Achille Falaise

http://pro.aiakide.net
achille.falaise@gmail.com

Ingénieur en linguistique-informatique

Informatique

Java, Perl, Python, R
HTML/CSS, PHP, JavaScript
Linux, Windows, MacOS
Apache, Tomcat, MySQL
Administration serveur Linux

Linguistique

Linguistique de corpus
Linguistique formelle
Statistiques textuelles

TALN

Analyse morphologique, lexicale, syntaxique et sémantique
Fouille de textes, apprentissage et traduction automatiques
Création, annotation, exploitation et diffusion de corpus textuels
Multilinguisme

Standards

TEI, OAXAL (Unicode, XML, SRX, TMX, XLIFF, SVG), OWL, RDF, LMF, DICT, GraphViz

Données

BD relationnelles, SQL, JSON, Redis, XML (DOM, SAX, XSLT)
Arbres, treillis, graphes

Langues

Français (langue maternelle), Anglais (~C1), Allemand (~A2), Japonais (~A1), Latin

Loisirs

Reconstitution historique, combat médiéval

Parcours professionnel

2018 2016	2 ans	**CNRS, laboratoire ICAR − UMR 5191** (IGR contractuel) Création, annotation, exploitation et diffusion d'un corpus de référence du français des XVIe-XVIIIe siècles (suite du projet précédent).
2016 2014	30 mois	**ENS de Lyon, laboratoire ICAR − UMR 5191** (IGR contractuel) Création, annotation, exploitation et diffusion d'un corpus de référence du français des XVIe-XVIIIe siècles.
2013	3 mois	**LIDILEM − EA 609** (IGR contractuel) Conception d'une base de données pour l'exploration de grands corpus arborés (>300M tokens).
2013 2012	1 an	**Grenoble-2** (ATER) Enseignement d'informatique en MIASS : développement Web, Java.
2012 2011	16 mois	**LIG-GETALP − UMR 5217** (IGR contractuel) Conception d'un environnement collaboratif de traduction de sites Web.
2010 2009	1 an	**LIG-GETALP − UMR 5217** (IGR contractuel) Conception d'une base de données pour la recherche multilingue dans des bases d'images de presse.
2009 2008	1 an	**LIDILEM − EA 609** (IGE contractuel) Annotation et conception d'IHM pour l'exploitation de corpus arborés.
2008 2005	3 ans	**Prosodie SA** et **LIG-GETALP − UMR 5217** (doctorant) Conception d'un outil d'assistance au dialogue en langue étrangère.
2006	100 h	**Grenoble-2** (vacataire) TP/TD d'informatique: Java et Système.
2004	4 mois	**CLIPS-GETA** (stage) Tchat et multilinguisme.
2003	3 mois	**Hyper-Horizon** (stage) Réalisation d'une application d'édition de devis multilingues.
2003	2 mois	**CEA-LETI** (stage) Décompilation et graphes algorithmiques.
2002	1 mois	**LIDILEM − EA 609** (stage) Réalisation d'une IHM pour un corpus de notes en langue seconde.
2001	3 mois	**CLIPS-GEOD** (stage) Étude de la prosodie des nombres pour la synthèse vocale.

Parcours académique

2009	D	**Doctorat** CIFRE, spécialité **Informatique** Université Grenoble-1
2004	M2R	**Master**, spécialité **Sciences cognitives** Institut National Polytechnique de Grenoble – *mention bien*
2003	M2P	**DESS Informatique Double Compétence** Université Grenoble-2 – *mention bien*
2002	M1	**Maîtrise Sciences du Langage**, spécialité **Industrie de la Langue** Université Grenoble-3 – *mention bien*

CV détaillé

Table des matières

Formation académique

2009 D **Doctorat**, spécialité **Informatique**
Université Grenoble I (UJF) – *cette université ne délivre pas de mention pour les thèses*
GETALP-LIG, en convention CIFRE avec la société Prosodie
Titre : *Conception et prototypage d'un outil web de médiation et d'aide au dialogue tchaté écrit en langue seconde*
Laboratoire : LIG (équipe GETALP)
Directeurs de thèse : Christian Boitet et Hervé Blanchon
Président du jury : Jean Caelen
Rapporteurs : Violaine Prince et Patrice Pognan
Examinateurs : Emmanuel Planas, Rémi Marand (Prosodie)

2004 M2R **Master Information, Cognition, Apprentissage** (ICA), spécialité **Sciences Cognitives**
Institut National Polytechnique de Grenoble (INPG) – *mention bien*
Sujet de stage : étude du tchat en langue maternelle et en langue étrangère
Responsables : Christian Boitet et Hervé Blanchon (GETA-CLIPS)

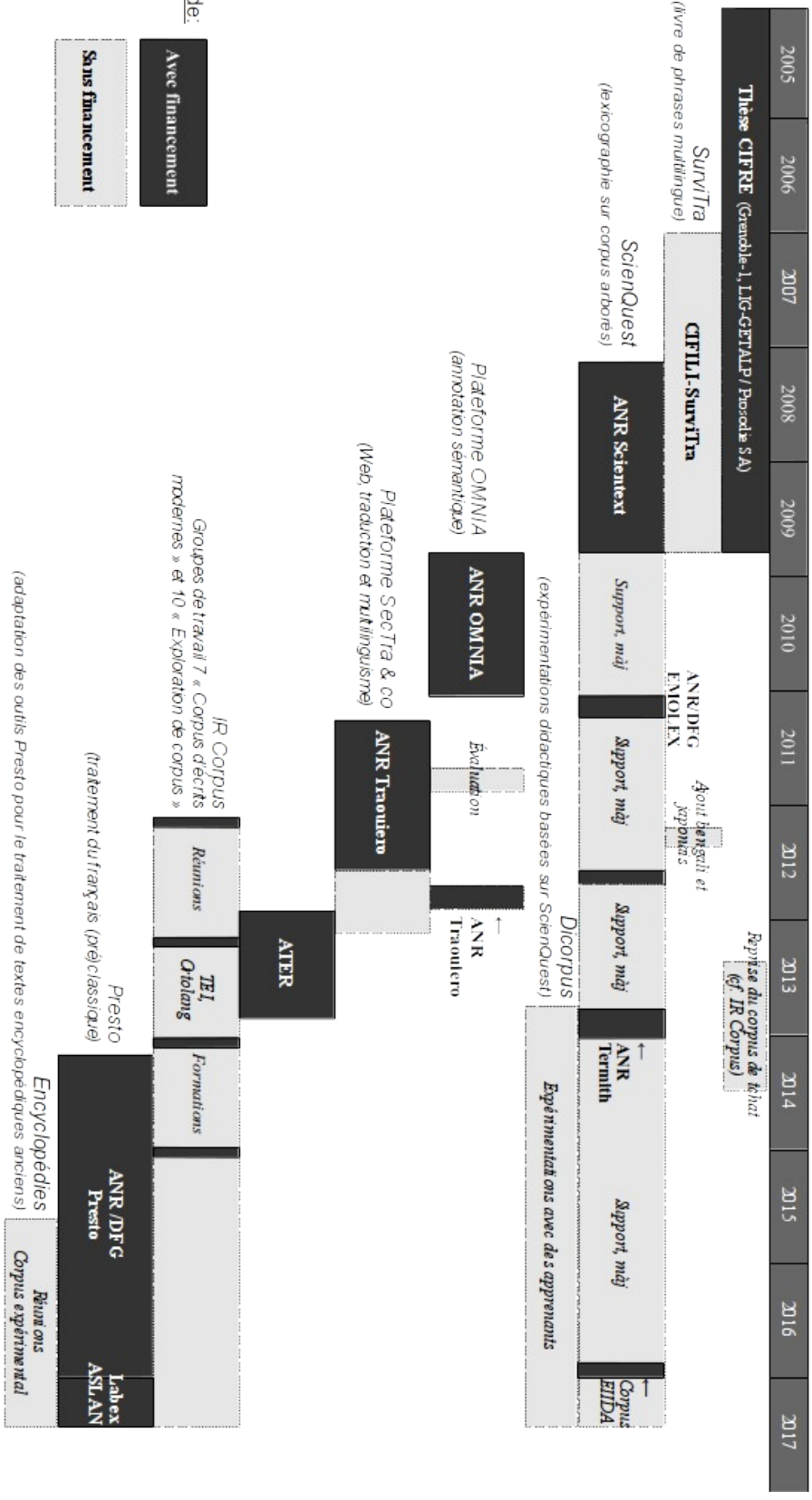
2003 M2P **DESS Double Compétence: Informatique et Sciences Sociales** (DCISS)
Université Grenoble II (UPMF) – *mention bien*
Sujet de stage : développement d'un générateur *web* de devis multilingues
Responsable : Jacques Bandet (société Hyper Horizon)
Sujet de projet : construction automatique de graphes algorithmiques à partir de code compilé
Responsables : Axel Bonnes, Philippe de Choudens, Jean-Pierre Krimm (LETI-CEA)

2002 M1 **Maîtrise Sciences du Langage**, spécialité **Industrie de la Langue** (IdL)
Université Grenoble III (Stendhal) – *mention bien*
Sujet de stage : conception de règles prosodiques pour la synthèse vocale des nombres
Responsable : Jean-François Sérignat (GEOD-CLIPS)

2001 L3 **Licence Sciences du Langage**, option **Traitement Automatique des Langues** (TAL)
Université de Nantes – *mention assez bien*

2000 L2 **DEUG Lettres modernes**
Université de Nantes – *mention assez bien*

1998 Bac **Baccalauréat général**, série **Littéraire**, spécialité **Mathématiques**
Lycée Léonard de Vinci, Montaigu (Vendée) – *mention assez bien*


Récapitulatif des projets (2005-2017)
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Koiné (assistance au tchat en langue étrangère)
These CIFRE (Grenoble-1, LIG-GETALP / Prosodie SA)
SurviTra (livre de phrases multilingue)
CIFILI-SurviTra
ScienQuest (lexicographie sur corpus arbores)
ANR Scientext
Support, màj
Support, màj
Support, màj
Support, màj
Ajout benguli et lepoma s
ANR/DFG EMOLEX
Reprise du corpus de tchat (cf. IR Corpus)
Plateforme OMNIA (annotation sémantique)
ANR OMNIA
Évaluation
ANR Traouiero
Plateforme SecTra & co (Web, traduction et multilinguisme)
ANR Traouiero
ATER
Dicorpus (expérimentations didactiques basées sur ScienQuest)
ANR Termith
Expérimentations avec des apprenants
Corpus EIDA
IR Corpus (Groupes de travail 7 « Corpus d'écrits modernes » et 10 « Exploration de corpus »)
Réunions
TEI, Ortolang
Formations
Presto (traitement du français (préclassique)
ANR /DFG Presto
Labex ASLAN
Encyclopédies (adaptation des outils Presto pour le traitement de textes encyclopédiques anciens)
Réunions
Corpus expérimental
Légende:
Avec financement
Sans financement

Mes activités s'inscrivent dans le domaine du traitement automatique des langues naturelles (TALN), et en particulier des corpus (textuels, oraux, multimodaux) : collecte de corpus, traitements de corpus, et création d'outils d'exploitation de corpus. Grâce à ma double compétence en informatique et sciences du langage, j'ai pu mener des collaborations sur les deux versants du TALN, aussi bien avec des équipes d'informaticiens que des équipes de linguistes.

TER de maîtrise (03/2002 – 06/2002) : étude de la prosodie des nombres

Lors de ce stage, effectué au **CLIPS-GEOD**[1] (Grenoble) sous la direction de Jean-François Sérignat, j'ai étudié la prosodie des nombres en français et démontré que lorsqu'elle constituait l'information saillante d'un énoncé, elle suivait un schéma caractéristique différent de celui d'un énoncé normal.

M2R (03/2004 – 06/2004) et doctorat (01/2005 – 09/2009) : aide au dialogue en langue seconde

Mon stage de M2R, dirigé par Christian Boitet, et mon doctorat, dirigé par Christian Boitet et Hervé Blanchon, se sont tous deux déroulés au **LIG-GETALP** (Grenoble), en convention CIFRE avec **Prosodie** (Boulogne-Billancourt) pour la thèse. Mon travail s'inscrit dans la thématique des aides au dialogue, en l'occurrence dans le cadre d'un dialogue en langue seconde[2].

Il existe aujourd'hui de nombreux outils destinés à aider la communication en langue étrangère, et consistant à intercaler un système informatique entre les locuteurs. Ce peut être par exemple un livre de phrases, ou un logiciel de traduction automatique. Pourtant ces solutions sont aujourd'hui peu utilisées, en regard des millions de locuteurs en situation de multilinguisme, et parfois en difficulté linguistique.

Afin de comprendre la raison de ce paradoxe, j'ai étudié les stratégies des locuteurs en situation "bruitée", par la langue (dialogue langue seconde) ou par le canal (tchat), et j'ai collecté à cette occasion le plus gros corpus de tchat actuellement disponible (24M mots). J'ai pu montrer que les locuteurs coopèrent énormément dans ces situations, et que cette coopération était mal prise en compte et perturbée, voire totalement inhibée, par les solutions actuelles de médiation par ordinateur.

J'ai alors étudié et décrit les problèmes rencontrés par les locuteurs à l'oral et à l'écrit, afin de déterminer des aides informatiques adaptées. La modalité orale est largement couverte dans la littérature[3], mais pour le dialogue écrit, mes observations, les premières à être basées sur un corpus de taille significative, constituent un apport important.

J'ai alors pu proposer des fonctionnalités d'aide utiles qui viennent épauler et prolonger les stratégies des locuteurs, au lieu de les contrarier. J'ai mis en évidence les problèmes d'implémentation de ces fonctionnalités et proposé des solutions (pour l'écrit), que j'ai implémentées.

Projet CIFLI-SurviTra (07/2007 - 06/2009) : livre de phrases et traduction automatique par pivot

Ce projet a été mené au **LIG-GETALP** en collaboration avec le laboratoire CFILT-IIT de Bombay. J'y ai participé, sous la direction de Georges Fafiotte. CIFLI-SurviTra (*Survival Translation assistant*) est une plate-forme destinée à favoriser l'ingénierie et la mise au point de composants de TA, à partir d'une mémoire de traduction formée de livres de phrases multilingues avec variables lexicales. En particulier, j'ai travaillé à consolider nos propres composants de TA, et ai travaillé à leur intégration avec ceux de nos partenaires indiens du **CFILT-IIT** au sein d'un prototype.

Projet Scientext (01/2008 - 06/2009) : création de corpus arborés et plateforme d'exploitation ScienQuest

J'ai mené ce projet ANR au laboratoire **LIDILEM** (Grenoble-3), sous la direction d'Agnès Tutin, en collaboration avec le LiCoRN (Bretagne sud), le LLS (Savoie), Didier Bourigault (Synomia) et Franck Sajous (ERSS-CLLE, Toulouse-2). Le projet vise à constituer un corpus d'écrits scientifiques variés, structurés, analysés syntaxiquement, en français et en anglais, permettant d'effectuer une étude linguistique du positionnement et du raisonnement dans l'écrit scientifique.

J'ai assuré le suivi informatique du projet, du pré-traitement du corpus jusqu'au développement de l'interface

1 Le CLIPS est un ancien laboratoire grenoblois dont la plupart des équipes ont rejoint le LIG. À cette occasion, l'équipe GEOD (du CLIPS) a fusionné avec l'équipe GETA (du CLIPS) pour former le GETALP (du LIG).

2 C'est à dire une langue qui n'est pas la langue maternelle.

3 Notamment pour l'anglais dans le cadre de l'*English as Lingua Franca* (*ELF*).

d'étude. La difficulté venait principalement de l'analyse syntaxique, qu'il fallait rendre accessible aux utilisateurs du système. Il existe des systèmes très puissants pour utiliser ce genre de corpus, mais il sont complexes, et ne sont utilisés que par une fraction des linguistes qui en auraient besoin. C'est pourquoi j'ai effectué un important travail d'étude ergonomique, avec de fréquentes étapes de test avec des utilisateurs, pour aboutir à un compromis acceptable entre simplicité et puissance. L'interface ScienQuest que j'ai développée suivant ces idées directrices est librement accessible[4].

J'ai aussi développé un langage de grammaire pour le moteur ConcQuest d'Olivier Kraif, utilisant les dépendances syntaxiques, les relations linéaires et des variables, et utilisé ce langage pour construire des grammaires d'extraction d'informations sémantiques ciblées avec Agnès Tutin. Enfin, la taille de certains corpus a vraiment constitué un défi pour certains traitements. Il a été initialement développé pour 1 corpus français (5M mots) et 2 anglais (1M et 33M mots).

Depuis la fin formelle du projet, j'ai travaillé à le rendre plus générique, à améliorer son ergonomie et enrichir ses fonctionnalités, et j'ai pu y intégrer d'autres corpus. Dans le cadre du projet informel Dicorpus (2013+), je développe une version allégée de ScienQuest pour les étudiants étrangers apprenant à rédiger en français scientifique.

Après la soutenance de ma thèse (09/2009), j'ai enchaîné sur un post-doctorat sur le projet OMNIA au GETALP, tout en intervenant régulièrement sur la plateforme ScienQuest.

Projet OMNIA (10/2009 – 09/2010) : analyse et indexation sémantique de dépêches multilingues pour la recherche d'information

Il s'agit d'un projet ANR, en collaboration avec le LIRIS et XEROX, effectué au **LIG-GETALP** sous la direction de Christian Boitet. Il s'agit de recherche d'images légendées en langue naturelle pour la prépresse. Pour cela, on annote les textes et les requêtes avec des lexèmes interlingues, puis on désambiguïse, et enfin on effectue une extraction de descripteurs permettant d'indexer les objets (images+légendes) dans une base de données ou dans l'ontologie utilisée pour l'extraction de contenu.

Je me suis chargé du processus d'analyse, d'annotation, de désambiguïsation interactive des textes et d'intégration des composants de désambiguïsation automatique et ontologiques, reliant ainsi entre elles plusieurs problématiques de l'équipe.

Dans le cadre de ce projet, j'ai travaillé sur :

- La lemmatisation de textes pour des langues avec ou sans séparateurs, en générant un treillis de lemmes, au format des graphes-Q d'Alain Colmerauer. Pour cette lemmatisation, j'ai travaillé avec les dictionnaires DELA et DELAF, et les plateformes de ressources lexicales PIVAX et Jibiki.
- L'annotation des textes précédemment lemmatisés avec des lexèmes interlingues (en l'occurrence, les *Universal Words* du consortium UNL).
- L'intégration du système de désambiguïsation lexicale à fourmis de Didier Schwab.
- L'annotation des textes interlinguisés avec des concepts issus d'une ontologie.
- L'indexation et la recherche d'information dans les textes ainsi analysés.

Ce projet posait une problème de passage à l'échelle particulièrement important. J'ai conçu une architecture d'intégration multitâches adaptée à cette difficulté, et développé certains composants (lemmatisation, extraction de contenu). J'ai donc revu, et amélioré certains algorithmes utilisés pour le projet. Enfin, le démonstrateur que j'ai développé a permis une évaluation de l'approche, à laquelle j'ai pris une part active.

Projet EMOLEX (12/2010 – 01/2011) : lexique multilingue des émotions

Ce projet est dirigé par Iva Novakova et Peter Blumenthal, et concerne l'étude du lexique des émotions. Il s'agit d'une collaboration entre le **LIDILEM** et les universités de Cologne et d'Osnabrück, sur financement ANR/DFG[5].

Je suis intervenu durant quelques semaines de vacations, pour ajouter de nouveaux corpus en allemand,

4 http://corpora.aiakide.net

5 *Deutsche Forschungsgemeinschaft*

français, anglais, russe et espagnol, annotés avec des analyseurs syntaxiques différents (XIP, Connexor et DeSR). Il fallait les intégrer à l'environnement existant, initialement conçu pour des corpus analysés avec Syntex, et ajouter des fonctionnalités d'analyse distributionnelle. Dans le cadre de se projet, je me suis rendu quelques jours à l'université de Cologne (département de romanistique) pour travailler avec Sascha Diwersy à l'intégration de composants d'analyse distributionnelle qu'il avait conçu (le Lexicoscope).

Projet Traouiero (01/2011 – 09/2012) : plateforme pour la traduction/post-édition de sites Web, opérationnalisation de linguiciels de TA

Le projet Traouiero (TRAduction : Outils Unifiés, Intégrables, Embarquables, et Ressources Opérationnelles) vise à l'opérationnalisation d'outils logiciels et de techniques et ressources linguicielles développés au **LIG-GETALP**, et utilisables à terme par deux futures sociétés de valorisation : AXiMAG (Grenoble, pour des passerelles de traduction collaborative de sites web) et Taranis (Paris, pour la traduction automatique ciblée sur des sous-langages ou de langues peu-dotées).

Sur ce projet, je suis responsable du développement et/ou de l'opérationalisation de plusieurs logiciels, que j'effectue avec la collaboration d'enseignants-chercheurs et de doctorants de l'équipe :

1. SECTra : il s'agit d'opérationaliser un prototype de gestionnaire de mémoires de traductions, comprenant des fonctions avancées de post-édition et d'évaluation. Après une analyse du logiciel existant, j'ai pu montrer la nécessité d'une réimplémentation, que j'ai mené en collaboration avec le doctorant Wang Lingxiao.
2. TRADOH : un méta-moteur de traduction, présentant une API unifiée pour appeler des systèmes de traduction tiers, et éventuellement d'en chaîner plusieurs, en fonction des langues demandées. J'ai élaboré ce système avec Võ Trung Hùng (Université de Danang, Vietnam), à partir de la version développée pour sa thèse ; j'ai en outre travaillé avec Tang Enya Kong et Chai Chong Chua à l'intégration du système Sistec pour l'anglais/malais développé à l'Université Multimedia (Cyberjaya, Malaisie). Enfin, un client Android pour TRADOH[6] a été développé par Antoine Le Maire lors d'un stage dans l'équipe.
3. SANDOH : un système statistique de reconnaissance des encodages et des langues d'un document textuel (éventuellement hétérogène), élaboré avec Võ Trung Hùng (université de Danang, Vietnam).
4. SegDoc : un segmenteur de pages Web, développé avec Ruslan Kalitvianski.
5. iMAG : ce logiciel vise à faire évoluer le processus traditionnel de traduction de sites web, dans lequel les sites sont traduits en une fois (et ne sont plus modifiables par la suite), pour introduire la notion d'accès multilingue, dans lequel le site est traduit au fur et à mesure de façon collaborative. Il s'agit essentiellement d'une tâche d'intégration des autres composants du projet : segmentation de pages Web (avec SegDoc), traduction des segments (avec TRADOH) et reconstruction de la page avec les segments traduits (à nouveau avec SegDoc).

Projet SurviTra-JF (01/2012) : livres de phrases pour la reconnaissance de la parole multilingue

Début 2012, j'ai participé à l'adaptation des livres de phrases de SurviTra, pour la construction d'un modèle de langage destiné à la reconnaissance de parole, dans le cadre d'un projet conjoint entre le LIG-GETALP et l'université de Kyōto.

ANR/DFG Presto (04/2014-11/2016) : traitement automatique du français en diachronie

Ce projet est dirigé par Denis Vigier et Peter Blumenthal, et concerne l'étude de l'évolution des prépositions en français du 16e au 21e siècle. Il s'agit d'une collaboration entre le laboratoire **ICAR** (ENS de Lyon) et l'université de Cologne, sur financement ANR/DFG[7]. L'ATILF est aussi fortement impliquée dans le projet.

Dans ce projet, j'analyse des textes en français des 16e-21e siècles, en recourant, pour la période 16e-18e siècles, à l'adaptation de ressources dérivées du français moderne (avec Gilles Souvay, IGR CNRS/ATILF), en associant des méthodes classiques utilisées pour l'informatisation des langues peu dotées (étudiées lors de mes passages au GETALP), et des méthodes utilisées actuellement pour le traitement des langues anciennes.

6 https://play.google.com/store/search?q=%22tradoh%22&c=apps
7 *Deutsche Forschungsgemeinschaft*

J'utilise ensuite ce corpus (23M tokens), en lien avec les autres membres du projet, selon une approche distributionnelle rendue possible par l'analyse préalable du corpus, pour étudier l'évolution du système prépositionnel du français. À cette fin, j'adapte, avec l'aide de Sascha Diwersy (MCF Montpellier-3), quelques méthodes statistiques utilisées en synchronie à ce cadre diachronique. Le financement sur ce projet s'arrêtant fin 2016, je poursuis actuellement ce travail sur un financement du Labex ASLAN, partagé entre le projet Presto, et des travaux de traitement et d'exploitation des autres corpus du laboratoire ICAR (actuellement, corpus diachronique Men'hir : menus de restaurant). Je suis par ailleurs associé à un projet concernant les écrits encyclopédiques des 18e-21e siècles (projet « Encyclopédies », partenariat entre le laboratoire ICAR et les universités de Neuchâtel et Chicago), qui a débuté en marge du projet Presto, et pourrait se développer sur un financement du Fonds national suisse de la recherche scientifique (FNS).

Collaborations internationales

L'activité d'un chercheur est parsemée de petites et de grandes collaborations. J'indique ici les principales collaborations impliquant des collègues étrangers.

Date	Langue de travail	Nature de la collaboration
2017	Anglais	Collaboration avec Guan Xiaojing (Institut d'études mandchoues, Académie des sciences sociales de **Pékin**), dans le cadre du Collegium de Lyon, pour l'aider à mettre en forme et à exploiter un corpus parallèle d'inscriptions lapidaires bilingues (mandchou/chinois).
2016-2017	Français	Collaboration avec l'Université de **Cologne**, l'Université de **Neuchâtel** et l'ARTFL (**Chicago**) sur les textes encyclopédiques anciens; association à un futur projet suisse financé par le FNS (demande déposée).
2015-2017	Français	Collaboration (en cours) avec Carine Skupien Dekens et Corrinne Rossari (Université de **Neuchâtel**) sur l'annotation de textes des 16^{e}-18^{e} siècles.
2015-2017	Français	Collaboration (en cours) avec l'ARTFL (**Chicago**) au sujet de l'annotation de l'Encyclopédie de Diderot et d'Alembert.
2009 2012	Anglais	Collaboration avec Sanjay Meena et Avishek Dan (*Indian Institute of Technology*, **Bombay**), pour l'intégration dans SurviTra du déconvertisseur indien de graphes sémantiques UNL vers le hindi (2009 et 2012).
2011	Anglais	Collaboration avec Võ Trung Hùng (Université de **Danang, Vietnam**), lors de son séjour à Grenoble en 2011, pour l'opérationalisation de TRADOH et SANDOH.
2011 2014-2017	Français	Collaboration avec Sascha Diwersy (Université de **Cologne**) pour l'intégration du Lexicoscope dans ScienQuest, lors de mon séjour à Cologne, en 2011, puis à nouveau sur le projet Presto de 2014 à 2017.
2011	Anglais	Collaboration avec Enya Kong et Chai Chong Chua (Université Multimedia, **Cyberjaya, Malaisie**) pour l'intégration de leur logiciel Sistec dans TRADOH, lors de mon séjour à Cyberjaya, puis du séjour d'Enya Kong à Grenoble, en 2011.
2010	Anglais	Collaboration avec Viacheslav Dikonov (*Institute for Information Transmission Problems*, **Moscou**) pour l'annotation sémantique et la désambiguïsation de textes, lors de son séjour à Grenoble, en 2010.

Montage de projets

Au-delà de la rédaction de quelques paragraphes pour des projets ANR ou FNS, je me suis impliqué dans plusieurs demandes de financement.

Date	Financement	Équipe d'accueil et contact	Thème
2016	Labex ASLAN	ICAR, CNRS, ENS de Lyon Denis Vigier	Corpus diachroniques et corpus multimodaux.
2016	CORLI/Ortolang	GREMUTS, Université de Grenoble Caroline Rossi	Corpus multilingue comparable « COP 21 ».
2014	CMIRA	RALI, Université de Montréal Guy Lapalme	Corpus pour l'aide à la rédaction en langue étrangère.
2013	JSPS	Université Waseda Yves Lepage	Traduction automatique par analogie.

Les deux projets les plus anciens sont des demandes de bourses auprès de la *Japan Society for the Promotion of Science* (JSPS), et de la région Rhône-Alpes (Coopération et Mobilité Internationale − CMIRA), que je n'ai malheureusement pas obtenues, mais qui témoignent d'une première expérience dans la recherche de financements.

Le projet « COP 21 » (2016) est une demande de financement modeste (2k€), initiée par Caroline Rossi, à laquelle j'ai fortement contribué (formulation des besoins, définition des objectifs, de la méthodologie, estimation et chiffrage des besoins). Cette demande de financement s'est vue couronnée de succès.

Le projet financé par le Labex ASLAN (2016) a été monté avec Denis Vigier afin de prolonger le projet Presto, en lien avec les objectifs du Labex et de l'équipe ICAR. Il s'agit d'un financement sur 2 ans (décembre 2016 - décembre 2018), sur lequel je suis actuellement salarié.

Responsabilités collectives

Comités de lecture

J'ai fait partie des comité de lecture de **RECITAL 2012**, **COLING 2012** (sur la thématique « *Word sense disambiguation* ») et **FDHN**[8] **2013**.

Comité d'organisation

J'ai été membre du comité d'organisation de **CLPS**[9] **2016**. Je me suis chargé du site Web, de l'inscription et de de l'accueil des participants, et de la présidence d'une session.

Contribution au Consortium Corpus Écrits et CORLI

Essentiellement sur la période 2011-2015, j'ai participé au consortium « corpus écrits[10] » (projet national, porté par l'ILF[11]) en particulier sur les groupes de travail[12] « exploration de corpus » et « nouvelles formes de communication écrite ». J'ai notamment contribué activement au corpus Comere[13] et dispensé une formation en linguistique de corpus outillée.

Depuis 2016, je participe au consortium HumanNum CORLI (Corpus, Langues et Interactions), qui succède au Consortium Corpus Écrits.

Organisation de séminaires

En 2007 et 2008, je me suis chargé de l'organisation des « journées des thésards » de l'équipe, qui voient les doctorants en 2ème année de thèse présenter leurs travaux (5 doctorants en 2007, où j'étais assisté par un autre doctorant, Sopheap Seng, et 3 doctorants en 2008).

Formations

En 2009, j'ai effectué 2 séances de formation à Scientext, sur l'outil de traitement de corpus que j'avais développé, pour les membres de l'axe 1 (analyses descriptives : syntaxe, sémantique et pragmatique) du LIDILEM. En 2014, j'ai effectué une nouvelle formation à cet outil, dans le cadre du Consortium Corpus Écrits, puis en 2015, une formation à TXM dans le cadre du projet Presto.

Administration d'outils informatiques

Du fait de mes compétences en administration Linux et développement Web (« *from scratch* » et CMS SPIP et XWiki), je suis fréquement sollicité pour la création et la maintenance de serveurs et de sites Web.

Pour le projet CIFLI-SurviTra, j'ai installé, configuré et maintenu le serveur Web de démonstration (Ubuntu, Apache).

Dans le cadre du projet Scientext, j'ai créé, déployé et administré le site Web du projet (en SPIP, toujours en service[14]), et installé 2 serveurs d'échange de fichiers, d'analyse et de test (Ubuntu, Apache, Vsftpd) du système développé. J'ai ensuite maintenu ces serveurs pendant toute la durée officielle du projet, puis au-delà tant qu'ils pouvaient s'avérer utiles. Lorsque le système développé fut suffisamment abouti, j'ai assuré son déploiement sur une machine virtuelle (Debian, Apache) à la Maison des Sciences Humaines (MSH) des Alpes. En 2012, j'ai assuré la migration du système vers un serveur dédié (à nouveau Debian et Apache) mis à disposition par la MSH.

Pour le projet OMNIA, j'ai installé, configuré et maintenu le principal serveur de calcul et le serveur de démonstration du projet (CentOS et Debian, Apache et Tomcat, Subversion).

8 Fouilles de données et humanités numériques
9 Changements linguistiques et phénomènes sociétaux
10 http://corpusecrits.corpus-ir.fr
11 Institut de Linguistique Française (CNRS, FR 2393)
12 http://corpusecrits.corpus-ir.fr/travaux-2/
13 http://hdl.handle.net/11403/comere
14 http://scientext.msh-alpes.fr

Pour EMOLEX, j'ai créé, déployé et administré le site Web du projet[15] (en SPIP).

Pour le projet Traouiero, j'ai effectué le déploiement et la maintenance des applications du projet sur un serveur d'applications Web (Debian, Apache), que j'ai co-administré avec Mathieu Mangeot (GETALP).

Dans l'équipe GETALP, j'ai administré, de 2011 à 2013, 3 serveurs de stockages (NAS).

Pour le projet Presto (2014-2017), j'administre le site Web du projet, et j'apporte mon aide à l'utilisation et au développement des outils TXM et Primestat.

Enfin, tout au long de ma carrière, j'ai contribué de manière importante à la réalisation d'outils d'exploitation de corpus, et à la conception de nombreux corpus. Ces contributions sont détaillées à la fin de la partie « Détail des travaux ».

Engagement associatif

Je suis membre depuis 2005 d'une association de reconstitution médiévale, comptant une cinquantaine de membres : Excalibur-Dauphiné. J'ai conçu le site Web[16] de l'association la même année avec l'aide d'un graphiste, et j'administre depuis ce site (SPIP) ainsi que le forum (phpBB) de l'association (~100 visiteurs uniques/jour), sur le plan informatique (nom de domaine, hébergement, mises à jour) et communautaire (gestion de la communauté des membres, rédacteurs et modérateurs). J'ai rédigé des documentations pour les rédacteurs, effectué des formations, et je prends part aux réunions de l'équipe de rédaction/modération (~1 par an).

J'ai en outre assuré les fonctions de secrétaire de l'association pendant 2 ans, et d'intendant (gestion du matériel) pendant 1 an.

15 http://www.emolex.eu
16 http://excalibur-dauphine.org

Détail des travaux

Plateformes Web en production

2017 2013	Auteur principal	Dicorpus : plateforme dictionairique en ligne pour la didactique des langues, basée sur ScienQuest. http://corpora.aiakide.net/dicorpus
2017 2009	Auteur principal	ScienQuest : plateforme en ligne pour l'étude de corpus arborés. http://corpora.aiakide.net
2005	Auteur principal	Plateforme pour la consultation du Corpus du français tchaté. http://corpora.aiakide.net/corpustchat2

Logiciels en production

2017	Auteur principal	CvsFreq : un outil simple pour calculer les fréquences absolues et relatives d'occurrences extraites de la base *Frantext*. http://corpora.aiakide.net/tools/cvsFreq

Prototypes, démonstrateurs

2017 2015	Coauteur	Presto : chaîne de traitement (normalisation, tokenisation, POS tagging, lemmatisation) pour le français moderne, classique et préclassique
2012	Auteur principal	iMAG : plateforme pour la traduction, centralisée ou communautaire, de pages Web
2011	Coauteur (opérationalisation)	SECTra : gestionnaire de mémoire de traduction, avec fonctions de post-édition et d'évaluation
2011	Coauteur (opérationalisation)	Tradoh : méta-moteur de traduction avec API unifiée pour l'appel de plusieurs de moteurs de TA
2011	Coauteur (opérationalisation)	Sandoh : système de reconnaissance d'encodages et de langues pour documents hétérogènes
2011	Coauteur (opérationalisation)	SegDoc : segmenteur de pages Web
2010	Auteur principal	OMNIA : chaîne de traitement pour l'indexation sémantique interlingue de documents multilingues
2009	Auteur principal	Koinè : assistant pour le tchat en langue seconde

Principaux corpus

Sauf mention contraire (corpus collectés en 2017), il s'agit de corpus libres.

2017	Auteur principal	Corpus diachronique issu du journal *Le Monde*, formé à partir des premiers paragraphes de 10 % des articles, diffusés gratuitement par le journal sur le Web, pour la période 1950-2015 (16,5 millions de mots). Ce corpus, collecté sous le régime de la copie privée, n'est pas redistribuable.
2017	Auteur principal	Corpus de commentaires touristiques *Tripadvisor*. Corpus collecté à partir du célèbre site d'évaluation touristique. Ce corpus, collecté sous le régime de la copie privée, n'est pas redistribuable.

2017	Auteur principal	Corpus d'offices de tourisme. Corpus collecté à partir de 772 sites Web d'offices de tourisme français (3,6 millions de mots). Ce corpus, collecté sous le régime de la copie privée, n'est pas redistribuable.
2017	Auteur principal	Corpus de guides touristiques, collectés à partir de la base *Wikitravel* (5,3 millions de mots).
2016	Coauteur (traitement)	Presto : corpus du français écrit en diachronie (XVIe-XXe siècle, 28 millions de mots, XML TEI). http://presto.ens-lyon.fr
2009 màj 2015	Coauteur (traitement)	Scientext : corpus d'écrits scientifiques (anglais et français, 41 millions de mots, XML TEI P5). http://scientext.msh-alpes.fr
2008	Coauteur (traitement)	CIFLI-Survitra : corpus de phrases « à trous » pour le tourisme, aligné en 5 langues + 1 langage formel.
2005 màj 2014	Auteur principal	Corpus du français tchaté (23 millions de mots, XML) – corpus passé en XML TEI CMC en 2014. http://corpora.aiakide.net/corpustchat2 https://www.ortolang.fr/#/market/corpora/comere

Chapitres d'ouvrages (recueils scientifiques)

2014, **Achille Falaise**. « Exploitation linguistique de corpus arborés d'écrits scientifiques à l'aide du logiciel ScienQuest », Agnès Tutin & Francis Grossmann (éd) *Autour du corpus Scientext : de la constitution d'un corpus d'écrits scientifiques à l'étude des marques du positionnement et du raisonnement*, Presses Universitaires de Rennes, 20 pages.

Revues internationales avec comité de sélection

2017, Sascha Diwersy, **Achille Falaise**, Marie-Hélène Lay, Gilles Souvay. Ressources et méthodes pour l'analyse diachronique, *Langages*, ISSN 1958-9549, numéro thématique : Prépositions françaises en diachronie.

2017, Danielle Leeman, **Achille Falaise**. Sur l'identité de la préposition *en* combinée avec des noms de lieu, *Langages*, ISSN 1958-9549, numéro thématique : Prépositions françaises en diachronie.

2011, **Achille Falaise**, Agnès Tutin, Olivier Kraif. Définition et conception d'une interface pour l'exploitation de corpus arborés pour non-informaticiens : la plateforme ScienQuest du projet Scientext, *Traitement Automatique des Langues*, ISSN 1965-0906, volume 52, numéro 3, *Ressources linguistiques libres*, 25 pages.

2003, M. Faraco, M.-L. Barbier, **A. Falaise**, S. Branca Rosoff. « Codage et traitement automatique de corpus pour l'étude de prises de notes en français langue première et langue seconde », revue *Arob@se* n°7, 21 pages.

Actes de conférences internationales avec comité de lecture

2017, Peter Blumenthal, Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, Gilles Souvay, Denis Vigier. Presto, un corpus diachronique pour le français des XVI^{e}-XX^{e} siècles, atelier « Les corpus annotés du français », *Actes de TALN 2017*, juin 2017, Orléans.

2017, Sascha Diwersy, **Achille Falaise**, Denis Vigier. *Étude de l'évolution sémantique des prépositions* à, en, dans, dedans *du français. Quel(s) apport(s) d'une périodisation automatique ?* 9èmes Journées Internationales de la Linguistique de corpus, juillet 2017, Grenoble.

2016, **Achille Falaise**, Danielle Leeman. « Prépositions et noms de régions anciennes : évolution des emplois et représentations socio-culturelles », *Actes de CLPS 2016*, Lyon. [article accepté dans les actes, à paraître]

2015, Sascha Diwersy, **Achille Falaise**, Marie-Hélène Lay, Gilles Souvay. « Traitements pour l'analyse du français préclassique », *Actes de TALN 2015*, Caen.

2015, **Achille Falaise**. « Intégration du corpus des actes de TALN à la plateforme ScienQuest », Actes de TALN 2015, Caen, démonstration, 2 pages.

2014, C. Frérot, C. Rossi, **A. Falaise**. « Integrating selected corpus data in the classroom: a case-study of English NPs for French students in specialized translation », *6th International Conference on Corpus Linguistics (CILC)*, Las Palmas de Gran Canaria, Espagne.

2013, M.-P. Jacques, L. Hartwell, **A. Falaise**. « TAL et linguistique de corpus pour aider la rédaction scientifique en anglais », *actes de TALN 2013*, Les Sables d'Olonne, 12 pages.

2013, **A. Falaise**. « Adaptation de la plateforme corporale ScienQuest pour l'aide à la rédaction en langue seconde », *actes de TALN 2013*, Les Sables d'Olonne, démonstration, 2 pages.

2012, **Achille Falaise**, Agnès Tutin, Olivier Kraif, David Rouquet. ScienQuest: a treebank exploitation tool for non NLP-specialists, demo paper, *proceedings of COLING 2012*, Mumbai, Inde, 10 pages.

2011, **Achille Falaise**, Agnès Tutin, Olivier Kraif. « Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques », actes de *TALN 2011*, Montpellier, 6 pages.

2011, S. Skaff, D. Rouquet, E. Dellandrea, **A. Falaise**, V. Bellynck, H. Blanchon, C. Boitet, D. Schwab, L. Chen, A. Saidi, G. Csurka, L. Marchesotti. « Multilingual search for graphic designers », *IVAPP 2011*, Algarve, Portugal.

2010, **A. Falaise**, D. Rouquet, D. Schwab, H. Blanchon, C. Boitet. « Ontology driven content extraction using interlingual annotation of texts in the OMNIA project », *workshop CLIA, COLING 2010*, Beijing, Chine, 9 pages.

2010, **Achille Falaise**, Agnès Tutin. « Approche onomasiologique de la phraséologie transdisciplinaire des écrits scientifiques : la recherche sémantique dans les textes dans le cadre du projet Scientext », démonstration, *Actes de TOTh 2010,* Annecy, 6 pages.

2009, G. Fafiotte, **A. Falaise**, J. Goulian. « CIFLI-SurviTra, deux facettes: démonstrateur de composants de TA fondée sur UNL, et *phrasebook* multilingue », démonstration, *Actes TALN 2009*, Senlis, 3 pages.

2007, C. Boitet, P. Bhattacharyya, E. Blanc, S. Meena, S. Boudhh, G. Fafiotte, **A. Falaise**, V. Vacchani. « Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction », *proceedings of SNLP 2007*, Pattaya (Thaïlande), 6 pages.

2005, **A. Falaise**. « Constitution d'un corpus de français tchaté », *actes de TALN et RÉCITAL 2005*, tome 1, pp. 615-624, Dourdan, 9 pages.

Communications internationales sans actes

2016, Thi Thu Hoai Tran, **Achille Falaise**. *Quelles stratégies pédagogiques pour une introduction des corpus en classe de langues ?* Langues sur objectifs spécifiques : perspectives croisées entre linguistique et didactique, novembre 2016, Grenoble.

2016, Thi Thu Hoai Tran, **Achille Falaise**. *Ergonomie des dictionnaires pour l'aide à la rédaction, état de l'art et propositions*, Journée « Work in progress » autour des discours scientifiques, Juin 2016, Grenoble.

2016, Sascha Diwersy, **Achille Falaise**, Gilles Souvay. *Traitement automatique du français en diachronie, retour d'expérience sur le projet Presto*, Conférence Changements linguistiques et phénomènes sociétaux, Lyon.

2015, S. Diwersy, **A. Falaise**. *Annotation du corpus PRESTO : création de ressources pour l'analyse du français classique.* Journées PRESTO, Cologne, Allemagne.

2014, V. Gossens, **A. Falaise**. *Le projet PRESTO : corpus et traitements*. Journée d'études Linguistique de corpus outillée : regards et expériences croisés, Neuchâtel, Suisse.

2014, A. Tutin, **A. Falaise**. « Expressions polylexicales dans le discours scientifique: une base de données lexicales basée sur corpus », *Europhras 2014*, Paris.

2013, A. Tutin, **A. Falaise**. « Multiword expressions in scientific discourse: a corpus-driven database », *eLex 2013*, Tallinn, Estonie.

2013, **A. Falaise**. « Written corpora for human beings », *3rd CAMELEON Workshop*, 20 décembre 2012, Porto-Alegre, Brésil.

Séminaires invités

2015, **A. Falaise**. *Annotation du français classique et pré-classique*, Séminaire Frantext, Nancy.

2014, F. Grossmann, **A. Falaise**. *Scientext : un corpus pour analyser l'écrit scientifique*, Séminaire Recherches linguistiques et corpus, STIH de l'Université Paris-Sorbonne.

Actes de conférences nationales avec comité de lecture

2011, David Rouquet, **Achille Falaise**. « Extraction d'information conceptuelle de textes, basée sur une annotation interlingue et guidée par une ontologie », atelier RISE 2011, conférence CORIA 2011, 15 pages.

2010, David Rouquet, **Achille Falaise**, Didier Schwab, Hervé Blanchon, Vallérie Bellynck, Christian Boitet, Emmanuel Dellandréa, Ningning Liu, Liming Chen, Alexandre Saidi, Sandra Skaff, Luca Marchesotti, Gabriela Csurka. « Classification multilingue et multimédia pour la recherche d'images dans le projet OMNIA », actes de l'*atelier MIRO:RISE, conférence INFORSID 2010*.

Séminaires et autres interventions

2017, **A. Falaise**. *Resources Creation for an Under-Resourced Language: Classical French*. Séminaire 40ans2ta, Grenoble, 1er juillet 2017.

2017, **A. Falaise**. *The Presto Project : Building a Treebank and Tools to Investigate Prepositions Evolution in French (16th-20th Centuries)*. Collegium de Lyon, 31 janvier 2017.

2015, **A. Falaise**, D. Vigier. *Complexités d'un corpus en diachronie du français. Le cas de Presto.* 4 juin 2015, Atelier « Cellule Corpus Complexes », Lyon.

2012, **A. Falaise**. Web site translation with interactive Multilingal Access Gateways, 2ème Atelier LICIA, 5-6-7 Septembre 2012, Grenoble.

2011, **A. Falaise**. *Integrating software and lingware in the OMNIA and AXiMAG projects, following a KISS approach*. 26 avril 2011, Universiti Malaysia Sarawak, Kuching, Malaisie.

2010, **A. Falaise.** *Démonstration du logiciel Scientext*. Journée d'études Scientext, Maison des langues, Grenoble.

2010, **A. Falaise**. *Démonstration du logiciel Scientext*. Inauguration du site d'écrits scientifiques Scientext. 19 mai 2010, Maison des Sciences Humaines des Alpes, Grenoble.

2008, **A. Falaise**. *Vers une interface « simple » pour la recherche lexico-syntaxique dans les corpus en ligne*. Journées d'étude : positionnement et raisonnement dans les écrits scientifiques. 19-20 juin 2008, Maison des Sciences Humaines des Alpes, Grenoble.

2008, R. Barr, **A. Falaise**, F. Grossmann, A. Tutin. *Autour du projet Scientext : État des lieux et perspectives. Journées d'étude : positionnement et raisonnement dans les écrits scientifiques*. 19-20 juin 2008, Maison des Sciences Humaines des Alpes, Grenoble.

2008, **A. Falaise**. *Pour une approche à « médiation faible »*. Table ronde du laboratoire LIG, « Conception Collaborative de Produits et de services » du thème « entreprise ouverte », février 2008, UFRIMA, Grenoble.

2007, **A. Falaise**. *Aide au dialogue en langue étrangère*, journée des doctorants du GETALP, Grenoble.

2006, **A. Falaise**. *Le projet Tradphone: un softphone au secours des locuteurs non-natifs aux prises avec des difficultés de communication en langue étrangère*. Affiche pour l'école d'été Dialogue et Interaction, Autrans.

2006, **A. Falaise**. *Le traitement de la langue au service de la téléconversation en langue étrangère*. Affiche pour l'école de printemps Traitement des Connaissances, Apprentissage et NTIC, Évian.

Rapports scientifiques et autres publications

2012, **Achille Falaise**, Valérie Bellynck, Christian Boitet, Lingxiao Wang. *Avancement de l'implémentation de SECTra-v3 et évolution d'outils associés (TRADOH, SANDOH), livrable ANR du projet Traouiero*, 38 pages.

2012, **Achille Falaise**, Ruslan Kalitvianski. Étude de la segmentation de documents et première version de SegDoc liée à SECTra-v3, *livrable ANR du projet Traouiero*, 21 pages.

2011, **A. Falaise**, V. Bellynck, C. Boitet. Consolidation de SECTra_w et rétro-ingénierie, *livrable ANR du projet Traouiero*, 8 pages.

2011, **A. Falaise**, V. Bellynck, C. Boitet. État des lieux de TRADOH, *livrable ANR du projet Traouiero*, 6 pages.

2010, David Rouquet, **Achille Falaise**, Didier Schwab, Christian Boitet, Valérie Bellynck. Extraction de contenu guidée par une ontologie dans des textes annotés par des lexèmes interlingues (Lien entre descripteurs interlingues et descripteurs d'une hiérarchie de classification des images), *livrable ANR du projet Traouiero*, 17 pages.

2010, **Achille Falaise**, Olivier Kraif, Agnès Tutin. *Syntaxe pour les grammaires en mode avancé dans Scientext*, disponible sur le site du projet, 5 pages.

2010, **Achille Falaise**, Agnès Tutin. *Mode d'emploi de l'interface guidée Scientext*, disponible sur le site du projet, 5 pages.

2008, **Achille Falaise**. *Spécifications d'une structure de données "pivot" pour CIFLI-SurviTra*, rapport interne LIG-GETALP, juillet 2008, 4 pages.

2006, **Achille Falaise**. *Les principaux systèmes de reconnaissance vocale*, rapport interne LIG-GETALP, 5 pages.

2006, **Achille Falaise**. *Traduction automatique en téléphonie*, rapport interne LIG-GETALP, février 2006, 10 pages.

2005, **Achille Falaise**. *Le CLIPS et la TA de parole*, rapport interne LIG-GETALP, mai 2005, 18 pages.

Contacts

Travaux actuels

Sascha Diwersy (MCF Sciences du langage, Université Montpellier-3, Praxiling), projets Émolex et Presto.
sascha.diwersy@univ-montp3.fr

Justine Lascar (IGE CNRS, ICAR, membre de la Cellule Corpus Complexes du laboratoire ICAR).
justine.lascar@ens-lyon.fr

Matthieu Quignard (CR CNRS, ICAR, membre de la Cellule Corpus Complexes du laboratoire ICAR).
matthieu.quignard@ens-lyon.fr

Daniel Valero (ASI CNRS, ICAR, responsable de la Cellule Corpus Complexes du laboratoire ICAR).
daniel.valero@ens-lyon.fr

Denis Vigier (MCF Sciences du langage, Université Lyon-2, ICAR), projet Presto.
Denis.Vigier@ens-lyon.fr

Travaux passés

Christian Boitet (Pr. Informatique, Université Joseph Fourier, Grenoble, LIG-GETALP), directeur de thèse et projets OMNIA et Traouiero.
Christian.Boitet@imag.fr

Olivier Kraif (MCF Informatique, Université Grenoble-Alpes, LIDILEM), projets Scientext et Émolex.
olivier.kraif@univ-grenoble-alpes.fr

Agnès Tutin (MCF HDR Sciences du langage, Université Grenoble-Alpes, LIDILEM), projet Scientext, Émolex et Termith.
agnes.tutin@univ-grenoble-alpes.fr