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

1.1. Résultats et auto-évaluation de l'unité

1. Rapport scientifique : auto-évaluation

Contexte

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) et le Centre d'Études et de Recherche en Traitement Automatique des Langues (CERTAL). L'année 2007 a été une année charnière pour l'équipe avec : (i) la clôture de projets initiés par le CRIM avant la fusion (TCAN CNRS DECO, SOCRATES-LINGUA 2 ALPCU, PUF e-Lexiques) ; (ii) l'obtention de deux projets ANR dont l'un comme porteur et (iii) la mise en place d'une activité d'encadrement doctoral. La stratégie scientifique de l'équipe s'en est trouvé modifiée et a donné lieu a une réorganisation de la recherche en 2009, suivant 4 axes révisés en assemblée générale en décembre 2011 (cf. *infra*, « b. Principaux résultats par axe »).

a. Activités et résultats

La vocation de l'ERTIM est d'être un acteur majeur de l'ingénierie et du TAL multilingue en France et en Europe. Son projet scientifique s'articule autour des thèmes suivants :

- la recherche en sémantique des textes pour les applications en traitement automatique des langues ;
- le développement de méthodologie pour l'ingénierie des textes et des documents numériques multilingues ;
- la production de ressources multilingues (lexicales, terminologiques, textuelles, didactiques).

Les champs disciplinaires dans lesquels l'équipe évolue sont ceux du traitement automatique des langues, des statistiques textuelles, de la terminologie et de l'ingénierie des connaissances, de la didactique, mais aussi de la linguistique générale (lexicologie textuelle, sémantique textuelle, morphologie lexicale).

En termes de champs d'application, la période a été marquée par les domaines sanitaires et médicaux, qui correspondent à une forte demande sociale. L'ERTIM bénéficie dans ces domaines d'une expertise reconnue à la fois d'un point de vue fouille de données, extraction d'information et analyse multilingue. Une problématique émergente fait également l'objet d'un investissement de l'équipe depuis deux ans : les humanités numériques.

Rayonnement et attractivité académiques

- **Participation aux manifestations scientifiques** du domaine (voir 3. Liste des publications et des productions).
- **Organisation d'événements scientifiques** :
  - depuis 2007 : formation doctorale/séminaire de recherche de l'ERTIM (qui accueille chaque année le séminaire de F. Rastier et des intervenants extérieurs issus du monde académique (D. Cardon, A. Jackiewicz, ou industriel, en conférence ou en atelier (6 à 8 séances par an) ;
  - en 2009 : Workshop *Du thème au terme. Émergence et lexicalisation des connaissances (TIA'09 Workshop, 20 novembre 2009,* IRIT, Toulouse) ;

- en 2011 : *9th International Conference on Terminology and Artificial Intelligence* (TIA'11, 8-10 novembre 2011, Paris). TIA'11 était la 9ème édition d'une conférence initialement francophone mais qui a évolué depuis 2007 en une conférence internationale. Les grands thèmes de la conférence étaient les suivants : acquisition et gestion de terminologie, terminologie et représentation des connaissances, terminologie et applications. La conférence s'est organisée autour de 2 conférences invitées, 14 présentations orales et 11 présentations affichées. Elle a accueilli environ 90 participants venus de 11 pays différents. Outre 2 workshops, les sessions ont porté sur les terminologies multilingues, l'extraction de termes à partir de textes, les liens entre textes et ontologie, la structuration terminologique, les relations entre termes, les terminologies pour un public non spécialiste. TIA 2011 a été organisée par l'ERTIM avec le soutien de l'INALCO, de l'Association française pour l'intelligence artificielle (AFIA), de l'Association pour le traitement automatique des langues (ATALA) et de l'association européenne pour les ressources linguistiques (ELRA).

- en 2011 : Workshop *Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity?* (TIA'11, 10 novembre 2011, Paris).

- en 2014 (en préparation, en partenariat avec l'Université Sorbonne Nouvelle Paris 3 — EA SYLED) : *12th International Conference on the Statistical Analysis of Textual Data*.

- **Coordination ou participation à des projets de recherche partenariale** :

  - *comme coordinateur :*

    - 2007-2010 : ANR MDCO C-MANTIC (Méthodologie et outils pour l'application de la sémantique de corpus au filtrage de masses documentaires) ;

  - *comme partenaire :*

    - 2007-2009 : ANR TECSAN VIGITERMES (Vers une meilleure détection du signal et gestion des connaissances en pharmacovigilance),

    - 2012-2015 : ANR CONTINT ACCORDYS (Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas de DYSmorphologie foetale).

- **Participation à l'élaboration du LABEX EFL (Empirical Foundations of Linguistics)** de Sorbonne Paris Cité, notamment l'axe 5 (Computational semantic analysis). Malgré sa note A à l'évaluation AERES précédente (A+ en qualité scientifique), l'ERTIM n'a pas été retenue comme équipe membre du LABEX car seules les UMR y ont été intégrées. La position du LABEX est toutefois ambiguë car il affiche le Master d'Ingénierie linguistique PLURITAL (notamment adossé à l'ERTIM, voir *infra*) comme une formation du LABEX alors qu'aucun des laboratoires supports n'en fait partie.

Interactions avec l'environnement social, économique et culturel

- **Collaborations avec les acteurs de l'industrie**. Sur la période considérée, l'ERTIM a poursuivi et développé, conformément à sa tradition, de nombreuse collaborations avec l'industrie par le biais de projets partenariaux : MONDECA (2007-2009), TEMIS (2007-2009), ANTIDOT (2012-2015), Asiathèque (2007), IMAGO Services (2007-2010, 2012) Green Management (2011) ou par le co-encadrement de doctorants en CIFRE (ARISEM, AMI Software, GEOL Semantics, SAMESTORY).

  À noter que l'ERTIM a initié en 2007 et assumé financièrement l'adhésion de l'INALCO au pôle de compétitivité CAP DIGITAL. Par ailleurs, deux anciens de la formation (dont 1 CIFRE ERTIM) ont créé une société d'ingénierie linguistique en 2012 : XIKO (agence conseil spécialisée dans l'analyse des conversations en ligne).

- **Adossement à l'enseignement en MASTER** : l'ERTIM est un des laboratoires d'adossement de la Spécialité Ingénierie Linguistique du MASTER Science du langage, dite « MASTER PLURITAL » (http://plurital.org), cohabilitée INALCO/Sorbonne Nouvelle Paris 3/Paris Ouest Nanterre La Défense). Cette formation a pour objectif de donner à des étudiants issus des cursus de langues ou de sciences du langage des bases solides qui leur permettent de s'orienter vers les métiers de l'ingénierie linguistique et choisir entre diverses perspectives : document électronique, ingénierie multilingue, traductique. Le Master est sans doute le plus important à Paris dans le domaine de l'ingénierie linguistique et du TAL (entre 70 et 80 inscriptions par an sur l'ensemble des 3 établissements ; plus de 70% des étudiants diplômés trouvent un emploi dans le domaine à l'issue du M2).

  Les compétences acquises sont à la fois linguistiques (méthodes et outils du TAL multilingue, unités de sens et multilinguisme, sémantique textuelle, traitement automatique de l'oral), statistiques (traitements quantitatifs sur corpus multilingues, textométrie, statistique textuelle) et informatiques (structuration de l'information, transformations, formats, codages, mémoires de traduction). Elles permettent aux étudiants de travailler dans le domaine du traitement de corpus multilingue et 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...).

Implication dans la formation par la recherche

- **Activité d'encadrement doctoral en plein essor**. 10 doctorants se sont inscrits depuis 2007, dont 9 sont financés. Convaincue que la recherche est professionnalisante, et forte de son expérience en Master Pro, l'ERTIM met tout en œuvre pour que l'activité doctorale soit une activité professionnelle comme les autres. Tous nos doctorants sont salariés, soit par le biais de convention CIFRE (5 thèses), soit par le biais de CDD académiques (2 ATER, 2 Contrats Doctoraux).

Principaux résultats par axe

Les travaux de l'équipe ont été répartis ci-dessous suivant les axes de recherches modifiés par l'AG du 7 décembre 2011[1]. Les résultats scientifiques ne sont pas tous détaillés, nous nous focalisons sur les projets phares de chaque axe. Les recherches non détaillées, en particulier les travaux doctoraux, poursuivent et approfondissent en général les recherches menées auparavant ou en parallèle dans les projets contractuels.

Axe 1 : Sémantique de corpus pour les applications TAL (resp. Mathieu Valette)

Cet axe vise à approfondir les propositions théoriques de la sémantique textuelle, en l'appliquant à l'ingénierie multilingue. Il s'agit notamment d'élaborer des méthodologies de traitement de corpus, de modéliser et de participer au développement d'outils de fouille de textes, d'analyse et d'interprétation de textes assistées. Les applications visées sont celles de la recherche d'information, la classification de documents et la fouille de textes.

Résultats scientifiques

À la suite du projet européen de filtrage de textes racistes à base de règles sémantiques PRINCIP[2] (2002-2004) [24], l'ERTIM a entamé une réflexion approfondie sur les possibilités d'exploiter la sémantique textuelle de F. Rastier[3] dans des applications ingénieriques de type Recherche d'Information, classification et fouille de textes. Cette réflexion s'est concrétisée par plusieurs publications (par exemple [25][57][58][77][123]) et **le projet ANR MDCO C-MANTIC (2007-2010)** qui a permis d'appliquer des méthodes issues de la sémantique des textes à un domaine d'application traditionnellement dominé par les techniques informatiques (notamment l'apprentissage artificiel). Si la linguistique est sollicitée indirectement, via le TAL, pour des prétraitements tels que la normalisation des textes, leur lemmatisation ou leur analyse syntaxique, le savoir-faire des linguistes en analyse sémantique des textes apparaît peu exploité dans un contexte où le sens, pourtant, est au cœur des applications.

L'ERTIM s'est donné comme objectif **d'éliciter, de modéliser et d'outiller le savoir-faire sémanticien**. Nous adoptons un positionnement proche de partenaires tels le SYLED-CLA^{2}T avec lequel nous collaborons ou l'équipe de l'ANR Textométrie (dont le SYLED-CLA^{2}T était partenaire), mais dans une perspective ingénierique et applicative et non seulement descriptive et académique[4]. Il en résulte les spécifications et la réalisation partielle d'outils destinés à (i) la constitution de corpus à partir de masses documentaires hétérogènes, (ii) l'analyse sémantique des corpus, (ii) la production de critères complexes pour le filtrage et la classification de textes. Pour lever ces verrous, une plateforme a été spécifiée et partiellement implémentée d'une part pour faciliter la production à la volée de multiples corpus contrastif, d'autre part, pour élaborer des critères de caractérisation multi-niveaux. L'architecture logicielle comprend actuellement un serveur d'application et de gestion des documents (Linguistic Processing Unit) qui pilote les enchaînements de traitements linguistiques multilingues commandées par le linguiste auquel sont interfacés des bibliothèques de traitements propres ou sous architecture UIMA. Une interface graphique qui bénéficiera des recherches menées sur les processus métiers pendant le projet C-Mantic (interface homme/machine, visualisation et manipulation) est à l'étude[5].

L'apport majeur de cet axe est de contribuer à doter une méthodologie linguistique, jusque là confinée au milieu académique, des prérequis techniques pour un outillage permettant de la placer en situation industrielle. Dans le cadre de l'ANR C-MANTIC, la méthodologie a été validée au moyen d'une application sanitaire et sociale : la caractérisation des textes du web relatifs à la consommation de tabac (anglais, chinois, français) [4][-54]. Le projet a, en outre, ouvert un chantier pour la fouille d'opinion se démarquant des applications standard (grilles interprétatives complexes et détection des signaux faibles), actualisé dans le cadre de plusieurs thèses en cours (par exemple [55][61][71][72]).

Réalisations correspondantes

- projet ANR MDCO C-MANTIC (2008-2010) (INALCO, LINA, LIMSI, INSERM) (langues traitées : français, anglais, chinois)

[1] Le dossier soumis initialement à l'AERES en 2005 rendait compte de la structure suivante : Problématique 1 : Enjeu du multilinguisme 1.1 Codage, transcodage des écritures 1.2 Grammaires plurielles et options TAL 1.3 Linguistique de corpus et multilinguisme 1.4 Représentations « objets » pour le TAL — Problématique 2 : Le document numérique 2.1 Formes et formats 2.2 Genres et discours.

[2] Projet PRINCIP (Plateforme pour la Recherche, l'Identification et la Neutralisation des Contenus Illicites et Préjudiciables sur Internet) du programme européen SAFER INTERNET 2002-2004, coordonné par Paris 6, avec l'INALCO, Dublin City University, Magdeburg Universität.

[3] F. Rastier, *La mesure et le grain. Sémantique de corpus*, Honoré Champion, Paris, 2011.

[4] Serge Fleury, responsable du Centre de Lexicométrie et d'Analyse Automatique des Textes (SYLED-CLA^{2}T) de Paris 3, est membre associé de l'ERTIM. Le SYLED a été partenaire du projet ANR Textométrie qui a donné lieu au logiciel TXM.

[5] Ces perspectives sont détaillées dans le document *Projet*.

- thèse Egle Eensoo « Caractérisation sémantique de textes pour la recherche d'information multilingue – recherche d'une méthodologie » (2008-2013) (langues traitées : français, estonien), dir. M. Valette

- thèse CIFRE (AMI Software) Aurélien Lauf « Evolution du « buzz » sur internet : identification, analyse, modélisation et représentation dans un contexte multilingue » (2010-2013) (langues traitées : français, japonais), dir. M. Valette

- thèse CIFRE (Samestory) Jugurtha Aït Hamlat « Analyse et structuration des récits d'expérience issus du web en français, anglais et arabe, en vue de leur classification thématique » (2010-2013) (langues traitées : français, anglais, arabe), dir. M. Valette

- thèse Nadine Pavot « Etude de lexicologie textuelle en japonais : analyse du discours des réseaux sociaux sur Internet » (2009-2013) (langue traitée : japonais), dir. M. Slodzian

- thèse CD Océane Ho Dinh « Méthodes et outils pour le traitement automatique du vietnamien. Application en Humanités Numériques : fouille comportementale sur le web social » (2011-2014) (langue traitée : vietnamien), dir. M. Valette

Autres

- expertise-conseil sur le projet collaboratif ALIENTO (Analyse Linguistique & Interculturelle des ENoncés sapientiels et Transmission Orient/occident) (MSH Lorraine / INALCO (CERMOM) / Université de Lorraine).

- soumission AXA Foundation 2010 : ARP (Accessing risk perception) (INALCO) — non financé

- projet Doxai (IRIT (Toulouse), INALCO) — projet en préparation

- co-organisation de la 12*th International Conference on the Statistical Analysis of Textual Data*, en partenariat avec l'Université Sorbonne Nouvelle Paris 3 — EA SYLED — en préparation.

Axe 2 : Acquisition de connaissances (resp. Pierre Zweigenbaum)

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.

Résultats scientifiques

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 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).

Des méthodes pour collecter des ressources terminologiques et les appliquer à l'extraction d'informations ciblées ont été mises au point. Ces méthodes ont été testées sur des articles scientifiques en japonais dans le domaine des médicaments (pharmacovigilance). L'ERTIM a entrepris une étude sur les sources d'information pertinentes et les ressources terminologiques bilingues disponibles au Japon en matière de pharmacovigilance. L'équipe a lancé une étude linguistique sur les articles médicaux issus de la base PUBMED et écrits en japonais, pour formaliser l'analyse et l'interrogation de ces textes. Cette étude a permis de cerner les difficultés présentées par la langue japonaise dans le cadre de l'analyse automatique : encodages, segmentation, repérage des termes. Puis, nous avons réuni un certain nombre de données lexicales et terminologiques permettant de repérer dans les textes en japonais les entités intéressantes, en particulier les données bilingues (français/japonais) de la base terminologique UMLS. Nous avons alors construit une première application qui permet le repérage automatique dans les textes des termes du Metathesaurus de l'UMLS, avec leur traduction en français pour permettre à des médecins pharmacovigilants français de connaître rapidement la présence ou l'absence d'informations médicales susceptibles de les intéresser dans les textes japonais. Enfin, des règles linguistiques spécifiques permettant de trouver dans les textes des informations sur des éléments non stables (comme la durée du traitement ou les antécédents du patient) ont été implémentées dans un prototype, en utilisant les outils de traitement automatique du japonais disponibles (Chasen et Cabocha) [26].

Ces travaux ont été menés dans le contexte du **projet ANR TECSAN VIGITERMES (2007-2009) qui** vise à améliorer l'environnement de travail des équipes de pharmacovigilance. Les outils et ressources développés sont intégrés sur une plate-forme de gestion des connaissances couplée à des outils de fouille de texte pour améliorer l'accès, l'analyse et la documentation des cas de pharmacovigilance. L'ERTIM était responsable du module permettant l'interrogation en ligne d'articles scientifiques en japonais pour y repérer les mentions d'effets indésirables suite à la prise d'un traitement médical, et a fourni entre autres un livrable sur le « text mining en japonais » en septembre 2008 [26]. Ces travaux se prolongent dans le **projet ANR CONTINT ACCORDYS (2012-2015)** qui part du constat que de nombreuses organisations disposent de gisements de connaissances inexploités, souvent peu structurés et hétérogènes. L'accès à la connaissance ainsi engrangée, mémoire des organismes, aiderait souvent à mieux traiter les cas présents. Leur sous-exploitation est due à la difficulté à d'y accéder de façon pertinente. Le projet ACCORDYS s'attaque à ce verrou scientifique par une approche combinée d'ingénierie des connaissances, de traitement de la langue (où l'ERTIM interviendra plus particulièrement) et de recherche d'information. ACCORDYS expérimentera ces méthodes dans un

domaine où l'accès aux cas passés est particulièrement utile et où son impact sociétal est important : le domaine de la santé, plus particulièrement celui des maladies rares. L'application choisie pour valider les hypothèses de recherche consiste en l'exploitation des données biomédicales hétérogènes (et des comptes rendus textuels correspondants) accumulées au cours des années par des spécialistes. Dans le contexte du diagnostic prénatal d'une malformation, la connaissance de situations antérieures « similaires » et résolues (c'est-à-dire vérifiées après l'autopsie du fœtus) est essentielle à l'orientation diagnostique. Notre hypothèse de recherche est la suivante : analyser les dossiers leur apportera une valeur ajoutée en termes de connaissances, permettant ainsi de construire une base de connaissances du domaine. Les dossiers seront structurés et enrichis par indexation à l'aide de ressources termino-ontologiques et par leur mise en relation avec des connaissances externes (publications, bases de données). Une telle base permettra l'accès à des cas similaires. Dans ACCORDYS les textes analysés seront d'une part des comptes rendus cliniques en français et d'autre part des publications scientifiques, principalement en anglais.

Réalisations correspondantes

- projet ANR TECSAN VIGITERMES (2007-2009) (INSERM, IM&Bio, INALCO, Rennes 1, LORIA, DSPIM, Mondeca, Temis, etc.) (langues traitées : français, anglais, japonais)
- projet ANR CONTINT ACCORDYS (2012-2015) (INSERM, INALCO, LIMSI, Antidot) (langues traitées : français, anglais)
- thèse CD Pierre Marchal « Extraction de lexiques bilingues français/japonais à partir de corpus parallèles et comparables » (2010-2013) (en codirection INALCO – Waseda University) (langues traitées : japonais), dir. Th. Poibeau
- thèse CIFRE (Arisem), Gaël Patin « Extraction interactive et non supervisée de lexique en chinois contemporain appliquée à la constitution de ressources linguistiques dans un domaine spécialisé » (2007-2012) (langue traitée : chinois), dir. P. Zweigenbaum
- thèse CIFRE (Arisem), Mani Ezzat « Passage de données non structurées à des données structurées : acquisition de relations entre entités nommées à partir de corpus » (2008-2013) (langue traitée : français), dir. Th. Poibeau
- thèse CIFRE (GEOL Semantics), Zhen Wang « Extraction en langue chinoise d'actions spatio-temporalisées réalisées par des personnes ou des organismes » (2010-2013) (langue traitée : chinois), dir. P. Zweigenbaum
- organisation de la *9th International Conference on Terminology and Artificial Intelligence* (TIA'11, 8-10 novembre 2011, Paris)

Autres

- soumission ANR-JST 2010 HYPERICUM/オトギリソウ属 (INALCO (FR), INSERM (FR), LINA (FR), LIMSI (FR), Tokyo University (JP)) — non financé
- soumission 7ème PCRD 2011 ICT DOCPROCHAIN (IMAGO Services (FR), INALCO (FR), ABBY (RU), Fundazione Bruno Kessler (IT), Aproged (FR), VOI (DE), EvidenceCube (BE)) — non financé

Axe 3 : Technologies éducatives et apprentissage des langues (resp. François Stuck)

Cet axe vise la conception et le développement finalisé de méthodes et d'outils d'apprentissage des langues fondés sur la création de ressources intégrant des techniques de corpus et de TAL.

Résultats scientifiques

La conception d'outils pour la didactique des langues est une tradition ancienne dans l'équipe qui a été porteuse de deux projets européens SOCRATES LINGUA depuis 1997. La période 2007-2012 a été marquée par le déploiement au sein de l'INALCO de la plateforme e-learning ALPCU issue du dernier projet européen (2005-2007) [105][106][113] inspirées de travaux antérieurs sur l'intercompréhension.

Ces travaux ont débouché sur un modèle de cours interactifs de langues étrangères proposant diverses activités tant de découverte que d'appropriation, basés sur la modularité des compétences langagières. Ce modèle met en œuvre par le biais de l'annotation de documents didactiques : (i) la multicanalité qui, en diversifiant et croisant les canaux d'apprentissage (textes, sons, vidéos, graphismes) contribue à la dissociation des compétences, mais aussi à leur rapprochement ponctuel et différencié (par ex. lecture pure d'un texte vs synchronisée avec son oralisation ; (ii) la multiréférencialité qui en proposant une multiplicité de points de vue (typiquement grammatical, lexical, communicatif ou culturel, mais aussi toute composante jugée pertinente) sur un objet didactique, renforce la différenciation des parcours d'apprentissage selon le temps et les profils des apprenants et leurs motivations ; (iii) une hypertextualité différenciée, distinguant les hypermots prototypiques d'un phénomène langagier et les diverses occurrences de ce dernier.

Aujourd'hui, ce modèle évolue vers la didactisation automatique de textes à l'aide d'outils de TAL (étiqueteurs morphosyntaxiques, ressources lexicales, etc.) dans le cadre d'outils d'aide à la lecture en Langue seconde (projet DEJA LU) et vers la didactisation de textes parallèles, notamment des variantes dialectales d'un même document (projet RED-Rrom).

Réalisations correspondantes

- adaptation de la **plateforme e-learning ALCPU**[6] en système de création et de gestion de contenus pédagogiques pour l'apprentissage des langues (CMS) (2010-2013) :

  refonte complète du moteur de rendu de textes didactisés issu d'ALPCU pour le rendre muti-écriture et multilingue, multi-navigateur et multi-système : utilisation des technologies XML, XSLT, JavaScript et CSS. (cf. http://maquettealpcu.crim.fr/nouveauxTD)

  Étude de faisabilité de son intégration dans un ENT (Moodle)

  - expertise technique et pré-montage pour la mise en place d'un projet d'enseignement du swahili pour la CMS ALPCU (2011)

  - Formation à la création de cours interactif et à l'usage de la plateforme ALPCU (début 2013)

- soutien technique et méthodologique au **projet européen RED-RROM** « Restoring the European Dimension of the Romani Language and Culture » (Projet du programme Education & Culture (2011-2015) du Centre d'étude Rromani de l'INALCO) : développement d'**outils** de didactisation semi-automatique (annotation lexicale et synchronisation sons/textes), adaptation du moteur de rendu à des textes parallèles en divers dialectes du rromani et expertise technique pour la future intégration d'un cours de rromani.

- thèse Nadia Makouar « Ressources textuelles pour la langue arabe et méthodologie de constitution de corpus avec la sémantique textuelle : application au e-learning » (2009-2012) (en codirection INALCO – Université Hassan II Mohammedia) (langue traitée : arabe littéral), dir. M. Valette

Autres

- 1ère soumission au programme LIFELONG LEARNING 2012 du projet européen DEJA LU (Dispositif multi-supports d'aide à la lecture et à la compréhension de textes en langue seconde) (INALCO, Hildesheim Universität, Tartu Universitet, Université Sorbonne Nouvelle Paris 3) (langues traitées : allemand, estonien, français, russe) — non financé, 2de soumission en préparation.

Axe transversal : corpus et multilinguisme (resp. Jean-Michel Daube)

Cet axe constitue le cœur de l'équipe et l'un de ses principales originalités dans le paysage scientifique français : l'ERTIM 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 (INALCO, Paris 3, Paris 10, Paris 7) et notamment apprenants ou locuteurs de langue orientale (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).

Pendant l'exercice 2007-2012 considéré, les langues sur lesquelles l'équipe a travaillé sont : allemand, anglais, arabe, estonien, français, japonais, mandarin, persan, swahili, rromani, russe, vietnamien.

Réalisations correspondantes

Les réalisations de l'axe transversal correspondent à la plupart des recherches menées dans les axes 1, 2 et 3. Quelques travaux isolés méritent d'être distingués dans la mesure où ils sont susceptibles de donner lieu à de nouveaux projets en cours d'élaboration :

- inventaire des outils pour les langues asiatiques :

  - état de l'art du TAL japonais (2010-2011)

  - état de l'art du TAL vietnamien (2012)

Autres

- soutien à d'autres équipes de l'INALCO (par exemple : adaptation de l'étiqueteur HUNPOS pour le persan, CERLOM)

- soumission ANR blanc 2009 AMITAL « Approches du Multilinguisme pour le Traitement Automatique des Langues » (SYLLABS, LINA, INALCO) — non financé

[6] Projet ALPCU (Apprendre les Langues nationales des Pays d'Europe centrale et orientale candidats à l'entrée dans l'Union européenne) du Programme européen SOCRATES-LINGUA 2 2005-2007, porté par l'INALCO (ERTIM).

b. Analyse des moyens de l'unité

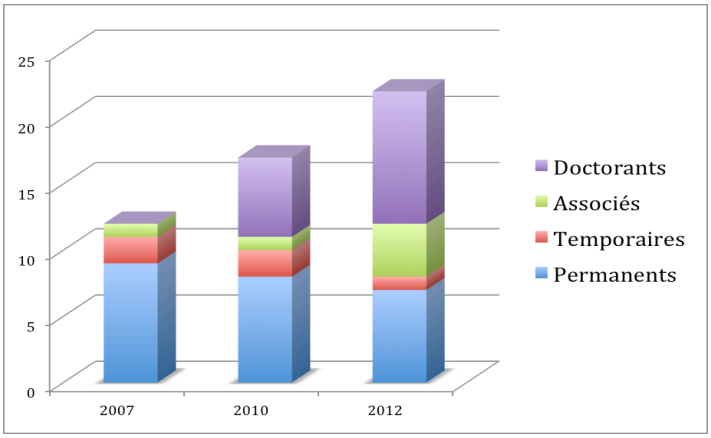
Ressources humaines

L'équipe a connu une importante baisse des effectifs de permanents avec le départ en retraite de François Rastier, DR détaché du CNRS (2009) Monique Slodzian, PU directrice de l'équipe (2010), Evelyne Bourion IE détachée du CNRS (2012). Seul le poste de Monique Slodzian a été renouvelé (recrutement en 2010). Cette baisse a été compensée numériquement par le recrutement régulier de doctorants, à telle enseigne qu'ils constituent actuellement la moitié des effectifs de l'équipe. La situation est toutefois fortement préjudiciable : plusieurs des permanents de l'équipe sont très investis dans des activités académiques diverses, ce qui limite leur capacité de publications (d'autant plus que les deux EC retraités de rang A étaient produisants). Cet assèchement des ressources permanentes contraste avec la forte attractivité de l'équipe, comme l'atteste notamment la croissance continue des inscriptions en doctorat.

Ce déséquilibre croissant entre personnel permanent et personnel non permanent accentue une des principales faiblesses de l'équipe : **la difficulté à capitaliser des connaissances et des compétences** qui sont souvent le fait des ressources humaines contractuelles et doctorales, par nature non pérennes (lire l'analyse SWOT dans le document Projet).

		1er janvier 2007	31 août 2010	30 juin 2012
Permanents	PU/DR	2	1	1
	MCF/CR	3	3	3
	IR	1	1	1
	IE	1	1	0
	PRAG	1	1	1
	ADA	½	½	1
Temporaires	PAST	½	½	½
	IR/IE	1	1	0
Associés	EC/CR/émérite/autre	1	1	4
Doctorants	CIFRE	0	3	5
	ATER/AER	0	½	2
	Contrat Doctoral	0	0	2
	Autres	0	2	1
Total (individus)		**12**	**17**	**22**

Evolution des personnels par corps et type contractuel



Evolution des personnels scientifiques par type contractuel

Les enseignants-chercheurs permanents

Ils sont au nombre de 4 depuis le départ en 2009 de F. Rastier (en détachement, non remplacé).

Les PU — M. Slodzian (2007-2010 puis émérite) et M. Valette (2010-) — sont produisants (publiants). Ils ont assuré successivement la direction de l'équipe. M. Valette est membre élu du CNU (7ème section) ; membre élu du Conseil Scientifique de l'INALCO. Il sera co-responsable (avec Chr. Bonnot) de la Mention Science du Langage et Langues Appliquées (SDLLA) de l'INALCO durant le quinquennal 2014-2018 et responsable de la spécialité Ingénierie Linguistique du Master SDLLA (dite PLURITAL, cohabilitée INALCO-Paris Sorbonne nouvelle - Paris Ouest Nanterre La

Défense). M. Valette est également membre à titre individuel du Labex Empirical Foundations of Linguistics (EFL, Sorbonne Paris Cité). F. Rastier (DR CNRS en détachement, 2007-2009) a une production très importante et d'envergure internationale.

M.-A. Moreaux (MCF) n'a pas été produisante sur la période considérée mais consacre l'intégralité de son temps recherche pour l'institut : elle est directrice du Centre des Systèmes d'Information et des Ressources Numériques (CIRN), correspondante C2i pour l'établissement et responsable de l'initiation à l'informatique en licence. Elle occupe aussi la fonction de directrice du Département TIM. M.-A. Moreaux sera en CRCT au second semestre 2012-2013. H. Rigot (MCF) est produisante ; elle présente également une importante activité de valorisation des productions scientifiques numériques auprès de l'institut (notamment la mise en ligne sur revue.org des revues éditées par l'établissement). M. Fanton (MCF) n'est pas produisant ; il sera à la retraite avant le contrat 2014-2018.

Autres personnels permanents

Jean-Michel Daube (PRAG), directeur adjoint du TIM, est produisant. Outre la publication de ressources terminologiques et lexicales [107][108][109], il est l'auteur de 2 articles sur l'exercice 2007-2012 [1][26]. Il est par ailleurs très investi dans l'équipe pédagogique qu'il anime, les relations avec les partenaires industriels et la gestion des projets contractuels. Il coordonne l'axe transversal.

François Stuck (IR) réalise ou participe à la réalisation de produits logiciels (plateforme e-learning) et de contenus pédagogiques pour l'apprentissage des langues [105][106][113]. Ses productions sont encore peu valorisées académiquement (en termes de publications) mais reconnues au niveau institutionnel (label européen pour ALPCU, valorisation de la plateforme par l'INALCO).

E. Bourion (IE CNRS en détachement, en retraite depuis mars 2012), titulaire d'un doctorat en linguistique a été très impliquée sur le projet ANR C-MANTIC et dans l'équipe pédagogique du TIM.

Personnels non permanents

P. Zweigenbaum (PAST à mi-temps, DR CNRS au LIMSI) présente une activité de publication importante. Ses recherches à l'ERTIM s'effectuent souvent dans le cadre de projets partenariaux (ANR C-MANTIC, ANR ACCORDYS).

R. Belmouhoub (IR 2007-2011, sur contrat ANR), a effectué une activité de soutien informatique stratégique dans une équipe de SHS à vocation technologique. Il a été recruté sur concours en 2011 par l'INALCO, initialement pour une affectation à mi-temps à l'ERTIM mais elle n'est depuis 1 an pas assumée car la charge de travail lié à son autre mi-temps est au Centre des Systèmes d'Information et des Ressources Numériques a été très sous-estimée. Son absence dans l'équipe s'avère pénalisante.

E. Eensoo (½ IE ANR 2007-2009, ½ ATER 2009-2011, ½ AER 2011-2012, IE ANR 2012-) actuellement doctorante, occupe de fait une fonction d'enseignante-chercheuse dans l'équipe. Elle est en particulier fortement impliquée dans le montage des projets collaboratifs, l'animation pédagogique et l'encadrement des étudiants.

P. Marchal (2010-2012) et O. Ho-Dinh (2011-2012) sont en contrats doctoraux. A ce titre, P. Marchal effectue des tâches de collecte et gestion de corpus multilingues, O. Ho-Dinh participe à la gestion du site web de l'équipe et supervise l'inventaire des outils pour les langues asiatiques (axe transversal).

Une dizaine de vacataires ou stagiaires ont également été recrutés ponctuellement (moins de 6 mois), principalement sur les contrats ANR.

L'ERTIM dispose d'un adjoint administratif : ½ ADA 2007-2011 (M. Kraskovetz) puis 1 ADA 2011-2012 (A. Bounoua).

Associés

M. Valette, CR CNRS jusqu'en 2010, était, par convention entre son employeur l'ATILF et l'INALCO, autorisé à effectuer une partie de ses recherches à l'ERTIM, soit 2 jours par semaine (2007-2010). Thierry Poibeau (CR puis DR CNRS) est directeur de deux thèses à l'ERTIM. Serge Fleury (MCF, Paris 3) est associé à l'équipe depuis 2011. Frédérique Segond (VISEO-Objet Direct) est associée depuis 2012 et candidate sur le poste de PAST que P. Zweigenbaum laissera vacant en septembre 2013.

Locaux et équipements

Tendue en 2007-2008 (année passée dans une cave du 2 rue de Lille), la situation s'est grandement améliorée avec le déménagement sur le site de Nogent-sur-Marne en 2008 puis le retour au 2 rue de Lille en 2012. L'équipe jouit depuis septembre 2008 de conditions de travail exceptionnelles dans le contexte parisien (environ 100m^2 alloués à l'équipe).

Crédits

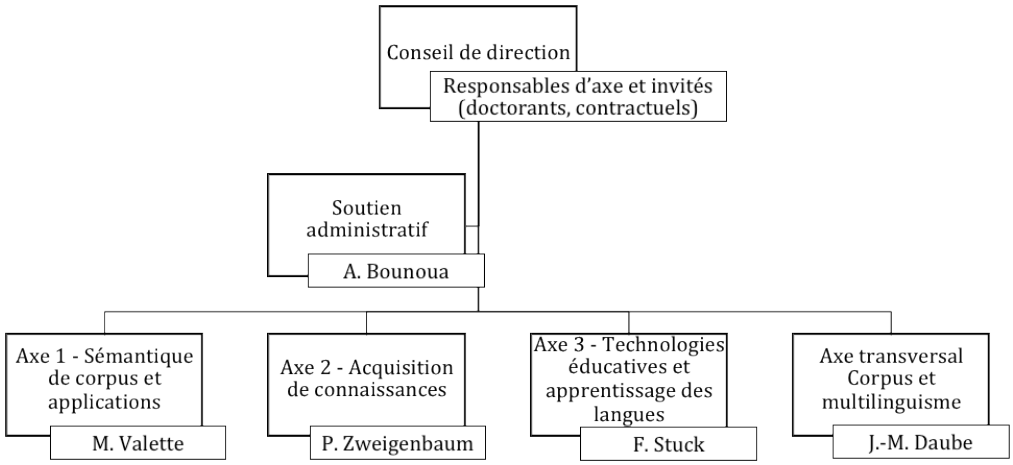
	Dotation INALCO	ANR*	CIFRE
2007	1 700	322 613	6 000
2008	1 700		13 000
2009	1 700		17 000
2010	10 000		16 000
2011	10 000		13 000
2012	10 000	146 759	9 000
Total	**35 100**	**469 372**	**74 000**

Vue d'ensemble sur les revenus de l'équipe
(hors salaires des permanents et contrats doctoraux)

** les montants annoncés sont ceux octroyés par l'ANR à la signature des contrats. La somme indiquée pour 2012 sera engagée le 1er septembre 2012.*

2. Organigramme fonctionnel et règlement intérieur

L'équipe a fonctionné sur la période considérée sans règlement intérieur. Elle est animée par un conseil de direction dont les membres sont les responsables d'axe et des invités réguliers (directrice du département TIM et 3 doctorants). Une AG était organisée une fois par an, à l'automne, ou suivant une décision du conseil de direction. Le conseil de direction se réunissait à la demande d'un de ses membres.



Organigramme de l'ERTIM

3. Liste des publications et des productions

NB : En ce concerne les personnels non permanents, seuls les travaux réalisés dans *le cadre explicite de l'ERTIM* ont été indiqués. Cela concerne notamment François Rastier (sa production, encore très abondante après son départ en retraite en 2009 n'a pas été prise en compte car n'ayant pas fait de demande d'éméritat, il n'est plus administrativement membre de l'équipe) ; Mathieu Valette (les publications 2007-2010, avant son recrutement PU, rendent compte de travaux effectués à l'ERTIM et non à l'ATILF) ; Pierre Zweigenbaum (2007-2010, PAST à mi-temps, ses recherches à l'ERTIM s'effectuent principalement dans le cadre de projets partenariaux ANR).

Articles dans des ou dans les bases de données internationales (ISI Web of Knowledge, Pub Med...).

Renseigné sans distinction dans la rubrique ci-après.

Articles dans des revues avec comité de lecture répertoriées par l'AERES ou dans des bases de données internationales.

1. **Daube, J-M.** (2008) *De la lexicologie textuelle multilingue outillée à la lexicographie numérique*, Syntaxe et Sémantique, Presses Universitaires de Caen, pp. 125-141.
2. Deléger, L., Merkel, M. & **Zweigenbaum, P.** (2009) Translating medical terminologies through word alignment in parallel text corpora. Journal of Biomedical Informatics, 42(4):692-701. Epub 2009 Mars 9.
3. Duteil,C., **Valette, M.**(2008) « Appropriation et réécriture : l'exploitation des faits divers par les sites Web racistes », *Récits et dispositifs du fait divers*, *Médias & Culture - La revue européenne des pratiques médiatiques et culturelles*, n° spécial – novembre 2008, L'Harmattan, Paris, 103-119.
4. **Eensoo-Ramdani, E., Bourion, E., Slodzian, M., Valette, M.** (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Les Cahiers du numérique*, volume 7, n°2, pp. 15-39.
5. **Rastier, F.** (2007) Communication, interprétation, transmission, *Semen*, 23, pp. 121-138.
6. **Rastier, F.** (2007) Signo y negatividad : una revolución saussureana, Puebla, *Topicos del Seminario*, 18, pp. 13-55.
7. **Rastier, F.** (2007) Passages, *Corpus*, 6, Université de Nice. Numéro dirigé par Bénédicte Pincemin, pp. 125-152.
8. **Rastier, F.** (2007) Le langage a-t-il une origine ? *Revue française de psychanalyse*, LXXI, 5, pp. 1481-1496.
9. **Rastier, F.** (2008) Doxa et sémantique de corpus, in G.-E. Sarfati, éd. Discours et sens commun, *Langages*, 170, pp. 54-68.
10. **Rastier, F.** (2008) Obscure référence, in P. Frath, éd., *Zeitschrift für Französische Sprache und Literatur*, 35, pp. 91-108.
11. **Rastier, F.** (2008) La linguistique comme science des textes, *Journal of Hermeneutic Study and Education of Textual Configuration*, I, 1, pp. 1-10.
12. **Rastier, F.** (2008) Sémantique du Web vs Web sémantique, *Syntaxe et sémantique*, 9, pp. 15-36. Numéro spécial Textes, documents numériques, corpus, coordonné par Valette, M.
13. **Rastier, F.** (2009) De Primo Levi au chic nazi — Entretien avec Georges Élia Sarfati, *Controverses*, 10, pp. 134-165.
14. **Rastier, F.** (2009) Tem a linguagem uma origem ?, *Revista Brasileira de Psicana'lise*, 43, 1, pp. 105-118. Traduit par Daisy Guttmann et Regina Campo Salgado.
15. **Rastier, F.** (2009) Euménides et pompiérisme — Refus d'interpréter, *Témoigner*, 103, pp. 171-190.
16. **Rastier, F.** (2009) Passages and Paths within the Intertext, *Belgian Journal of Linguistics*, 23, pp. 7–29.
17. **Rastier, F.** (2009) Heidegger aujourd'hui, ou le Mouvement réaffirmé, *Labyrinthe*, 33, 2009-2, pp. 71-106.
18. **Rastier, F.** (2009) Sémiotique et sciences de la culture, *Acta semiotica et linguistica*, 33, 1, pp. 35-64.
19. **Rastier F,** (2009) Sêmantica dos textos e semiótica, *Acta semiotica e linguistica*, vol. 14, 2, pp. 27-49. Réédition de Semântica dos textos e semiótica, in Cortina, A. et Marchezan, R. C. (éds). *Razões e sensibilidades: a semiótica em foco,* Araraquara: Laboratório Editorial FCL/UNESP, 2004, pp. 11-32.
20. **Rastier, F., Valette, M.** (2009) « De la polysémie à la néosémie », *Le français moderne*, S. Mejri, éd., *La problématique du mot*, 77, 97-116.
21. **Rigot, H.** (2007) « (En)-quête de l'autre. Recherches qualitatives et corpus numériques ». Revue Recherches qualitatives, Hors série no°3, 2007. http://www.recherche-qualitative.qc.ca/hors_serie_v3/RigotFINAL2.pdf
22. **Rigot, H.** (2009) « Le nouveau contrat textuel des humanités et des sciences de la société », Distance et savoirs, Vol. 7, 2009/3, pp. 457-478.
23. **Rigot, H.** (2010) « D'une esthétique de la réception à une pragmatique de la décision » dans *L'Acte éditorial. Publier à la Renaissance et aujourd'hui*. Éd. Classiques Garnier, 2010, pp. 205-223.
24. **Valette, M.** (2009) « Détection automatique des documents racistes et xénophobes sur Internet. L'apport de la sémantique de corpus », *Sciences du langage et demandes sociales, actes du colloque 2007 de l'ASL*, textes réunis et présentés par Christian Hudelot et Christine Jacquet-Pfau, Lambert-Lucas, Paris. ISBN : 978-2-35935-018-0
25. **Valette, M., Slodzian, M.** (2008) « Sémantique des textes et Recherche d'information », *Extraction d'information : l'apport de la linguistique*, A. Condamines & Th. Poibeau, éd., *Revue Française de Linguistique Appliquée*, volume XIII-1 – juin 2008), 119-133.

Articles dans des revues sans comité de lecture.

26. Bousquet, C., Amardheil , F., **Daube , J-M.**, Delamarre , D., Duclos, C., Lanne, S-G., Jaulent, M-C., Lillo-Le Louët , A., Toussaint, Y. (2011) *Vers une meilleure détection du signal et gestion des connaissances en pharmacovigilance : le projet VigiTermes*, IRBM 32 (2011) 158–161

27. **Rastier, F.** (2009) Témoigner et traduire – Sur *Ulysse à Auschwitz*, entretien avec Gaëtan Pégny, *La mer gelée*, 6, pp. 74-85.

28. **Rastier, F.** (2009) Bezeugen und übersetzen – Über *Odisseus in Auschwitz*, Gespräsch zwischen Gaëtan Pégny und François Rastier, Übersetzung von Rüdiger Fischer, *La Mer gelée*, 6, pp. 74-85.

Conférences données à l'invitation du comité d'organisation dans un congrès national ou international.

29. **Bourion, E**. (2009) « Le dictionnaire, objet culturel », Journée d'études : Perspectives en lexicographie , Université de Chypre, Nicosie, 5 décembre 2009, Actes à paraître.

30. Deléger, L., Merabti, T., Lecrocq, T., Joubert, M., **Zweigenbaum**, P. et Darmoni, S. (2010) A twofold strategy for translating a medical terminology into French. In Proc AMIA Symp, pages 152-156.

31. Deléger, L. et **Zweigenbaum**, P. (2009) Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In Fung et al. (editors), Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-Parallel Corpora, pages 2-10, Singapore, August 2009.

32. Deléger, L. & **Zweigenbaum**, P. (2008) Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In Proceedings AMIA Annual Fall Symposium 2008, pages 146-150, Washington, DC, November 2008. AMIA.

33. **Rastier, F.** (2007) Saussure et la science des textes, *Documents du colloque Révolutions saussuriennes*, Genève, 2007, pp. 81-90.

34. **Rastier, F.** (2007) Saussure au futur – Ecrits retrouvés et nouvelles réceptions, in Montserrat Lopez Diaz et Maria Montes Lopes, Perspectives fonctionnelles : emprunts, économie et variations dans les langues, Saint Jacques de Compostelle, Axac, pp. 73-79. Actes du XXVIIIe congrès de la Silf.

35. **Rastier, F.** (2007) Saussure et la science des textes, in Matsuzawa, Kazuhiro, éd. Saussure et la science des textes, *Proceedings of the Ninth International Conference Studies for The Integrated Text Science*, pp. 61-80.

36. **Rastier, F.** (2007) Traduction japonaise du précédent, par Michihiro Nagata, in Matsuzawa, Kazuhiro, éd. Saussure et la science des textes, *Proceedings of the Ninth International Conference Studies for The Integrated Text Science*, pp. 147-166.

37. **Rastier, F.** (2008) Que cachent les données textuelles ?, Conférence invitée, *Actes des IXe JADT*, Presses Universitaires de Lyon, édités par Serge Heiden et Bénédicte Pincemin, tome I, pp. 13-26.

38. **Rastier, F.** (2008) Des « données » aux documents, conférence invitée, in Interactions et usages autour du numérique, in Maryvonne Holzem et Eric Trupin, éds., Actes du onzième colloque international sur le document numérique, Rouen, Uropia, p. 222-243.

39. **Zweigenbaum**, P. (2009) Multilingualism and medical information processing. In Languages in Biology and Medicine (LBM 2009), page 6, Jeju, Korea. Keynote speech.

Communications avec actes dans un congrès international.

40. **Dutrey, C.**, Peradotto, A., Clavel, Chl. « Analyse de forums de discussion pour la relation clients : du Text Mining au Web Content Mining », », *Actes des 11èmes Journées Internationales d'Analyse statistique des Données Textuelles (JADT 2012)*, 13-15 juin 2012, Liège, pp. 445-457. [C. Dutrey, Master 2 promotion 2010-2011].

41. **Ezzat, M.,** Poibeau, T. (2011) « A new scheme for annotating semantic relations between named entities in corpora », Proceedings of Recent Advances in Natural Language Processing, pages 275–281, Hissar, Bulgaria, 12-14 September 2011.

42. Hagège, C, **Marchal, P.**, Gicquel, Q., Darmoni, S., Pereira, S., Metzger. M.-H. (2010) Linguistic and temporal processing for discovering hospital acquired infection from patient records. *Proceedings of the ECAI 2010 conference on Knowledge representation for health-care (KR4HC'10)*.

43. **Lauf, A.**, **Valette, M.**, Khouas, L. (2012) « Analyse du graphe des cooccurrents de deuxième ordre pour la classification non-supervisée de documents », *Actes des 11èmes Journées Internationales d'Analyse statistique des Données Textuelles (JADT 2012)*, 13-15 juin 2012, Liège, pp. 577-589.

44. **Moreaux, M.-A.,** Détection automatique des unités polylexicales de l'allemand, in Th. Gallèpe et M. Dalmas (eds.) *Déconstruction - Reconstruction, autour de la pensée de J-M Zemb, Rencontre internationales de linguistiques - Tours, 20 et 21 mai 2009*, Lambert-Lucas, Paris, 2011.

45. **Patin, G.** (2010). *Unsupervised Chinese Lexicon Extraction on a Domain Specific Corpus: Method and Evaluation.* In proceeding Coling'10, China, Beijing, 963-971.

46. Proux, D., **Marchal, P.**, Segond, F. Kergourlay, I., Darmoni, S. Pereira, S. Gicquel, Q., Metzger. M.H. (2011) Natural Language Processing to Detect Risk Patterns related to Hospital Acquired Infections. *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing (RANLP 2009).*

47. **Rastier, F.** (2007) Saussure et la science des textes, *Documents du colloque Révolutions saussuriennes*, Genève, pp. 81-90.

48. **Rastier, F.** (2007) Saussure au futur – Ecrits retrouvés et nouvelles réceptions, in Montserrat Lopez Diaz et Maria Montes Lopes, Perspectives fonctionnelles : emprunts, économie et variations dans les langues, Saint Jacques de Compostelle, Axac, pp. 73-79. Actes du XXVIIIe congrès de la Silf.

49. **Rastier, F.** (2007) Saussure et la science des textes, in Matsuzawa, Kazuhiro, éd. Saussure et la science des textes, *Proceedings of the Ninth International Conference Studies for The Integrated Text Science*, pp. 61-80.

50. **Rastier, F.** (2008) Que cachent les données textuelles ?, Conférence invitée, *Actes des IXe JADT*, Presses Universitaires de Lyon, édités par Serge Heiden et Bénédicte Pincemin, tome I, pp. 13-26.

Communications avec actes dans un congrès national.

51. **Bourion , E, Makouar, N.,** (2010) Emergence de concepts et « négociation lexicale » entre acteurs : le champ de la finance islamique, Atelier « Du thème au terme. Emergence et lexicalisation des connaissances », TIA 2009, Toulouse, 18-20 novembre 2009.

52. **Ezzat, M.** (2010) «Acquisition de grammaire locale pour l'extraction de relations entre entités nommées », TALN 2010 (RECITAL), Montréal, 19-23 juillet 2010.

53. Gicquel , Q., Proux, D., **Marchal, P.**, Hagège ,C., Berrouane ,Y., Darmoni, S., Pereira,S. , Segond, F., Metzger, M.-H. (2011). *Evaluation d'un outil d'aide a l'anonymisation des documents médicaux basé sur le traitement automatique du langage naturel*, *Proceedings of 14èmes Journées Francophones d'Informatique Médicale (JFIM2011).*

54. Ke, G., **Zweigenbaum, P**. (2009) « Catégorisation automatique de pages web chinoises : documents spécialisés vs grand public sur le tabagisme », *Proceedings CORIA 2009*, pages 203-218.

55. **Lauf, A.,** Khouas, L., **Valette, M.** (2011), « Calcul de l'autorité des pages Web au sein de leurs communautés respectives - Propositions pour une contextualisation de l'information », Premier atelier Extraction et Contextualisation des Connaissances, en association avec les 22èmes Journées francophone d'Ingénierie des Connaissances (ExCoCo - IC 2011), Chambéry, 2011 (actes électroniques uniquement).

56. **Makouar, N.** (à paraître) « Les "mots" des révolutions : Étude contrastive d'un corpus de journaux arabes » (Colloque « Les ondes de choc des révolutions arabes », 4 février 2012, Paris) Ed. *Inalco*.

57. **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.

58. **Valette, M.** (2010) « Des textes au concept. Propositions pour une approche textuelle de la conceptualisation », *Actes des 21es Journées francophones d'Ingénierie des Connaissances (IC'2010)* (8-11 juin 2010), Nîmes Sylvie Despres, éd., Publication de l'Ecole des Mines d'Alès, pp. 5-16.

Communications orales sans actes dans un congrès international ou national.

59. **Bourion**, **E**. (2009) « Sémantique et contraste de corpus multilingues », Journée « Sciences de la communication et informatique multilingue », Paris, PIRSTEC, 6-8 octobre 2009.

60. **Bourion, E., Aït-Hamlat, J.** (2011) Subjectivité et sentiments : l'éclairage de la sémantique de corpus, *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Slodzian, M.; Valette, M.; Aussenac-Gilles, N.; Condamines, A.; Hernandez, N. & Rothenburger, B. (Eds.)

61. **Daube, J.-M.**, **F. Stuck** (2009) « Construire et utiliser des ressources multilingues », *Journée d'étude Sciences de la communication et informatique multilingue*, PIRSTEC, 7 octobre 2009, Maison de la Recherche, Paris.

62. **Dutrey, C.**, Glorieux, F., **Yankova, G**, Thullier, S. (2011) « Le dictionnaire comme corpus : problèmes, réalisations, expérimentations », *Journée d'étude Lexicographie et Informatique (Université de Cergy-Pontoise)*, janvier 2011 [C. Dutrey et G. Yankova, Master 2 promotion 2010-2011

63. **Eensoo-Ramdani, E.** (2011). Les mots des sentiments : questions émergentes, *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Slodzian, M.; Valette, M.; Aussenac-Gilles, N.; Condamines, A.; Hernandez, N. & Rothenburger, B. (Eds.)

64. **Lauf A.** (2012), "Analyse du graphe des plus proches voisins cooccurrentiels pour détecter des sous-thématiques", *Colloque international : la cooccurrence — du fait statistique au fait textuel*, Besançon, février 2012.

65. **Rastier, F.** (2007) Collège de philosophie, CHU La Salpêtrière, Université de Reims (mai et octobre), Université Lyon II (deux conférences), Université de Bologne, Maison Henri Heine (Paris), Université de Lisbonne, Université de Genève, Fondation Bull, Université de Sienne, SFP (Paris), Université libre internationale de Moldavie, Université Nationale (Chisinau). Université d'Aix-en-Provence. Université de Reggio Emilia et Modena, Université d'Orléans.

66. **Rastier, F. (**2008) Université d'Amiens, Université de Sienne, ENS de Lyon, Université Paris VIII, EHESS, Institut Cervantès (Berlin), Université de Namur, Université des sciences humaines (Tunis) (3 conférences), Facultés universitaires Saint-Louis (Bruxelles), Université de Lyon II (2 conférences). Université de Limoges. Université de Bergen.

67. **Rastier, F. (**2009) Académie Royale de Belgique (Bruxelles), Université des sciences sociales (Tunis), Université d'Arras, Université de Munich (Seeon). Université de Lyon II. Université de Lille III. Université de la Corogne.

68. **Rastier, F. (**2011) Ontologies et folksonomies : même combat ?, *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Slodzian, M.; Valette, M.; Aussenac-Gilles, N.; Condamines, A.; Hernandez, N. & Rothenburger, B. (Eds.)

69. **Slodzian, M.** (2011) Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity? — Introduction and Argument, *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Slodzian, M.; Valette, M.; Aussenac-Gilles, N.; Condamines, A.; Hernandez, N. & Rothenburger, B. (Eds.)

70. **Patin, G.** (2009) « Extraction de Lexique dans un corpus spécialisé en chinois contemporain ». RECITAL, France, Senlis.

Communications par affiche dans un congrès international ou national.

71. **Eensoo, E., Valette, M.** (2012) « Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments », Georges Antoniadis, Hervé Blanchon, Gilles Sérasset, éd., *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Volume 2: *TALN*, 4-8 juin 2012, Grenoble, pp. 367-374.

72. **Lauf A.** (2012) « Sous-graphes de cooccurrences pour la détection de thématiques dans un corpus de taille moyenne », CORIA 2012 (journée jeunes chercheurs), Bordeaux, mars 2012, poster.

Directions d'ouvrages ou de revues.

73. Fung, P., **Zweigenbaum, P.**& Rapp, R. (2009) editors. Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-Parallel Corpora. Association for Computational Linguistics, Singapore, August 2009.

74. Rapp, R., **Zweigenbaum, P.**, & Sharoff, S. (2010) editors. *Proceedings of the 3nd Workshop on Building and Using Comparable Corpora*, LREC 2010. ELDA, Valetta, Malta, May 2010.

75. **Rastier, F**. (2007) *Corpus en Lettres et Sciences sociales — Des documents numériques à l'interprétation*, Texto ! (pour l'édition numérique) et CALS-Presses universitaires du Mirail (pour l'édition papier). Co-direction avec Michel Ballabriga.

76. **Rastier, F.** (2009) *Plurilinguisme, interculturalité et emploi : défis pour l'Europe* (édition, en collaboration avec François-Xavier d'Aligny, Astrid Guillaume, Babette Nieder), Paris, L'Harmattan.

77. **Slodzian, M., Valette, M.**, éd. (2010) *Du thème au terme. Émergence et lexicalisation des connaissances (TIA'09 Workshop, 20 novembre 2009, Toulouse)*, *CEUR Workshop Proceedings*, vol. 579. ISSN : 1613-0073.

78. **Slodzian, M.**, **Valette, M.**, Aussenac-Gilles, N., Condamines, A., Hernandez, N. & Rothenburger, B. (Eds.) Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity? / Ontology and Lexicon: new insights, *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Paris.

79. **Valette, M., éd.** (2008b) *Textes, documents numériques, corpus. Pour une science des textes instrumentée, Syntaxe & Sémantique*, n°9/2008.

80. **Zweigenbaum, P.**, Gaussier, E. & Fung, P., (2008) editors. Proceedings LREC Workshop Building and using comparable corpora. ELRA, Marrakech, Morroco, 2008.

81. **Zweigenbaum, P.**, Rapp, R. & Sharoff, S. (2011) ed. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. Association for Computational Linguistics*, Portland, Or., June 2011.

82. Kageura, K., **Zweigenbaum, P.** (2011) ed. *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence* (TIA'11), http://tia2011.crim.fr/Proceedings/index.html, 152 pages.

Outils de recherche, corpus de recherche, cohortes.

Sans objet.

Ouvrages scientifiques (y compris les éditions critiques et les traductions scientifiques).

— Chapitres d'ouvrage

83. **Rastier, F.** (2007) Conditions d'une linguistique des normes, in Siouffi, Patrick et Steuckart, Agnès, éd., *Les linguistes et la norme — Aspects normatifs du discours linguistique*, Berne, Lang, pp. 3-20.

84. **Rastier, F.** (2007) Indices et parcours interprétatifs, in Denis Thouard, éd. *L'interprétation des indices*, Lille, Presses du Septentrion, pp. 123-152.

85. **Rastier, F.** (2007) Les langues sont-elles des instruments de communication ?, in Fernandez-Vest Marie- Rastier F, Madeleine, éd., *Combat pour les langues du monde — Hommage à Claude Hagège*, Paris, L'Harmattan, pp. 421-432.

86. **Rastier, F.** (2007) Semantica interpretativa, in Paolucci, C. éd. *Studi di semiotica interpretativa*, Milan, Bompiani, pp. 203-286.

87. **Rastier, F.** (2007) articles : acte de langage, allotopie, cohésion, domaine sémantique, impression référentielle, interprétation, noème, taxème, trope, asémanticité, in *Semiotica. Dizionario ragionato della teoria del linguaggio*, Milan, Mondadori.

88. **Rastier, F.** (2007) La linguistique comme science des textes — Des corpus aux systèmes, Romanische Syntax in Wandel, Tübingen, Narr, pp. 499-512. [*Festschrift Wulf Oesterreicher*].

89. **Rastier, F.** (2008) Warten auf Valentin Temkine, in Pierre Temkine, éd. *Warten auf Godot. Das Absurde und die Geschichte*, Matthes & Seitz, Berlin, 2008 (En attendant Valentin Temkine, traduction allemande par Tim Trzaskalik), pp. 43-94.

90. **Rastier, F.** (2008) Silence des disparus et dialogue des œuvres, pp. 443-466 ; in Mesnard, P. et Thanassekos, Y. éds, *Primo Levi à l'œuvre*, Paris, Kimé.

91. **Rastier, F.** (2008) Rhétorique et interprétation des figures, in Sémir Badir et Jean-Marie Klinkenberg, éd., *Figures de la figure. Sémiotique et rhétoriques générales*, Limoges, Presses Universitaires de Limoges, pp. 81-101.

92. **Rastier, F.** (2008) Croc de boucher et rose mystique — Le pathos sur l'extermination, in Michael Rinn, éd. *Émotions et discours — L'usage des passions dans la langue*, Presses universitaires de Rennes, pp. 249-273.

93. **Rastier, F.** (2008) Le langage sans origine ou l'émergence du milieu sémiotique, in Régine Delamotte-Legrand et coll., éds., Dialogues, mouvements discursifs, significations, *Hommage à Frédéric François*, Cortil-Wodon, EME, pp. 207-222.

94. **Rastier, F.** (2008) L'ipallage e Borges, in Migliore, Tiziana, *Argomentare il visibile. Esercizi di retorica dell'immagine*, Bologne, Esculapio, pp. 93-120.

95. **Rastier, F.** (2008) "Retórica e interpretación de las figuras", en H. Beristáin y G. Ramírez Vidal (compiladores). *Las figuras del texto*. México: UNAM, 2008 (Bitácora de retórica 26), pp. 47-73.

96. **Rastier, F.** (2009) Traduction et linguistique des textes, in Tatiana Milliaressi, *La traduction : philosophie, linguistique et didactique*, Travaux et recherches, Université de Lille 3, p. 35-38.

97. **Rastier, F.** (2009) Préface à la troisième édition, *Sémantique interprétative*, Paris, PUF, pp. I-XIV.

98. **Rastier, F.** (2009) Éloge paradoxal du plurilinguisme, in 2009 f., pp. 15-28.

99. **Rastier, F.** (2009) La sémiotique des cultures, in Driss Ablali et Dominique Ducard, Vocabulaire des études sémiotiques et sémiologiques, Paris, Champion, 2009, pp. 89-95 [avec la collaboration de Carine Duteil Mougel].

100. **Rastier, F.** (2009) Entretien avec F. Stjernfeld et P. Bundgaard, éds., *Signs and Meaning — Five Questions*, s. l., Automatic Press, pp. 139-152;

101. **Rastier, F.** (2009) Pour un remembrement de la linguistique : enquête sur la sémantique et la pragmatique, in Dominique Verbeken, éd. *Entre sens et signification — Constitution du sens : points de vue sur l'articulation sémantique-pragmatique*, Paris, L'Harmattan, pp. 251-278.

102. **Rastier, F.** (2009) Naturalisation et culturalisation, in *L'évolution aujourd'hui : à la croisée de la biologie et des sciences humaines*, Bruxelles, Académie Royale de Belgique, pp. 231-250.

103. **Rigot, H.** (2008) « Activités de recherche en SHS et systèmes d'information organisationnels : cas limite ou prototype ? » dans L'information dans les organisations : dynamique et complexité, sous la dir. de Christiane Volant. Presses Universitaires François Rabelais.

104. Véronis, J., Hamon, O., Ayache, C., **Belmouhoub, R.**, Kraif, O., Laurent, D., Nguyen, T. M. H., Semmar, N., **Stuck, F.**, Zaghouani, W. (2008). La campagne d'évaluation Arcade II. In Chaudiron, S. & Choukri, K. (Eds.) *L'évaluation des technologies de traitement de la langue* (pp. 47-69). Paris: Hermès-Lavoisier.

— Ouvrages

105. Baranová, E., Krecková, V., Lemay, D. et Pognan, P. (2007) *Découvrir et pratiquer le slovaque*, L'Asiathèque, Paris (ISBN : 978-2-91-525546-1) [Méthode de langue interactive (CD-ROM) réalisée sous la supervision de **François Stuck** dans le cadre du projet européen ALPCU].

106. Bernard, A., Fournier, P., Horvat, S. et Jesensek, M. (2007) *Découvrir et pratiquer le slovène*, L'Asiathèque, Paris (ISBN : 978-2-91-525547-8) [Méthode de langue interactive (CD-ROM) réalisée sous la supervision de **François Stuck** dans le cadre du projet européen ALPCU].

107. Chraïbi, S., **J.-M. Daube** (2007) *E-lexique français-anglais-arabe des relations internationales*, PUF, Paris, ISBN : 978-2-13-055979-5, 224 pages.

108. **Daube, J.-M.** (2007) *E-lexique français-anglais des médias*, PUF, Paris, ISBN : 978-2-13-055977-1, 112 pages.

109. **Daube, J.-M.** (2007) *E-lexique français-anglais de la vie politique française*, PUF, Paris, ISBN : 978-2-13-055978-8, 97 pages.

110. **Rastier, F.** (2009) *Ulisse ad Auschwitz — Primo Levi, il superstite*, Liguori, coll. Profili, Naples. Traduction par Rossella Saetta-Cottone et Daria Francobandiera de Ulysse à Auschwitz, Paris, Cerf, 2005.

111. **Rastier, F.** (2010) traduction arabe de *Arts et sciences du texte*, Casablanca, Toubkal. Traducteur : Driss El Khattab.

112. **Rastier, F.** (2010) *AÇÃO E SENTIDO — POR UMA SEMIÓTICA DAS CULTURAS*, Editora Universitária da UFPB, João Pessoa-PB. Traduction de *L´action et le sens pour une sémiotique des cultures*, Maria de Fátima Barbosa de Mesquita Batista.

113. Vrinat, M. Krasteva, T. et Tchoukanova, Y. (2007) *Découvrir et pratiquer le bulgare*, L'Asiathèque, Paris (ISBN : 978-2-91-525545-4) [Méthode de langue interactive (CD-ROM) réalisée sous la supervision de **François Stuck** dans le cadre du projet européen ALPCU].

114. **Zweigenbaum, P** & Demner-Fushman, D. (2009) Advanced literature mining tools. In David Edwards, David Hansen, and Jason Stajich, editors, Bioinformatics: Tools and Applications, chapter 21, pages 347-380. Springer, 2009.

Publications de vulgarisation. Productions artistiques théorisées (compositions musicales, cinématographiques, expositions, installations...).

115. **Rastier, F. (**2007 m) Muerte espectáculo y éxito anunciado, Barcelone, *La Vanguardia*, 15.11.07, Cultura/s, p. 4.

116. **Rastier, F. (**2007) Primo Levi aujourd'hui. Avec Ruth Scheps, Radio suisse romande. 13 mai 2007, 20-22h.

117. **Rastier, F. (**2007) Télévision : Entretien, Télévision Moldave, 15 octobre 2007, 21h.

118. **Rastier, F. (**2007) Mort-spectacle et succès annoncé, *Bulletin de la Fondation Auschwitz*, 97, pp. 111-113 [compte rendu de Jonathan Littell, *Les Bienveillantes*, Paris, Gallimard, 2006].

119. **Rastier, F. (**2008) A Primo Levi, [traduction d'un poème de Giovanni Abbate], in Mesnard, P. et Thanassekos, Y. éds, *Primo Levi à l'œuvre*, Paris, Kimé, p. 21.

120. **Rastier, F. (**2009) L'angoisse du Surhomme, *Controverses*, 12, pp. 275-278. Compte rendu de Rafael Cutillas, *Viuvre mata*, Fonoll, Barcelone, 2006.

121. **Valette, M.** (2009) *5/5=1, 5 trans-fusions,* soirée organisée par Christophe Huysman et Jacques André à La ménagerie de Verre, Paris XIe. Entretien sur le racisme (40'). 14 mars 2009.

122. **Valette, M.** (2010) *Tire ta langue !*, France Culture, émission d'Antoine Perraud. Entretien sur le racisme (25') (7 mars 2010)

Publications de transfert. Autres productions : bases de données, logiciels enregistrés, rapports de fouilles, guides techniques, catalogues d'exposition, rapports intermédiaires de grands projets internationaux, etc.

123. **Valette, M.** (2009) *Approche textuelle du lexique*, mémoire pour l'Habilitation à Diriger des Recherches, Institut National des Langues et Civilisations Orientales, Paris.