DILTEC), Paris 7 (EILA) et Paris 6 (Master Ingénierie de la Formation en Ligne) pour l'élaboration d'une preuve de concept (projet soumis fin octobre à l'appel blanc de l'USPC), (iii) avec les universités de Tartu et de Hildesheim pour la réalisation d'une application dans le cadre du programme Lifelong Learning 2013 de la Commission Européenne.

c. Stratégie de mise en œuvre

On synthétise ici, en guise de conclusion, quelques aspects de la stratégie prévue pour atteindre nos objectifs, complémentaires de ceux indiqués pp. 3-4 pour contrer les faiblesses et risques observés dans l'analyse SWOT.

[11] Lire par exemple Hansen, R.D. (1985). Cognitive economy and commonsense attribution procession. In J.H. Harvey & G. Weary (Eds.), Attribution : Basic issues and applications, Orlando, Academic Press, pp. 65-85.

[12] Lire par exemple Miltsakaki, E. and Troutt, A. (2008). Real Time Web Text Classification and Analysis of Reading Difficulty. Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, at the 46th Meeting of the Association for Computational Linguistics and Human Language Technologies, Columbus, OH.

Des partenariats toujours plus nombreux — Outre les partenariats historiques et fortement invétérés (notamment avec l'INSERM et le LIMSI[13]), l'ERTIM entreprend des rapprochements concrets avec plusieurs équipe de l'USPC, dans le cadre LABEX EFL (LIPN, LATTICE) et hors LABEX (SYLED/CLA^{2}T, DILTEC, CLILLAC-ARP/EILA) (lire *supra*).

Par ailleurs, le Master Ingénierie Linguistique (PLURITAL), dont l'assise universitaire est déjà exceptionnelle depuis 2002 (cohabilitation Paris 3, Paris 10, INALCO) est appelé à élargir encore ses points d'appui (mutualisation Paris 7, Master Linguistique Informatique). Un Projet Pédagogique Emergent « Constitution d'un pôle des sciences et pratiques de la traduction », lauréat d'un appel de l'USPC, est en cours de réalisation qui inclut plusieurs équipes de formation de l'INALCO (dont le TIM), de Paris 3 (dont l'ESIT) et Paris 7 (EILA). Il vise à optimiser l'offre en matière de formation en traduction/traductologie à Paris. Enfin, l'équipe de formation du Master Ingénierie Linguistique/PLURITAL travaille actuellement à une meilleure intégration des laboratoires supports à la formation (ERTIM, SYLED, MODYCO, LATTICE).

La montée en puissance du doctorat — Depuis le dernier contrat quadriennal, le nombre de thèses en cours est passé de deux à dix(avec de nombreux refus, faute d'encadrants disponibles). Neuf de ses thèses sont financées (5 CIFRE, 2 CD, 2 ATER ou contrat ANR). L'ERTIM conçoit en effet la recherche doctorale comme une professionnalisation, dans la continuité du Master Ingénierie Linguistique. L'objectif de la formation doctorale est de donner à des étudiants issus des cursus de langues ou de sciences du langage les outils intellectuels et techniques leur permettant de s'orienter *immédiatement* vers les métiers de l'ingénierie linguistique, qu'ils soient académiques ou industriels (tous nos doctorants ont suivi une formation L, M, ou LM à l'INALCO). Mais elle conçoit le doctorat également comme un instrument de veille : l'interaction régulière entre l'équipe et les entreprises partenaires CIFRE permet en effet de prendre connaissance des problématiques émergentes dans l'industrie, notamment dans le secteur extrêmement mouvant de l'Internet et de la société de l'information. Ce dialogue permet aussi de prendre en compte la forte demande sociale en matière de traitement du multilinguisme et des langues stratégiques.

Enfin, les thèses professionnalisantes sont l'occasion d'éprouver et, le cas échéant, de valider les propositions théoriques des sciences du langage et, par conséquent, de diffuser et de valoriser de nouvelles procédures méthodologiques et algorithmiques dans la société civile et l'industrie. A cet égard, les interactions entre doctorants CIFRE et doctorants académiques (ATER, Contrats Doctoraux, ANR), au sein des séminaires organisés à leur intention, peuvent être l'occasion de débats opposant « principe de réalité » (faisabilité industrielle) et « principe de plaisir » (modèles scientifiquement satisfaisants).

Le recrutement (espéré sinon probable) de Frédérique Segond sur le poste de PAST vacant en 2013, ouvre de nouvelles perspectives en termes de collaborations avec l'industrie. L'élaboration de partenariat « conception contre développement » est également envisagée à l'initiative de F. Segond.

Recherche de moyens — Dans le contexte actuel, l'équipe n'envisage pas de modifier sa stratégie de financement basée à plus de 90% sur des fonds contractuels (ANR, CIFRE), comme l'indique le bilan. Elle souligne à nouveau l'impératif que constituent pour elle le recrutement de permanents ou tout du moins le passage en CDI de ses emplois contractuels comme E. Eensoo, qui travaille à l'ERTIM depuis 5 ans.

[13] Une proportion significative de l'équipe enseignante de notre Master est constituée de membres du LIMSI.