# Curriculum Vitæ Étendu

François Yvon
LISN, CNRS & Univ. Paris Saclay

Novembre 2022

## Table des matières

# 1 Curriculum Vitæ court

**Expérience professionnelle**

| | |
|---|---|
| 2013- | Directeur de recherche au CNRS, chercheur au LIMSI (LISN depuis janvier 2021) dans le groupe "Traitement du Langage Parlé". Entre juin 2013 et décembre 2019, j'ai été directeur du LIMSI (UPR CNRS 3251). |
| 2007-2013 | Professeur au département d'informatique de l'UFR des Sciences de l'Université Paris-Sud, à Orsay ; chercheur au LIMSI/CNRS |
| 1996-2007 | Maîitre de Conférences au département informatique et réseaux de l'École Nationale Supérieure des Télécommunications (ENST / Télécom Paris), Paris. |
| 2001-2002 | Séjour sabbatique au centre de recherche d'IBM (Yorktown Heights, NY). |
| 1992-1996 | Ingénieur de recherche, ARECOM. Responsable de la participation de l'ENST au projet Européen ONOMASTICA. |
| 1989-1992 | Chef de Projet, Systèmes d'information. Centre de recherche européen des laboratoires Wyeth, Paris. |

**Titres et Diplômes**

| | |
|---|---|
| 2006 | Habilitation à diriger des recherches de l'Université Pierre et Marie Curie : *Des apprentis pour le traitement automatique des langues* (soutenue le 28/11/2006). |
| 1996 | Thèse de doctorat de l'ENST (spécialité Informatique et Réseaux) : *Apprentissage par analogie* (soutenue le 14/05/1996). |
| 1992 | Diplôme d'études Approfondies (Mathématiques discrètes) - Université René Descartes (Paris 5). |
| 1989 | Diplôme d'ingénieur - École Nationale Supérieure de Statistique et d'Administration Economique (ENSAE) |
| 1987 | Diplôme d'ingénieur, École Polytechnique |

# 2 Synthèse de la carrière

La première partie de ma carrière s'est déroulée au sein de l'ENST Paris (aujourd'hui Télécom ParisTech), que j'ai rejoint en 1992, pour y entamer une thèse de doctorat sous la direction d'Alain Bonnet. J'avais auparavant travaillé trois ans comme chef de projets dans l'industrie pharmaceutique. Au terme de ma thèse, j ai été recruté comme Maitre de Conférences à l'ENST, et je suis demeuré dans ces fonctions sans interruption jusqu'en septembre 2007. Seule parenthèse, j'ai eu l'occasion de prendre une année sabbatique, que j'ai passé à New York en tant que visiteur scientifique du laboratoires d'IBM T.J. Watson, à Yorktown Heights (en 2001-2002). Pendant ces quelques 10 années à l'ENST, j'ai porté (avec Jean-Louis Dessalles) les activités d'enseignement et de recherche en traitement automatique des langues, entouré d'une petite équipe de doctorants et post-doctorants. J'ai également assuré, au fil de l'eau, d'autres enseignements en intelligence artificielle, ainsi que des enseignements plus fondamentaux en informatique (optimisation, théorie des langages). Ma recherche a porté essentiellement sur l'application de méthodes d'apprentissage probabiliste à des problèmes de traitement automatique des langues, de traitement automatique de la parole et de fouille de textes. J'ai également contribué au développement d'une formalisation générale du cadre d'apprentissage par analogie. Durant ces années à l'ENST, j'ai encadré ou co-encadré 9 thèses. J'ai enfin assuré un certain nombre de responsabilités collectives au sein de l'ENST Paris et du GET/Institut Télécom, en tant qu'élu des personnels dans plusieurs conseils et comités statutaires de l'établissement (voir détail section 6.5).

En 2007, j'ai saisi l'opportunité qu'offrait la création d'un poste de professeur à l'Université Paris-Sud sur les thématiques de la traduction automatique pour rejoindre le LIMSI/CNRS. J'anime donc depuis septembre 2007, au sein du groupe « Traitement de la Langue Parlée », les travaux en traduction automatique au LIMSI (puis au LISN). Cette activité a fédéré au cours des années un groupe d'environ une quinzaine de personnes (environ 5 permanents et un peu plus de non-permanents) et touche tous les aspects de la traduction automatique : de l'alignement des phrases et des mots à la l'évaluation, en passant par les modèles de traduction et les algorithmes d'apprentissage et de décodage pour la traduction de parole et de textes écrits. Un fait particulièrement marquant de cette période a été le développement du premier système

de traduction entièrement basé sur des modèles neuronaux, qui s'est montré le plus performant des systèmes participant au challenge organisé durant la conférence WMT'2012. Cette activité s'est depuis 2015 élargie aux méthodes d'apprentissage cross-lingues et plus généralement au traitement des langues multilingue, avec des applications pour la linguistique de terrain. Durant cette période, mon activité d'enseignement était répartie entre l'UFR des Sciences et l'école d'ingénieur de l'Université (l'IFIPS, depuis devenue PolyTech Paris-Sud). En tant que professeur, j'ai exercé enfin diverses responsabilités administratives et pédagogiques au sein du département d'enseignement de Paris-Sud (voir détail section 6.2).

Entre juillet 2013 et décembre 2019, j'ai assumé les fonctions de directeur du LIMSI, UPR 3251 du CNRS, en étant détaché au CNRS sur un poste de Directeur de Recherche. Durant cette période, j'ai eu la responsabilité de la gestion quotidienne de l'activité d'un laboratoire accueillant entre 200 et 250 personnels (permanents et non-permanents) dans trois bâtiments d'une surface totale d'environ 800 m$^2$ et gérant un budget consolidé de plus de 15 ME. J'ai impulsé ou accompagné un certain nombre de changements dans l'organisation des cellules administratives et techniques du laboratoire, pour aller dans le sens d'une meilleure mutualisation des activités de soutien et d'une meilleure utilisation des outils informatiques. J'ai également conduit à son terme un important programme d'investiment au sein du laboratoire, tout en prenant des responsabilités importantes dans les structures d'animation de la recherche (RTRA Digiteo, LabEx Digicosme, Département STIC) qui se sont progressivement mises en place dans le cadre du projet d'Université Paris-Saclay. J'ai en particulier contribué à mettre en place le Comité d'Ethique de la Recherche de l'Université (à partir de 2017), et à déployer une grappe de calcul mutualisée entre les laboratoires d'informatique du plateau (Saclay-IA). Cette activité a été évaluée positivement par le HCERES en décembre 2018.

Depuis janvier 2020, je suis chercheur CNRS au LIMSI (devenu LISN en 2021). Mes activités de recherche en traduction automatique s'intéressent plus particulièrement aux questions qui se posent dans un cadre industriel : traduction multidomaine, traduction avec contraintes terminologiques et mémoires de traduction, en collaboration en particulier avec la société Systran ; je poursuis en parallèle mes travaux sur le transfert cross-lingue pour améliorer le traitement automatique des langues peu dotées.

À l'extérieur du laboratoire, je me suis principalement investi dans le milieu du traitement automatique des langues et de la traduction automatique. Cette implication est essentiellement de nature scientifique et a pris en premier lieu la forme d'une implication dans les conférences et les revues scientifiques de ce domaine. En second lieu, je me suis également impliqué dans les réseaux européens de coordination des activités en traitement automatique des langues : je suis ainsi membre du Conseil d'Administration du réseau META-NET [1], point de contact technique pour la France dans le cadre du projet CEF-AT ELRC [2] (*European Language Resource Coordination*) et au sein du projet *European Language Grid* [3] ; j'ai enfin été responsable de l'expertise sur le français au sein du projet *European Language Equality* [4]. Depuis 2020, je suis membre du bureau la section européenne de *l'Association for Computational Linguistics*, principale société savante dans le domaine du traitement des langues. Je suis également membre du Conseil Scientifique du GDR CNRS « Traitement Automatique des Langues », enfin membre du bureau exécutif du pôle de Compétitivité francilien Cap Digital depuis 2019, après avoir été membre du Conseil d'Administration de 2016 à 2018.

# 3 Production scientifique récente

## 3.1 Développement des activités de recherche (2018-2022)

Depuis 2007, ma recherche se développe autour de la traduction automatique (TA) et du traitement des langues multilingue. Parmi les contributions remarquables de l'équipe que j'anime, mentionons en particulier des travaux pionniers sur la traduction neuronale [11] et sur le transfert cross-lingue pour l'analyse syntaxique en dépendances [14, 10, 2].

Sur la période récente, je me suis concentré principalement sur quatre axes : l'alignement de mots et des applications pour la documentation des langues en danger [5, 9, 6] ; le transfert cross-lingue pour les langues peu dotées [15, 7] ; la traduction automatique de haute qualité avec des ressources (listes de termes, mémoires de traduction) [8, 12, 13] ; les applications au sous-titrage automatique TA [3, 16]. En particulier, nous avons proposé dans [9] une méthode originale et extrêment efficace pour produire des alignements de mots de

1. http://www.meta-net.eu/
2. http://www.lr-coordination.eu/
3. http://live.european-language-grid.eu
4. https://humane-ai.net

manière entièrement non supervisée pour un très grand nombre de langues. Un autre travail important est l'étude réalisée dans [13], qui formalise le cadre de l'évaluation des systèmes de traduction multidomaine.

Durant cette période, j'ai publié 10 articles dans des revues nationales et internationales et 33 communications dans des conférences internationales.

En plus des partenariats directement liés aux activtés contractuelles décrites ci-dessous, je collabore principalement avec LMU Munich (H. Schütze) en Allemagne, et en France avec Univ. de Paris (N. Kübler, G. Wisniewski, N. Ballier) et l'UGA (L. Besacier). En 2022, j'ai également collaboré avec FBK (Trento) et Charles University (Prague) dans le cadre de l'accueil de doctorants au laboratoire pour des séjours de longue durée. Entre 2021 et 2022, j'ai contribué à la collaboration « Big Science » visant au développement d'un grand modèle de langue multilingue ouvert ; en plus du suivi pour le compte du CNRS de ce projet collaboratif, je me suis particulièrement impliqué dans l'évaluation des capacités multilingues du modèle BLOOM.

Mon principal projet actuel s'intéresse à la TA pour la rédaction et la publication scientifique (par exemple pour des textes dans domaine bio-médical, sur lequel des travaux sont en cours). Traiter des documents complets en domaine de spécialité permet d'aborder des questions encore peu traitées en TA neuronale : intégration de ressources lexicales et terminologiques, génération de longs textes cohérents, etc. C'est un sujet que je soutiens également au sein du groupe de travail 'Traduction Scientifique' mis en place par le MESRI en 2020, et qui résonne avec diverses initiatives récentes au sein du consortium ISTEX et de *l'Association for Computational Linguistics*. J'ai obtenu en 2022 un financement de l'ANR pour le projet collaboratif MaTOS (*Machine Translation for Open Science*), qui permettra de continuer à travailler sur ces thématiques en partenariat avec l'INIST, Inria, et l'Université de Paris-Cité.

### 3.2 Recherche contractuelle et partenariale

Au niveau Européen, j'ai assumé le portage pour le LISN de la participation au réseau ITN Mirror (2015-2019) ; je suis actuellement responsable de la participation du CNRS au sein du réseau Humane-AI [5], un des quatre réseaux en Intelligence Artificielle labellisé en 2019. Sur le thème du traitement des langues, je suis point de contact national français pour les projets '*European Language Resource Coordination*' (ELRC-1 et 2, dans le programme CEF/AT, depuis 2016), pour le projet '*European Language Grid*' (depuis 2019) ; enfin porteur de la participation du LISN au projet '*European Language Equality*', (2020-2021), qui s'est traduite par la rédaction d'un état de l'art sur les technologies des langues en France [1].

Au niveau national, j'ai contribué au projet BPI Rosetta (2019-2021) en tant que responsable du WP sur le sous-titrage en Français, au projet ANR Franco-Allemand *Computational Language Documentation 2020* (responsabilité de deux WP) et au projet RAPID-DGA SOULT (2020-2022, avec la société Lingua Custodia), dont j'ai été porteur pour le LIMSI-LISN.

J'ai enfin participé aux travaux de deux groupes de travail mis en place l'un par le MESRI sur le multilinguisme et la science ouverte, qui a donné lieu à un rapport en 2020 [4], et qui continue ses travaux avec la mise en place opérationnelle d'un soutien au multilinguisme par la traduction automatique ; l'autre par l'Inria dans le cadre du Plan Stratégique en IA sur les grands modèles de langue, remis en juin 2022.

### 3.3 Médiation scientifique

Je suis régulièrement sollicité pour des interviews à la radio ou dans la presse (généraliste ou professionnelle) concernant la traduction automatique ou plus généralement le traitement des langues. J'ai ainsi récemment participé aux émissions de France Culture « Autour de la question » (en 2017), « Danse avec les mots » (en 2018), « La Méthode Scientifique » (en 2018, puis de nouveau en 2022).

Je participe à des tables rondes lors de conférences scientifiques ou professionnelles en traitement des langues ou en IA – par exemple, la conférence '*France is AI*' en 2019 et 2020. J'ai également collaboré à une exposition à la Cité des Sciences en 2020 sur les langues en danger.

## 4 Activités scientifiques

5. `https://www.humane-ai.eu/`

## 4.1 L'apprentissage structuré pour le traitement des langues

Mon activité de recherche porte sur les applications en **traitement automatique des langues** de **de techniques d'apprentissage automatique**, que ces techniques reposent sur des mécanismes de nature symbolique, ou bien sur les principes de l'inférence statistique. Les motivations de ces approches empiriques (ou à base de corpus) ont fait l'objet de controverses fameuses en linguistique, ainsi que d'une épaisse littérature en linguistique informatique, qui met en avant la robustesse des approches à base de corpus, leur portabilité etc. Je ne reviens pas sur ces arguments qui sont largement discutés dans le premier chapitre de mon mémoire d'habilitation [78].

Concernant les approches symboliques, ma contribution principale porte sur les **approches à base d'analogie**, que j'ai initialement développé pour l'apprentissage de la conversion orthographique-phonétique et pour l'acquisition de la morphologie. J'ai développé dans ma thèse, en collaboration avec V. Pirrelli, un cadre d'apprentissage qui exploite non pas les similarités de surface entre objets linguistiques, mais des similarités « du second ordre », qui impliquent des relations entre quatre objets. Dans ce modèle, l'inférence repose essentiellement sur la possiblité de résoudre des équations analogiques de la forme « A est B comme C est à ? ». J'ai posé un cadre formel et proposé des algorithmes efficaces pour résoudre les équations analogiques sur des collections d'objets variées tels que des séquences ou des arbres [**Stroppa06formal**, 80, 65, 66]. Ce travail s'est poursuivi, en collaboration avec P. Langlais (U. Montréal), pour étendre ces techniques au cadre de la traduction automatique [41, 39, 40]. En parallèle, en collaboration avec V. Pirrelli [61], puis avec N. Stroppa, [67] nous avons mis en évidence la pertinence linguistique de cette démarche en particulier pour modéliser l'apprentissage de régularités morphologiques. Ces méthodes, constituent aujourd'hui des approches susceptibles d'offrir des alternatives et/ou des compléments aux modèles purement statistiques, qu'ils soient probabilistes ou neuronaux, qui dominent aujourd'hui sur la scène de la traduction automatique et plus généralement en traitement automatique des langues. La découverte et la résolution d'analogies continue de faire l'objet de développements fondamentaux en IA symbolique [62]) comme en traitement automatique des langues [51].

L'autre facette principale de mes travaux porte sur l'utilisation de **méthodes statistiques de catégorisation et de classification** (voir [69, 63]), ainsi que sur **les modèles statistiques de séquences**, utilisés aussi bien pour des tâches d'acquisition lexicale que pour des tâches d'orientation plus « syntaxique » : modèles de langue pour la reconnaissance automatique de la parole [49, 47, 48], automates et transducteurs stochastiques pour la correction automatique [79] ; modèles de Markov cachés (HMMs) et champs conditionnels aléatoires (CRFs) [64, 44, 45] pour les tâches d'étiquetage de séquences ; analyse statistique en dépendances [38, 6], segmentation automatique de documents [52] ; identification des entités nommées [53] ; etc. Mes travaux en traitement de la parole ont bénéficié d'une longue collaboration avec G. Gravier (IRISA) ; nombres de travaux sur les modèles probabilistes en TAL se sont développés en tandem avec O. Cappé. Dans ces dernières années, je me suis beaucoup intéressé aux problématiques de transfert entre domaines ou entre langues, avec des applications en étiquetage en partie du discours ou en analyse syntaxique [57, 38, 5]. Ces résultats récents de ces travaux sont détaillés ci-dessous.

En ce qui concerne les champs applicatifs, j'ai principalement travaillé sur quatre domaines :

- l'acquisition de connaissances lexicales (phonologie, conversion graphème-phonème, analyse morphologique) : mon travail de doctorat portait sur la conversion orthographique-phonétique ; ce travail s'est depuis déplacé vers la morphologie computationnnelle, la correction orthographique/lexicale, en particulier dans le cadre de l'analyse automatique des textes électroniques. J'ai ainsi développé un outil de « traduction » du langage SMS vers le français standard [37, 79] ;
- la gestion de collections documentaires : recherche, classement, classification, segmentation, filtrage, extraction d'information dans les archives de documents textuels ;
- le traitement de la parole (synthèse et reconnaissance vocale) : j'ai pris une part active dans le développement de deux systèmes de reconnaissance vocale (l'un dans le cadre du projet Sirocco ; l'autre, pendant mon séjour sabbatique au centre de recherche d'IBM à New York) ; j'ai également travaillé sur le développement de modules linguistiques pour la synthèse de la parole ;
- la traduction automatique par méthodes statistiques, qui est le thème que je développe depuis mon arrivée au sein du LIMSI. Nos activités en traduction statistique sont multiformes et touchent à tous les aspects de la traduction : l'alignement de phrases [76] et de mots [42, 68, 54], l'apprentissage de modèles, en particulier neuronaux [18, 46, 4], ou encore la question des mesures de confiance [71] et de l'évaluation [14]. Ils sont également développés ci-dessous.

Cet intérêt pour les approches à base de données en traitement des langues s'accompagne, presque natu-

rellement, de préoccupations relatives à la mesure des performances des systèmes : j'ai ainsi pris part à l'organisation et participé à plusieurs campagnes d'évaluation nationales de systèmes de traitement de la langue et de la parole, ainsi que plus récemment, de systèmes de traduction automatique (par exemple les campagnes internationales organisées dans le cadre de la série d'ateliers sur la traduction statistique [6] qui sont organisés chaque année depuis 2006.

J'ajoute enfin que ces travaux trouvent des débouchés au travers de nombreuses collaborations académiques et industrielles (voir ci-dessous la section 4.5).

## 4.2 Les progrès de la traduction automatique neuronale (2014-2022)

## 4.3 Systèmes de Traduction Automatique

L'amélioration des systèmes et des modèles de traduction automatique (TA) est restée au centre de mes préoccupations de la période et j'ai continué de contribuer activement au développement d'architectures computationnelle pour la TA, en explorant deux pistes principales. La première correspond à l'étude de systèmes *plus interactifs et plus réactifs*, capables de proposer de nouvelles formes de collaboration humain-machine en TA. Dans le cadre de la thèse Li Gong, nous avons étudié des systèmes statistiques capables de traduire à la volée sans aucun apprentissage préalable. Ces systèmes exploitent des techniques d'échantillonnage qui interrogent des structures de données efficaces (arbres de préfixes) pour réaliser des estimations *locales et adaptés* des paramètres de modèles statistiques [27] ; en prenant en compte les corrections des utilisateurs, ils s'adaptent et s'améliorent de manière continuelle [28].

Avec Julia Ive, nous nous sommes intéressés à l'identification et à la correction *ex ante*, plutôt qu'*ex post*, des difficultés de traduction d'un document. Nous avons ainsi montré, en travaillant sur les textes de spécialité en domaine médical [30, 32], qu'il était possible d'identifier avec une bonne confiance ces difficultés, et que leur correction préalable était une alternative réaliste et potentiellement plus efficace qu'une post-édition [31].

Le travail de Jitao Xu s'est concentré autour de la question de l'écriture bilingue, visant à concevoir de nouvelles architectures permettant de traiter des entrées mélangeant fragments en langues source et cible (par exemple correspondant à une prétraduction, à un segment issu d'une mémoire de traduction ou encore d'un état antérieur du texte cible) [74, 73]. Ces travaux ouvrent eux-aussi des perspectives nouvelles pour concevoir de nouveaux systèmes de traduction plus interactifs, et plus facilement capables de prendre en charge les besoins d'utilisateurs (partiellement) bilingues.

Le second axe de mes recherches a été la poursuite des travaux sur les architectures neuronales pour la TA, qui ont, au cours de la période, radicalement transformé l'état de l'art en TA à base de corpus et pratiquement entièrement éliminé du paysage les méthodes antérieures (TA « à base de segments »). Un premier résultat marquant, quoiqu'un peu tardif, a été la définition de nouvelles fonctions objectif pour la TA neuronale ; obtenu dans le cadre de la thèse Quoc-Khanh Do, ce résultat montrait la possibilité d'apprendre de manière complètement discriminante tous les paramètres d'un réseau neuronal, en utilisation des techniques empruntées à l'apprentissage de fonctions de réordonnancement [19, 4]. J'ai également été conduit, dans le cadre du projet H2020 QT21 [7] à m'intéresser à la traduction de langues dites morphologiquement riches : plusieurs contributions ont ainsi été réalisées avec Franck Burlot relatives à la neutralisation automatique de traits morphologiques (en langue souce) inutiles pour la traduction [15], à la spécification d'une architecture neuronale permettant d'exploiter des traits morphologiques en langue cible [9], voire à les réintroduire ex-post [10], enfin à l'exploitation de données monolingues en traduction [17]. Un produit dérivé de cette activité a été la définition d'un nouveau cadre pour évaluer les compétences morphologiques des systèmes de TA ; initialement développé pour la traduction vers le tchèque et le letton [14], ce cadre a été généralisé au français dans [16], puis étendu à l'allemand, au turc et au finnois [13].

Récemment, la communauté de la TA s'est résolument tourné vers la conception de systèmes capables de prendre en charge de multiples langues ou domaines [22, 21]. C'est sur cette dernière question qu'a porté la thèse de M.-Q. Pham, qui a en particulier permis de mieux formaliser la question de l'aprentissage multidomaine [59], ainsi que de proposer diverses évolutions des architectures Transformer de l'état de l'art pour les rendre robustes et efficaces en présence de données hétérogènes [58, 60].

6. `http://www.statmt.org`
7. `http://www.qt21.eu/`

Les travaux en TA sont valorisés à travers des collaborations suivies avec les entreprises du secteur (Reverso, Systran, Lingua Custodia, voir la section 4.5), par des participations régulières à des campagnes d'évaluation pour la traduction de texte (campagnes d'évaluation WMT [20, 56, 50, 3, 12, 1]) ou, plus épisodiquement, pour la traduction de parole (campagnes d'évaluation IWSLT [11]), ainsi que par la production de donnée linguistiques [77, 33].

**Autour de la Traduction** Si la TA fournit une application exemplaire pour étudier des tâches complexes d'apprentissage supervisé pour des données structurées, un problème très voisin, celui de l'alignement de bitextes, nous a fourni un cadre pour étudier les méthodes d'apprentissage non-supervisé. Le projet ANR/TransRead [8], que j'ai coordonné, a ainsi mis en valeur l'intérêt de calculer des alignements de phrases [75] et de mots [76] très sûrs y compris pour des bi-textes difficiles pour des tâches de lecture bilingue (thèse de Y. Xu), et permis la réalisation d'un premier prototype de liseuse électronique bilingue [81]. Un autre résultat de ce projet a été la conception de nouvelles métriques pour évaluer l'alignement de mots [77].

Un autre contexte est celui du projet ANR-DFG/BULB [9], qui nous amène à étudier des alignements entre du signal de parole (dans des langues mal documentées, voire sans système d'écriture stabilisé) et des séquences de mots (en français) [2]. Ces traitements posent en réalité un double problème : celui de découverte automatique d'unités signifiantes et celui de l'alignement de ces unités avec les mots d'une langue connue (thèse de P. Godard). Sur le premier axe, nous avons travaillé à la fois sur la question des représentations optimales pour la segmentation automatique [23], sur l'utilisation d'informations tonales pour la segmentation en mots [25], enfin sur l'utilisation de grammaires stochastiques [10] exploitant des connaissances expertes, dont nous avons montré l'utilité et grâce auxquelles nous avons pu considérablement améliorer l'état de l'art de la segmentation automatique [24]. Sur le second axe, nous avons principalement essayé de détourner pour l'alignement des méthodes de traduction neuronales [26], avec des résultats encore insuffisants par rapport aux méthodes monolingues. Ce travail s'est poursuivi avec la thèse de S. Okabé, financée par un projet qui prolonge BULB : les premiers résultats issus de ce travail portent sur l'exploitation de ressources auxiliaires (dictionnaires, grammaires) pour apporter une supervisation faible pour des tâches de segmentation en mots et morphèmes [55].

## 4.4 Contributions à l'apprentissage de modèles structurés (cross-lingues)

Mes travaux en apprentissage automatique s'intéressent exclusivement à des problèmes d'apprentissage structurés correspondant à des traitement » de bas niveau » des textes et documents (segmentation, normalisation et correction orthographique, étiquetage en parties du discours ou en chunks, segmentation morphologique, parsing), avec comme ambition de développer des méthodes, par exemple en matière d'adaptation au domaine, qui pourront ensuite être transférées à des problèmes de TA. En plus des méthodes neuronales étudiées dans le cadre de mes travaux en TA. J'ai principalement travaillé sur trois familles de modèles : les champs markoviens globalement normalisés (thèse de N. Pécheux), les modèles bayésiens non-paramétriques pour la segmentation morphologique (déjà évoqués supra dans le cadre de la thèse de P. Godard), les méthodes d'apprentissage par imitation (thèse de E. Knyazeva). Concernant la première famille de techniques, si nous avons goblament échoué à les rendre opérants dans un contexte de TA (ce qui était l'ambition annoncée par [43]), nos efforts nous ont permis de mieux cerner les difficultés posées par ces modèles, à savoir d'une part l'introduction de données de segmentation latentes, d'autre part le problème des références non atteignables. Ces travaux ont conduit à profondément remanier le code du logiciel Wapiti, et préparé la contribution essentiellement algorithmique résumée dans [45], qui étend de diverses manières le cadre des CRFs standard (en particulier en introduisant des tests rationnels arbitraires sur les séquences d'étiquette). Ces modèles surpassent l'état de l'art en étiquetage morphosyntaxique pour plusieurs langues morphologiquement riches. Si, pour des raisons voisines, nos efforts pour appliquer l'apprentissage par imitation au cadre de la TA n'ont pas été fructueuses, ils ont débouché sur plusieurs contributions sur des tâches structurées plus simples : une démonstration de l'intérêt du décodage en ordre libre [36] ; une application originale de l'apprentissage structuré à la reconnaissance de personnages [35].

Une activité importante en traitement des langues multilingue a porté sur le transfert de connaissances ou de ressources depuis une langue bien documentée et/ou bien outillée vers une autre langue moins bien

8. `http://www.limsi.fr/TransRead`
9. `http://www.bulb-project.org`
10. Plus précisément des *adaptor grammars*.

dotée, un problème qui a pris une importance accrue ces dernières années. En plus d'applications pratiques évidentes, ce cadre permet de reposer des questions fondamentales du point de vue linguistique (l'existence d'universaux linguistiques ou du moins de propriétés partagées par de nombreuses langues) comme du point de vue statistique (l'apprentissage faiblement supervisé, l'adaptation au domaine). Dans ce contexte, nos travaux (thèse de N. Pécheux et de L. Aufrant ; projet DGA Rapid/Papyrus), ont étudié les deux principaux modes de transfert cross-lingue : transfert de modèles délexicalisés et transfert d'annotations avec des applications variées : étiquetage morpho-syntaxique, alignement de mots, analyse syntaxique en dépendances. Nos contributions les plus significatives ont porté sur le transfert de connaissances syntaxiques, où nous avons proposé diverses méthodes efficaces pour le transfert cross-lingue qui s'appuient sur des ressources linguistiques expertes relativement faciles à obtenir : dictionnaires et règles de génération morphologiques extraits de wiktionnaires [70, 57], informations typologiques telles que documentées par exemple dans le WALS (*World Atlas of Language Structures*) [5]. Une analyse plus fine des difficultés et des compromis du transfert syntaxique dans un cadre cross-lingue (mono- et multi-source) est donnée dans [8], qui rejoint dans ces préoccupations les travaux du projet BULB déjà mentionné supra. Ces recherchent se développent aujourd'hui en collaboration avec l'université Ludvig-Maximilian de Munich (H. Schütze), nos efforts portant sur l'apprentissage et l'exploitation de représentations massivement multilingues [34, 29].

Il est enfin à noter que les travaux en analyse syntaxique sur le transfert se sont accompagnées de plusieurs contributions autour de l'apprentissage d'analyseurs à transition, exploitant en particulier le concept d'oracle non-déterministe de Goldberg [] pour développer des cadres d'apprentissage non-standard tels que l'apprentissage partiel [38], apprentissage de structures discontinues [7] ou améliorer l'apprentissage pour des recherches en faisceau [6].

## 4.5 Partenariats industriels et valorisation

Depuis le début de ma thèse, financée par un contrat européen (programme LRE), j'ai conduit en parallèle avec mes activités académiques une activité de recherche plus contractuelle, en relation avec des industriels. L'activité en traduction du LIMSI est depuis 2007 financée par un certain nombre de projets nationaux et internationaux, et impliquent également des partenaires industriels privilégiés (Systran, Reverso/Softissimo, Lingua Custodia, Vocapia Research pour la traduction de parole).

Au niveau européen, mon activité récente s'est traduite par la participation active à plusieurs réseaux de recherche : le réseau Humane-AI autour de l'IA en interaction avec l'humain, au sein duquel je représente le CNRS ; le réseau Meta-NET en traitement des langues, autour duquels se sont structurés les projets *European Language Equality* et *European Language Grid*, ainsi que la poursuite d'une démarche au long cours pour coordonner la production de ressources linguistiques au sein des actions *European Language Resource Coordination*. Un projet H2020/QT21 [11] (2015-2018), auquel le LIMSI a pris part au côté des meilleurs groupes européens de traduction automatique. J'ai coordonné au LIMSI ce projet qui nous a permis d'une part de progresser sur la traduction entre langues européennes, en particulier les langues morphologiquement complexes, ainsi que de développer, en partenariat avec les autres groupes de recherche du projet, des systèmes de traduction au meilleur niveau de l'état de l'art.

Dans les années récentes j'ai participé à plusieurs projets ANR nationaux et internationaux. Mentionnons en particulier le projet ANR/TransRead, que j'ai coordonné, et qui visait à étudier le design et les fonctionnalités d'outils pour la lecture bilingue. Ce projet a été sélectionné parmis les réalisations marquantes de l'ANR en 2016, et a été présenté publiquement aux journées ANR du Numérique en 2016 ; plusieurs démonstrations ont également été présentées à l'occasion du Meta-Forum 2016 (à Lisbonne) et au Forum 2016 de la DGT (à Bruxelles). Le projet franco-Allemand BULB a donné lieu à une collaboration fructueuse avec des équipes de linguistiques de terrain, et a été valorisé par une riche production scientifique, ainsi que par la production de plusieurs corpus annotés pour des langues bantoues peu dotées. Ces initiatives se prolongent aujourd'hui dans le projet ANR-DFG *Computational Language Documentation*, qui rassemble peu ou prou les mêmes partenaires. À partir de 2023, et pour 4 ans je coordonnerai le projet MaTOS (*Machine Translation for Open Science*).

Dans un passé plus ancien, j'ai été très actif au sein du programme Quaero [12], au côté des équipes de Hermann Ney à Aix-la-Chapelle et d'Alex Waibel à Karlsruhe, ainsi que dans le réseau d'excellence T4ME [13], qui a

11. `http://www.qt21.eu/`
12. `http://www.quaero.org`
13. `t4me.dfki.de`

pour ambition de fédérer les efforts de recherche de la communauté en matière d'ingéniérie multi-lingue et de traduction automatique. T4ME a muté en META-Net [14], et j'ai été élu au bureau exécutif de cette association en 2014.

Le tableau 1 liste l'ensemble des contrats dans lesquels j'ai été impliqués, ainsi que des collaborations auxquelles ces contrats ont donné lieu. J'ai fait figurer en gras ceux dans lesquels mon implication est ou a été la plus forte (porteur pour le LIMSI-CNRS ou pour l'ENST).

| | |
|---|---|
| **2020-2021** | Projet ELE visant à élaborer une feuille de route pour aboutir à l'égalité des langues dans l'espace numérique. Collaboration pilotée par ADAPT (Irlande), impliquant une cinquantaine de partenaire européens. |
| **2020-2024** | Projet ANR/DFG CLD 2025, sur l'outillage de la linguistique de terrain pour documentater les langues en danger. Collaboration avec KIT, LIG (Grenoble), LPP (Paris 3) et Lacito (CNRS). |
| **2020-2022** | Projet SOULT, financement DGE, sur l'amélioration de la traduction neuronale par utilisation de lexiques et terminologies. Collaboration avec Lingua Custodia |
| 2018-2022 | Projet Rosetta , BPI « Grands défis du Numérique » sur la production automatique de sous-titres et la traduction en LSF. Collaboration avec Systran SA, FranceTV, MocapLab, Lutin (Paris 8) |
| **2015-2018** | Projet H2020/QT21, sur la traduction automatique de haute qualité. Collaboration avec DFKI, KIT, RWTH (GE), Uni. Edimbourg, Uni. Sheffield (UK), Uni. Amsterdam (NL), DCU (IR), FBK (IT) etc |
| **2015-2018** | Projet ANR/DFG BULB, sur l'outillage des linguistes de terrain pour la documentation de langues en danger. Collaboration avec KIT, Leibniz ZAS, Berlin (GE), LIG (Grenoble), LPP (Paris 3) et LLACAN (CNRS). |
| 2016-2020 | Projet ANR Parsiti, sur le parsing et la traduction d'enoncés bruités issus des réseaux sociaux, avec LIPN (Paris-Nord), LLF (Paris-Diderot), Inria Paris-Centre |
| **2015-2016** | Projet DGA/RAPID Papyrus, portant sur l'extraction d'informations multilingues. Collaboration avec la société Systran. |
| **2012-2016** | Projet ANR TRANSREAD, portant sur les interfaces de lecture bilingue. Collaboration avec ILJ/CEDRIC/CNAM et Reverso/Softissimo. |
| **2012-2013** | Projet DGA/RAPID Rapmat, portant sur les interfaces mobiles pour la traduction de la parole. Collaboration avec la société Vocapia. |
| **2011-2012** | *Award* décerné par Google pour développer de nouvelles méthodes d'alignement de corpus multilingues. |
| 2010-2013 | Participation aux activités du réseau d'excellence européen T4ME, notamment au premier pilier qui concerne les activités de recherche en traduction automatique |
| **2007-2012** | Participation au programme Quaero sur le thème de la traduction automatique, en collaboration avec Systran, avec l'Université d'Aix-la-Chapelle et l'Université de Karlsruhe. |
| **2009-2012** | Projet ANR TRACE, portant sur la correction d'erreur en traduction automatique. Collaboration avec Réverso/Softissimo. |
| 2009-2012 | Projet SAMAR (pôle de compétitivité « Cap Digital ») portant sur la réalisation d'une plate forme pour le traitement automatique de dépêches de presse en langue arabe. Collaboration en particulier avec l'Université de Caen. |
| **2007-2010** | Projet ANR CroTAL, portant sur les champs aléatoires conditionnels en traitement des langues. Collaboration avec l'Université d'Orléans, l'Université de Paris Nord, l'Université de Lille 3. |
| **2006–2009** | Participation au Projet Infom@gic du pôle de compétitivité « Cap Digital » (IDF), notamment dans le sous-projet 2.12 (Méthodes statistiques pour l'extraction d'information). Rédaction de deux rapports techniques. |
| 2006–2009 | Participation aux activités du Réseau d'excellence K-Space, notamment dans le sous-projet 3.3 (outils pour l'indexation de textes). Participation à la rédaction d'un état de l'art sur l'indexation de données multimedia. |

14. www.meta-net.eu

| | |
|---|---|
| **2003–2006** | Contrat de Recherche FT R&D sur la fouille de données textuelles. Participation au développement d'un logiciel implémantant des méthodes de clustering probabiliste. |
| 2005 | Contrat de Recherche Thalès TRT. Mise en place d'un système de reconnaissance automatique de la parole |
| 2005 | Contrat de Recherche FT R&D sur la fouille de traces d'interactions sur le Web. Supervision de développement logiciel, rédaction d'un rapport technique. |
| **2005** | Participation au programme CNRS « Traitement des Connaissances, Apprentissage et Nouvelles Technologies de l'Information et de la Communication », en collaboration avec l'IRISA (Rennes), le LRI (Université d'Orsay) et l'ERSS (Toulouse) sur l'analogie en traitement automatique des langues. |
| **2002-2003** | Contrat de recherche avec « Droit In Situ ». Développement d'un prototype d'un logiciel pour l'indexaction automatique d'enregistrement de conférences juridiques. |
| **2000-2001** | Action de recherche concertée INRIA, en collaboration avec l'IRISA (Rennes). Développement du logiciel (public) de reconnaissance vocale Sirocco |
| **1997-1999** | Contrat Ministère de la Recherche »Industrie des Langues », en collaboration avec le LIMSI (Orsay) , l'Institut pour Communication Parlée (Grenoble), l'Institut de Phonétique de Paris. Constitution d'un dictionnaire phonétisé pour le Français, aujourd'hui diffusé via ELRA/ELDA. |
| **1996-1997** | Contrat de recherche avec SwissCom. Développement d'un lexique phonétisé de noms propres pour des applications de reconnaissance vocale. |

TABLE 1: Activités de recherche contractuelle

## 4.6 Collaborations internationales

J'ai effectué un séjour en tant que *visiting scientist* au centre de recherche T.J. Watson d'IBM à Yorktown Heights en 2000-2001, dans l'équipe de G. Zweig. Cette collaboration a été importante pour le développement de mes activités en reconnaissance de parole dans la période 2000-2005.

Entre 2007 et 2013 j'ai collaboré avec P. Langlais (DIRO, Univ. Montréal) sur les méthodes à base d'analogies en traitement des langues ; cette collaboration a donné lieu à des séjours croisés, le dernier au printemps 2013 pour un mois invité à Montréal.

J'ai été invité à participer à un *Dagstuhl Seminar* organisé en en février 2014 par A. Fraser (LMU, Munich) sur le thème du traitement automatique des langues morphologiquement riches [15], thème qui est a été au cœur des préoccupations du projet H2020 QT21, auquel participaient également des membres des principaux centres de traduction européens (RWTH-Aachen, FBK-Trento, Univ. Edimbourg, KIT Karlsruhe, Univ. Charles à Prague, Univ. Amsterdam, DFKI-Berlin, etc).

Des collaborations avec ces différentes équipes se sont développées dans le cadre de ce projet, en particulier avec KIT (J. Niehues, S. Stüker) sur le thème des réseaux neuronaux, des réordonnancements en traduction automatique, mais également sur celui du traitement des langues peu dotées (projet BULB, puis CLD).

J'ai également été membre de l'*Advisory Board* du projet H2020 HiML (*Health in my language*) et collabore dans ce contexte avec le *Center for Information and Language Processing* de LMU Munich ainsi qu'avec le groupe de traduction de l'Université d'Edimburgh sur les problèmes de la traduction automatique dans le domaine bio-médical, qui fait l'objet en France de plusieurs collaborations avec le Centre Cochrane Français (une thèse CIFRE, ainsi que des actions croisées dans le cadre du RTN MIROR). J'ai à ce titre été invité une semaine au *Center for Advanced Studies* de Munich en mai 2016.

J'ai également participé à un second *Dagstuhl Seminar* en janvier 2017, organisé par H. Schütze (LMU Munich) et C. Dyer (DeepMind) sur le thème des représentations neuronales à base de caractères. Ce séminaire a permis d'initier une collaboration avec H. Schütze qui se développe autour des représentations multilingues.

## 4.7 Publications

Le tableau 2 présente une vision comptable de mes publications depuis le début de ma carrière et sur la période de référence. Les références complètes figurent dans la liste des travaux qui accompagne ce rapport

15. `http://www.dagstuhl.de/14061/`.

d'activité et sont accessibles sur HAL [16] ou depuis d'autres serveurs de publications.

Je n'ai pas fait figurer dans ce tableau mes maigres activités en tant qu'éditeur : elles comprennent l'édition des actes de la conférence IWSLT (en 2014). J'ai précédemment publié un livre collectif (« Méthodes Statistiques pour l'Accès à l'Information Textuelle », co-édité avec Eric Gaussier, paru en 2011 chez Hermès-Lavoisier, et dont une traduction en anglais est également parue chez Wiley en 2012) ; des actes de conférences (2 en 2010), et des numéros de la revue TAL (4 sur la période 2006-2010) et une en 2012. Je n'ai pas non plus dénombré ci-dessous les rapports techniques et les communications sans actes, qui figurent également dans la liste de travaux jointe.

| | 1995-2022 | 2018- |
|---|---|---|
| Revues internationales | 28 | 4 |
| Revues nationales | 13 | 5 |
| Chapitres de livres | 17 | 1 |
| Conférences internationales avec actes | 137 | 36 |
| Conférences nationales avec actes | 53 | 8 |

Table 2 – Bilan chiffré de l'activité de publication

Trois articles de conférence ont été primés [19] et [72] qui ont reçu le prix du meilleur article à la conférence TALN en 2014 et 2021, et [15] qui a reçu le *best paper award* de la conférence européenne en traduction automatique (EAMT) en 2017.

## Références


[1] Sadaf Abdul Rauf, José Carlos Rosales Núñez, Minh Quang Pham et François Yvon. "LIMSI @ WMT 2020". In : *Proceedings of the Fifth Conference on Machine Translation*. Online : Association for Computational Linguistics, nov. 2020, p. 801-810.

[2] Gilles Adda et al. "Breaking the Unwritten Language Barrier : The BULB Project". In : *Proceedings of the SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages*. Yogyakarta, Indonesia, 2016, p. 8-14.

[3] Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Ophélie Lacroix, Elena Knyazeva, Thomas Lavergne, Guillaume Wisniewski et François Yvon. "LIMSI@WMT16 : Machine Translation of News". In : *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, août 2016, p. 239-245.

[4] Alexandre Allauzen, Quoc Khanh Do et François Yvon. "A comparison of discriminative training criteria for continuous space translation models". In : *Machine Translation* 31.1-2 (2017), p. 19-33. issn : 1573-0573.

[5] Lauriane Aufrant, Guillaume Wisniewski et François Yvon. "Zero-resource Dependency Parsing : Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge". In : *Proceedings of the 26th International Conference on Computational Linguistics : Technical Papers*. COLING 2016. Osaka, Japan, déc. 2016, p. 119-130.

[6] Lauriane Aufrant, Guillaume Wisniewski et François Yvon. "Don't Stop Me Now! Using Global Dynamic Oracles to Correct Training Biases of Transition-Based Dependency Parsers". In : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain : Association for Computational Linguistics, 2017, p. 318-323.

[7] Lauriane Aufrant, Guillaume Wisniewski et François Yvon. "Exploiting Dynamic Oracles to train Projective Dependency Parsers on Non-Projective Trees". In : *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL. New Orleans, LO, 2018, 5p.

[8] Lauriane Aufrant, Guillaume Wisniewski et François Yvon. "Quantifying training challenges of dependency parsers". In : *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Santa Fe, New Mexico, USA : Association for Computational Linguistics, 2018, p. 3191-3202.


16. `https://cv.archives-ouvertes.fr/francois-yvon`

[9] Franck Burlot, Mercedes Garcìa-Martìnez, Loïc Barrault, Fethi Bougares et François Yvon. "Word Representations in Factored Neural Machine Translation". In : *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*. Copenhagen, Denmark : Association for Computational Linguistics, sept. 2017, p. 20-31.

[10] Franck Burlot, Elena Knyazeva, Thomas Lavergne et François Yvon. "Two-Step MT: Predicting Target Morphology". In : *Proceedings of the International Workshop on Spoken Language Translation*. IWSLT'16. Seattle, USA, 2016.

[11] Franck Burlot, Matthieu Labeau, Elena Knyazeva, Thomas Lavergne, Alexandre Allauzen et François Yvon. "LIMSI at IWSLT'16: MT Track". In : *Proceedings of the International Workshop on Spoken Language Translation*. IWSLT'16. Seattle, USA, 2016.

[12] Franck Burlot, Pooyan Safari, Matthieu Labeau, Alexandre Allauzen et François Yvon. "LIMSI at WMT'17". In : *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark : Association for Computational Linguistics, sept. 2017, p. 257-264.

[13] Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen et François Yvon. "The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English". In : *Proceedings of the Third Conference on Machine Translation*. Belgium, Brussels : Association for Computational Linguistics, oct. 2018, p. 550-564.

[14] Franck Burlot et François Yvon. "Evaluating the morphological competence of Machine Translation Systems". In : *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*. Copenhagen, Denmark : Association for Computational Linguistics, sept. 2017, p. 43-55.

[15] Franck Burlot et François Yvon. "Learning Morphological Normalization for Translation from and into Morphologically Rich Language". In : *Prague Bulletin of Mathematical Linguistics* 108 (2017), p. 49-60.

[16] Franck Burlot et François Yvon. "Évaluation morphologique pour la traduction automatique: adaptation au français". In : *Conférence sur le Traitement Automatique des Langues Naturelles*. TALN. Rennes, France : ATALA, 2018, 14 pages.

[17] Franck Burlot et François Yvon. "Using Monolingual Data in Neural Machine Translation : a Systematic Study". In : *Proceedings of the Third Conference on Machine Translation*. Belgium, Brussels : Association for Computational Linguistics, oct. 2018, p. 144-155.

[18] Josep Maria Crego et François Yvon. "Factored bilingual n-gram language models for statistical machine translation". In : *Machine Translation* (2010), p. 1-17.

[19] Quoc Khanh Do, Alexandre Allauzen et François Yvon. "A Discriminative Training Procedure for Continuous Translation Models". In : *Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon, Portugal, sept. 2015, 7p.

[20] Quoc Khanh Do, Teresa Herrmann, Jan Niehues, Alexander Allauzen, François Yvon et Alex Waibel. "The KIT-LIMSI Translation System for WMT 2014". In : *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA, juin 2014, p. 84-89.

[21] Angela Fan et al. "Beyond English-Centric Multilingual Machine Translation". In : *Journal of Machine Learning Research* 22.107 (2021), p. 1-48.

[22] Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural et Kyunghyun Cho. "Zero-Resource Translation with Multi-Lingual Neural Machine Translation". In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas : Association for Computational Linguistics, nov. 2016, p. 268-277.

[23] Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Hélène Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland et François Yvon. "Preliminary Experiments on Unsupervised Word Discovery in Mboshi". In : *Proceedings of the Annual Conference of the International Speech Communication Association*. San francisco, CA, 2016, p. 3539-3543.

[24] Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Hélène Maynard et Annie Rialland. "Adaptor Grammars for the Linguist: Word Segmentation Experiments for Very Low-Resource Languages". In : *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Brussels, Belgium, oct. 2018, p. 32-42.

[25] Pierre Godard, Kevin Löser, Alexandre Allauzen, Laurent Besacier et François Yvon. "Unsupervised Word Segmentation : does tone matter ?" In : *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. CICLING 2018. Hanoï - VN, 2018, 11p.

[26] Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Bérard, François Yvon, Aline Villavicencio et Laurent Besacier. "Unsupervised Word Segmentation from Speech with Attention". In : *Proceedings of the International*. InterSpeech 2018. Hyderabad, India, sept. 2018.

[27] Li Gong, Aurélien Max et François Yvon. "Improving bilingual sub-sentential alignment by sampling-based transpotting". In : Proceedings of the International Workshop on Spoken Language Translation (IWSLT). IWSLT 2013 (6-7 déc. 2013). Heidelberg, Germany, 2013, 8 pages.

[28] Li Gong, Aurélien Max et François Yvon. "Incremental Development of Statistical Machine Translation Systems". In : *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*. Sous la dir. de Marcello Federico, Sebastian Stüker et François Yvon. Lake Tahoe, CA, 2014, p. 214-222.

[29] Ayyoob Imani, Lütfi Kerem Senel, Masoud Jalili Sabet, François Yvon et Hinrich Schuetze. "Graph Neural Networks for Multiparallel Word Alignment". In : *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland : Association for Computational Linguistics, mai 2022, p. 1384-1396.

[30] Julia Ive, Aurélien Max et François Yvon. "LIMSI's Contribution to the WMT'16 Biomedical Translation Task". In : *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, août 2016, p. 469-476.

[31] Julia Ive, Aurélien Max et François Yvon. "Reassessing the proper place of man and machine in translation : a pre-translation scenario". In : *Machine Translation* (oct. 2018). issn : 1573-0573.

[32] Julia Ive, Aurélien Max, François Yvon et Philippe Ravaud. "Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations". In : *International Conference on Language Resources and Evaluation - Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016 2016)*. Portorož, Slovenia, mai 2016, p. 8.

[33] Julia Ive et François Yvon. "Parallel Sentence Compression". In : *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. COLING 2016. Osaka, Japon, 13/12 au 16/12 2016, p. 11.

[34] Masoud Jalili Sabet, Philipp Dufter, François Yvon et Hinrich Schütze. "SimAlign : High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings". In : *Findings of the Association for Computational Linguistics : EMNLP 2020*. Online : Association for Computational Linguistics, nov. 2020, p. 1627-1643.

[35] Elena Knyazeva, Guillaume Wisniewski, Hervé Bredin et François Yvon. "Structured Prediction for Speaker Identification in TV Series". In : *Proceedings of the 16th Annual Conference of the International Speech Communication Association (InterSpeech)*. Dresden, Germany, sept. 2015, 5 pages.

[36] Elena Knyazeva, Guillaume Wisniewski et François Yvon. "Apprentissage par imitation pour l'étiquetage de séquences: vers une formalisation des méthodes d'étiquetage "easy-first"". In : *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*. Caen, France : ATALA, juin 2015, (12).

[37] Catherine Kobus, François Yvon et Géraldine Damnati. "Normalizing SMS : are two metaphors better than one ?" In : *Proceedings of the International Conference on Computational Linguistics (COLING)*. Manchester, UK, 2008, p. 441-448.

[38] Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski et François Yvon. "Frustratingly Easy Cross-Lingual Transfer for Transition-Based Dependency Parsing". In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, juin 2016, p. 1058-1063.

[39] Philippe Langlais et François Yvon. “Scaling up analogical learning”. In : *Proceedings of the International Conference on Computational Linguistics (COLING)*. Manchester, UK, 2008, p. 49-52.

[40] Philippe Langlais et François Yvon. “Issues in Analogical Inference Over Sequences of Symbols: A Case Study on Proper Name Transliteration”. In : *Computational Approaches to Analogical Reasoning: Current Trends*. Sous la dir. d’Henri Prades et Gilles Richard. Springer-Verlag Berlin Heidelberg, 2014, p. 59-82.

[41] Philippe Langlais, François Yvon et Pierre Zweigenbaum. “Improvements in Analogical Learning : Application to Translating multi-Terms of the Medical Domain”. In : *Proceedings of the Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*. Athens, Greece, 2009, p. 487-495.

[42] Adrien Lardilleux, François Yvon et Yves Lepage. “Generalizing sampling-based multilingual alignment”. In : *Machine Translation* 27.1 (2013), p. 1-23.

[43] Thomas Lavergne, Alexandre Allauzen, Josep Maria Crego et François Yvon. “From n-gram-based to CRF-based Translation Models”. In : *Proceedings of the Sixth ACL Workshop on Statistical Machine Translation*. Edinburgh, Scotland, 2011, p. 542-553.

[44] Thomas Lavergne, Olivier Cappé et François Yvon. “Practical Very Large Scale CRFs”. In : *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden : Association for Computational Linguistics, 2010, p. 504-513.

[45] Thomas Lavergne et François Yvon. “Learning the Structure of Variable-Order CRFs: a finite-state perspective”. In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2017. Copenhagen, Denmark : Association for Computational Linguistics, 2017, p. 433-439.

[46] Hai-Son Le, Alexandre Allauzen et François Yvon. “Continuous Space Translation Models with Neural Networks”. In : *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Montréal, Canada : Association for Computational Linguistics, juin 2012, p. 39-48.

[47] Hai-Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain et François Yvon. “Structured Output Layer Neural Network Language Model”. In : *Proceedings of ICASSP*. 2011, p. 5524-5527.

[48] Hai-Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain et François Yvon. “Structured Output Layer Neural Network Language Models for Speech Recognition”. In : *Audio, Speech, and Language Processing, IEEE Transactions on* 21.1 (2013), p. 197-206.

[49] Shiuan-Sung Lin et François Yvon. “Discriminative training of finite-state decoding graphs”. In : *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech)*. Lisbon, Portugal, sept. 2005, p. 733-736.

[50] Benjamin Marie et al. “LIMSI@WMT’15 : Translation Task”. In : *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, sept. 2015, p. 145-151.

[51] Tomas Mikolov, Wen-tau Yih et Geoffrey Zweig. “Linguistic Regularities in Continuous Space Word Representations”. In : *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Atlanta, Georgia : Association for Computational Linguistics, juin 2013, p. 746-751.

[52] Hemant Misra, François Yvon, Olivier Cappé et Joemon Jose. “Text segmentation : A topic modeling perspective”. In : *Information Processing and Management* 47.4 (2011), p. 528-544. issn : 0306-4573.

[53] Erwan Moreau, François Yvon et Olivier Cappé. “Robust Similarity Measures for Named Entities Matching”. In : *Proceedings of the International Conference on Computational Linguistics (COLING)*. Manchester, UK, 2008, p. 593-600.

[54] Anh Khoa Ngo Ho et François Yvon. “Generative latent neural models for automatic word alignment”. In : *The Association for Machine Translation in the Americas*. Florida, United States, oct. 2020.

[55] Shu Okabe, Laurent Besacier et François Yvon. “Weakly Supervised Word Segmentation for Computational Language Documentation”. In : *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland : Association for Computational Linguistics, mai 2022, p. 7385-7398.

[56] Nicolas Pécheux, Li Gong, Quoc Khanh Do, Benjamin Marie, Yulia Ivanishcheva, Alexandre Allauzen, Thomas Lavergne, Jan Niehues, Aurélien Max et François Yvon. "LIMSI @ WMT 14 Medical Translation Task". In : *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA, juin 2014, p. 246-253.

[57] Nicolas Pécheux, Guillaume Wisniewski et François Yvon. "Reassessing the value of resources for cross-lingual transfer of POS tagging models". In : *Language Resources and Evaluation* (2016), p. 1-34.

[58] Minh Quang Pham, Josep Maria Crego, François Yvon et Jean Senellart. "A Study of Residual Adapters for Multi-Domain Neural Machine Translation". In : *Proceedings of the Fifth Conference on Machine Translation*. Online : Association for Computational Linguistics, nov. 2020, p. 615-626.

[59] Minh-Quang Pham, Josep Crego et François Yvon. "Revisiting Multi-Domain Machine Translation". In : *Transactions of the Association for Computational Linguistics* 9.0 (2021), p. 17-35. issn : 2307-387X.

[60] Minh-Quang Pham, François Yvon et Josep Crego. "Latent Group Dropout for Multilingual and Multidomain Machine Translation". In : *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States : Association for Computational Linguistics, juill. 2022, p. 2469-2481.

[61] Vito Pirrelli et François Yvon. "The hidden dimension : a paradigmatic view of data-driven NLP". In : *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)* 11.3 (1999), p. 391-408.

[62] Henri Prades et Gilles Richard, éd. Springer-Verlag Berlin Heidelberg, 2014, p. 59-82.

[63] Loïs Rigouste, Olivier Cappé et François Yvon. "Inference and Evaluation of the Multinomial Mixture Model for Text Clustering". In : *Information Management and Processing* 43.5 (jan. 2007), p. 1260-1280.

[64] Nataliya Sokolovska, Thomas Lavergne, Olivier Cappé et François Yvon. "Efficient Learning of Sparse Conditional Random Fields for Supervised Sequence Labeling". In : *IEEE Journal of Selected Topics in Signal Processing, Special issue on 'Statistical Learning Methods for Speech and Language Processing'* 4.6 (2010), p. 953-964.

[65] Nicolas Stroppa et François Yvon. "An analogical learner for morphological analysis". In : *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL'2005)*. Ann Arbor, MI, 2005, p. 120-127.

[66] Nicolas Stroppa et François Yvon. "Du quatrième de proportion comme principe inductif: une proposition et son application à l'apprentissage de la morphologie". In : *Traitement Automatique des Langues (TAL)* 47.1 (2007).

[67] Nicolas Stroppa et François Yvon. "Proportions in the lexicon : (re)discovering paradigms". In : *Lingue e Linguaggio* VI.2 (2007), p. 201-226.

[68] Nadi Tomeh, Alexandre Allauzen et François Yvon. "Maximum-entropy word alignment and posterior-based phrase extraction for machine translation". In : *Machine Translation* 28.1 (2014), p. 1-38.

[69] Romain Vinot et François Yvon. "Improving Rocchio with weakly supervised clustering". In : *European Conference on Machine Learning (ECML)*. Cavtat-Drubrovnick (Croatie), 2003, p. 456-467.

[70] Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham et François Yvon. "Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning". In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, oct. 2014, p. 1779-1785.

[71] Guillaume Wisniewski, Anil Kumar Singh et François Yvon. "Quality estimation for machine translation: some lessons learned". In : *Machine Translation* 27.3 (2013), p. 213-238.

[72] Guillaume Wisniewski, Lichao Zhou, Nicolas Ballier et François Yvon. "Biais de genre dans un système de traduction automatiquen euronale : une étude préliminaire". In : *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*. Sous la dir. de Pascal Denis, Natalia Grabar, Amel Fraisse, Rémi Cardon, Bernard Jacquemin, Eric Kergosien et Antonio Balvet. Lille, France : ATALA, 2021, p. 11-25.

[73] Jitao Xu, Josep Crego et François Yvon. "Bilingual Synchronization : Restoring Translational Relationships with Editing Operations". In : *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*. Abou Dabi, United Arab Emirates, déc. 2022.

[74] Jitao Xu et François Yvon. “One Source, Two Targets: Challenges and Rewards of Dual Decoding”. In : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online et Punta Cana, Dominican Republic : Association for Computational Linguistics, nov. 2021, p. 8533-8546.

[75] Yong Xu, Aurélien Max et François Yvon. “Sentence Alignment for Literary Texts”. In : *Linguistic Issues in Language Technology* 12.6 (oct. 2015), 25 pages.

[76] Yong Xu et François Yvon. “A 2D CRF Model for Sentence Alignment”. In : *9th Workshop on Building and Using Comparable Corpora (BUCC)*. Portorož, Slovenia : European Language Resources Association, mai 2016.

[77] Yong Xu et François Yvon. “Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts”. In : *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC 2016)*. Sous la dir. de Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk et Stelios Piperidis. Portorož, Slovenia : European Language Resources Association (ELRA), mai 2016, p. 10.

[78] François Yvon. “Des apprentis pour le traitement automatique des langues”. Thèse de doct. Mémoire d’habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris, 2006.

[79] François Yvon. “Rewriting the orthography of SMS messages”. In : *Natural Language Engineering* 16.2 (2010), p. 133-159.

[80] François Yvon, Nicolas Stroppa, Arnaud Delhay et Laurent Miclet. *Solving analogical equations on words*. 2004-D005. École Nationale Supérieure des Télécommunications, 2004.

[81] François Yvon, Yong Xu, Marianna Apidianaki, Clément Pillias et Pierre Cubaud. “TransRead : Designing a Bilingual Reading Experience with Machine Translation Technologies”. In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*. San Diego, California, juin 2016, p. 27-31.

# 5 Formation initiale et doctorale

## 5.1 Enseignements en formation initiale

Depuis le début de ma carrière, mes enseignements se répartissent à part égale entre informatique générale (compilation, programmation, algorithmique) et des enseignements de spécialité en intelligence artificielle et particulièrement en traitement automatique des langues et en apprentissage automatique. Cette répartition n’a que peu évolué à l’occasion de ma nomination à l’Université Paris-Sud, même si les pratiques pédagogiques et le contenu des cours ont eux changé de manière très significative, ce qui m’a conduit à renouveler en 2007 pratiquement l’intégralité de mes enseignements. Depuis 2012 (délégation, puis détachement et intégration au CNRS), mon activité d’enseignement s’est considérablement réduite, et s’est concentrée sur des interventions au niveau Master 2 dans des cours spécialisés en traitement des langues.

Je distingue dans la suite ces trois périodes, en commençant par la plus récente.

### 5.1.1 En tant que chercheur CNRS (2013-)

Durant mon mandat de directeur du LIMSI (2013-2019), mon activité d’enseignement a été réduite à un cours d’introduction aux modèles graphiques probabilistes offert aux étudiants de M2 de différents Masters (Master Informatique / IAC à Paris-Sud, Master Mathématique et Applications / Statistiques et Machine Learning à Paris-Sud, Master DataScience à l’Ecole Polytechnique). Il représentait un volume de 31h Eq TD.

Mes activités d’enseignement actuelles se déroulent principalement dans le cadre du Master d’Informatique de l’Université Paris-Saclay. Entre 2019 et 2022, j’ai également été chargé de cours à l’École Polytechnique, où j’ai enseigné un cours de traitement des langues (NLP) pour les étudiants du Master ’*DataScience*’ de l’Institut Polytechnique de Paris, et, dans une version condensée, aux étudiants du Master X-HEC ’*DataScience for Business*’. Cette charge s’est accompagnée de suivi de projets d’élèves (deux groupes en 2019, 4 groupes en 2020, 3 en 2021). En 2020 également, j’ai créé un nouveau cours *Multilingual Natural Language Processing*

à l'Université Paris-Saclay. Tous ces cours ont lieu en anglais. Un bilan comptable pour la période récente est dans le tableau ci-dessous.

| Cours | Public | 18-19 | 19-20 | 20-21 | 21-22 |
|---|---|---|---|---|---|
| Graphical models for text mining | M2 Info (Paris-Saclay) | 31,5h | 31,5h | 31,5h | 31,5h |
| Natural Language Processing | M2 DataScience (IPP) | - | 31,5h | 31,5h | 31,5h |
| Multilingual NLP | M2 IA (Paris-Saclay) | - | - | 27h | 27h |

En 2021, je suis intervenu dans l'école d'été '*Data Science*' de l'Ecole Polytechnique ; en 2022 j'ai été invité à donner un cours sur l'apprentissage pour les données linguistiques dans le cadre de l'Ecole de Printemps d'Informatique Théorique (EPIT, au CRIM de Lumigny).

J'interviens enfin ponctuellement dans des conférences / formations destinées à des étudiants en linguistique / traductologie : en 2020 dans les journées d'études 'Qualité en traduction' (Univ. Lille) et '*Artificial Intelligence & Intercultural Intelligence*' (CIUTI/ISIT) ; en 2021 dans les Masters de traduction de Bordeaux / Montaigne et de Louvain la Neuve).

### 5.1.2 En tant que professeur à l'Université Paris-Sud (2007-2013)

À l'Université Paris-Sud, j'ai d'abord eu une activité d'enseignant universitaire standard, répartie entre d'une part l'UFR des sciences, dans les différentes filières de Master 1 et Master 2, et d'autre part l'école d'Ingénieur de l'Université (IFIPS, aujourd'hui Polytech Paris-Sud). J'y ai enseigné deux cours d'informatique générale : « Introduction à la programmation », dans la filière »Compétence complémentaire en Informatique », et « Compilation », à l'école d'Ingénieur. Le reste des enseignements correspond à cours de spécialisation dans mes domaines d'expertise : traitement de la parole, fouille de texte, traitement des langues, apprentissage automatique, à Polytech et dans les différents parcours de Master d'Informatique de l'UFR. Tous ces cours sont au niveau M (deuxième ou troisième année du cycle ingénieur à Polytech, master 1 ou 2 à l'UFR). En 2010-2011, j'ai notamment développé un nouveau cours de niveau M1 (24 de cours + 24h de Travaux pratiques) constituant une introduction au traitement automatique des langues axée sur les aspects algorithmiques.

Le tableau 3 donne, à titre indicatif, les heures enseignées par niveau et par type de cours ; il n'intègre pas les décharges horaires dont je bénéficie au titre des responsabilités administratives et pédagogiques (en 2007-2008 : 20 h ; en 2008-2009 : 35h ; en 2009-2010 45h ; en 2010-2011 55h ; et en 2011-2012 40h).

| | CM | TP | TD | Stages |
|---|---|---|---|---|
| 2007-2008 | 101 | 36 | 7 | 8 |
| 2008-2009 | 86 | 21 | 7 | 15 |
| 2009-2010 | 58 | 16 | 47 | * |
| 2010-2011 | 100 | 24 | 12 | * |
| 2011-2012 | 115 | 42 | - | * |

TABLE 3 – Répartitition des enseignements (U Paris Sud) - en eq. TD

### 5.1.3 À l'ENST / Télécom ParisTech (1996-2007)

De 1996 à 2007, je suis intervenu dans les enseignements d'informatique de l'ENST, mon activité se répartissant à peu près également entre les cours de première année du cycle ingénieur (théorie des langages, algorithmique et théorie des graphes, base de l'optimisation numérique) et les divers cours de spécialisation en intelligence artificielle qui se sont succédés au fil des réformes (introduction à l'intelligence artificielle, Prolog, Lisp, représentation des connaissances, traitement automatique du langage, de la parole, apprentissage automatique et reconnaissance des formes) effectués en seconde ou troisième année. Les charges d'enseignement à l'ENST dépendent en dernière instance des choix des élèves, selon les flux dans les différentes options, ce qui rend difficile d'établir des comparaisons avec un service d'universitaire. À titre d'exemple, je donne dans le tableau ci-dessous une image fidèle de mes enseignements pour l'année 2005/2006 ce qui représente (hors encadrement de projets, de stages et de thèse) environ une centaine d'heures de cours, soit une charge considérée comme plutôt élevée par rapport aux standards de l'établissement (voir le tableau 4). La première année de l'ENST correspond à un niveau L3, les deux dernières années à un niveau M.

| Intitulé | Niveau | Volume |
|---|---|---|
| Théorie des langages | 1ere année ENST | 16,5h + 12h (cours) |
| Structures données, algorithmique | 1ere année ENST | 20.5h (cours) |
| Optimisation | 1ere année ENST | 15h (cours) |
| Analyse et Grammaires | 2/3eme année ENST | 9h (cours) |
| Traitement des Langues | 2/3eme année ENST | 12h (cours), 12h (TPs) |
| | module ATHENS(*) | 6h (cours), 9h (TPs) |
| | M2 UPMC / ICATAL | 14h (cours) |
| Divers | 2/3eme année ENST | 7.5 (cours) |
| Total | | 112.5h (cours), 21h (TPs) |

(*) En anglais

TABLE 4 – Enseignements 2005/2006

L'enseignement à Télécom ParisTech évoluant très rapidement, j'ai eu souvent à créer des nouveaux enseignements. Je suis en particulier à l'origine de la « restauration » d'un cours de théorie des langages (en première année du cycle ingénieur), que j'ai assuré entre 2002 et 2014, et pour lequel j'ai écrit avec Akim Demaille d'épaisses notes de cours, qui servaient aussi bien aux étudiants de Télécom qu'à ceux de l'EPITA (et, nous l'espérons, à d'autres, car ces notes sont en libre service sur le Web [17]).

Je suis enfin intervenu de manière marginale dans des formations professionnelles, et de manière beaucoup moins marginale dans le suivi de projets de fin d'études (une dizaine) et de stages ingénieurs (autant).

### 5.2 Direction et animation de formations, dont partenariats internationaux

À mon arrivée à l'Université Paris-Sud j'ai pris en charge (de 2007 à 2013) la responsabilité de la spécialité « Compétence complémentaire en Informatique » (CCI), en collaboration avec Hervé Delacroix. Cette formation accueillait chaque année entre 20 et 30 étudiants ayant au moins un M1 dans une autre discipline scientifique, et qui souhaitent acquérir une double compétence en informatique. En tant que responsable, j'ai eu à faire évoluer et à mettre en place la maquette de la formation pour le contrat quadriennal 2010-2014.

Pour le contrat quadriennal 2010-2014, j'ai également activement participé à la conception de la nouvelle spécialité du Master Recherche en Informatique « Intelligence, Apprentissage, Cognition » ; j'ai notamment été responsable responsable du suivi des stages jusqu'en 2012.

### 5.3 Formation doctorale

J'ai encadré ou co-encadré 9 thèses à Télécom Paris (école doctorale EDITE) et 16 à l'Université Paris-Sud / Paris-Saclay (école doctorale EDIPS, puis ED-STIC). Six doctorants ont soutenu leur thèse dans la période récente. Je supervise actuellement 3 thèses au sein de l'ED-STIC, dont une en contrat CIFRE (avec la société Systran), une financé sur contrat de recherche et la dernière financée par une bourse doctorale de l'Université Paris-Saclay.

Mes anciens étudiants sont aujourd'hui employés dans l'enseignement supérieur ou dans des emplois de recherche et développement, dans le secteur privé, à l'étranger ou en France, au sein de grands groupes (Google, Amazon, ByteDance, Orange Labs, Airbus, Safran, etc), ou de PMEs (Quantmetry).

Un détail des thèses encadrées est donné dans la section A.1.

## 6 Responsabilités Collectives

### 6.1 En tant que directeur du LIMSI

J'ai été nommé officiellement directeur du LIMSI [18] le 1er juillet 2013, après avoir été étroitement associé à la direction du laboratoire depuis le début de l'année 2013 ; j'ai été confirmé à la direction de l'Unité pour

17. www.limsi.fr/Individu/yvon/classes/thl/thl-2.pdf
18. http://www.limsi.fr.

le contrat quinquenal 2015-2020. Le LIMSI était une unité propre du CNRS, associée par convention avec l'Université Paris-Sud (ainsi que, jusque fin 2013, avec l'Université Pierre et Marie Curie) qui abritait environ 110 personnels permanents, dont environ 30 ITA et un nombre équivalent de chercheurs CNRS ; le laboratoire accueillait également un nombre respectable de doctorants (actuellement environ 70), de post-doctorants et de stagiaires, avec un effectif qui fluctuait annuellement entre 200 et 250 personnes.

**Actions de transformation du LIMSI** Avec l'aide des deux directeurs adjoints, A. Vilnat et C. Tenaud, responsable chacun d'un département scientifique, j'ai pris une large part à l'élaboration du rapport d'évaluation AERES de 2013, et j'ai supervisé l'écriture de trois rapports scientifiques en 2014, 2015 et 2016, et j'ai enfin piloté la préparation et l'accompagnement côté laboratoire de l'évaluation HCERES de fin 2018. Tous ces rapports sont disponibles publiquement sur le site du laboratoire [19].

En plus de l'activité quotidienne de gestion scientifique, humaine, administrative et financière d'un laboratoire propre du CNRS, réparti sur des locaux d'une surface totale d'environ 8000 m$^2$, et doté d'un budget consolidé de plus de 15 ME, un certain nombre de chantiers de plus long terme ont été initiés ou conduits à leur terme pendant mes deux mandats. Parmi les principaux :

- l'aménagement d'une nouvelle salle machine (livrée en juillet 2014, pour un budget total de 700 KE) et l'installation de nouvelles capacités de calcul. Nous avons en parallèle travaillé à simplifier et d'uniformiser de la gestion des comptes utilisateurs, ainsi qu'à mieux mutualisé les moyens de calcul du laboratoire ;
- l'implémentation de la Politique de Sécurité des Systèmes d'Information (PSSI) au sein du laboratoire. L'intégralité des locaux du LIMSI étant placé sous le régime de la ZRR (Zone à Restriction d'accès) depuis 2014, la PSSI s'applique au laboratoire avec des règles de sécurité maximales :
- la refonte intégrale du site Web du laboratoire (intranet et internet) pour passer à des standard plus modernes de gestion du site. J'ai également impulsé la mise en place d'une politique de communication sur les réseaux sociaux, avec l'ouverture d'un compte Twitter (@LimsiLab, qui compte plus de 1 000 abonnés) et d'un compte LinkedIn.
- la diffusion (sur HAL) de la production scientifique du laboratoire, avec la création d'une collection LIMSI [20] qui compte aujourd'hui plus de 3 500 entrées.
- une refonte (dématérialisation) des procédures d'arrivée et de départ du laboratoire ;
- réorganisation de l'activité des ITA en soutien de la recherche et création de deux entités mutualisées au niveau du laboratoire en charge respectivement des plateformes informatiques et des salles d'expérimentation en IHM ; d'autre part des montages expérimentaux, en particulier pour les activités en mécanique-énergétique ;
- la construction d'un nouveau bâtiment destiné à abriter l'ensemble des activités en traitement des langues du laboratoire (pour une surface totale d'environ 2500 m$^2$ budget total de 8ME). Débuté en 2008, ce projet au long court s'est achevé par la livraison du bâtiment aux équipes de recherche à l'hiver 2017, et par son inauguration officielle en mars 2018. Un dernier chantier est en préparation, qui concerne l'aménagement d'un auditorium dans le rez-de-chaussée du bâtiment 510, une opération qui mettra un point définitif à la construction de ce bâtiment, en exploitation pour ses zones expérimentales depuis 2010.
- la préparation, avec une douzaine de laboratoires français, d'un projet d'Institut Carnot sur les Sciences Cognitive, qui a été labellisé par l'ANR "Tremplin Carnot" en juillet 2016. Après trois ans de fonctionnement, et sur la base des bons résultats obtenus durant cette période, le Tremplin Carnot Cognition est devenu en 2019 un institut Carnot de plein exercice [21]. J'ai représenté le LIMSI au sein du Comité de pilotage de cet Institut Carnot.

**Le LIMSI et le plateau de Saclay** Une des mes principales activités en tant que Directeur du LIMSI a été l'accompagnement de la difficile gestation de la nouvelle Université Paris-Saclay (UPSay), dont la fondation (en 2015) a marqué le début d'un ensemble de recompositions thématiques et institutionnelles de grande ampleur, qui ont finalement conduit à un renouvellement de l'IDEX au printemps 2018 sur un périmètre plus restreint, mais également sans doute plus cohérent, et devrait déboucher en 2020 à la fusion entre l'Université Paris-Sud et l'Université Paris-Saclay. Les premiers effets de la création de la COMUE Paris-Saclay se sont manifestés dans le domaine de l'enseignement (création d'une École doctorale en Sciences et Technologies

19. https://www.limsi.fr/fr/rapports-d-activite
20. https://hal.archives-ouvertes.fr/LIMSI
21. http://www.institut-cognition.com/

de l'information et la Communication (STIC) au niveau du site, mise en place de diplômes de Masters délivrés directement par l'Université Paris-Saclay). Au niveau de la recherche, l'effet principal de l'IdEx a été d'amplifier les collaborations entre équipes au sein des laboratoires de l'UPSay, en bénéficiant du cadre et des financements redistribués via les LabEx Digicosme et LaSips, ainsi que les objets propres de l'IdEX (l'Institut pour la Société Numérique (ISN), l'Institut pour le Controle et la Décision (Icode), le *Center for Data Science*, etc).

En tant que directeur du LIMSI, je me suis également fortement investi dans la construction de l'Université Paris-Saclay : j'ai été ainsi successivement impliqué dans les instances de gouvernance (comité de pilotage, conseil scientifique, etc) du RTRA Digitéo, du LabEx Digicosme [22], du département STIC de l'Université Paris-Saclay, du *Center for Data Science* [23], ainsi que plus récemment de l'institut de Convergence DataIA [24]. En plus de la coordination de facto de la réflexion stratégique sur le domaine de l'Interaction Humain-Machine au sein du département STIC, j'ai été particulièrement moteur dans deux projets importants pour l'Université.

Le premier concerne la mise en place du cercle Polethis [25], instance d'animation de l'enseignement, de la recherche et de la réflexion sur l'éthique et l'intégrité scientifique à Paris-Saclay. J'ai initié la mise en place (dès 2016) d'un Comité d'Ethique de la Recherche de l'Université Paris-Saclay (CER-PS), qui est devenu en 2018, suite à la création du Polethis, une instance officielle de l'Université Paris-Saclay. Je suis chargé de mission au sein du Polethis, en charge de coordonner les activités du CER-PS. Ce comité, dont j'ai contribué à formaliser le fonctionnement et à définir les procédures et les outils de gestion, est opérationnel depuis 2017 et a, entre 2017 et 202 réalisé l'expertise de plus de 70 protocoles émanant de nombreux établissements de l'Université Paris-Saclay. Il a également organisé deux journées de sensibilisation en décembre 2017 et en mars 2019.

J'ai d'autre part assuré le portage d'un projet régional (Ile-de-France) SESAME « Saclay-IA » d'un montant total de 1.3 ME, visant à équiper les équipes de recherche en IA du département STIC (au sein du LRI, du LIMSI, du L2S, du CEA-LIST, etc) d'une plateforme de calcul GPU mutualisée. L'achat de cet équipement réalisé grâce à un cofinancement Paris-Sud / CNRS / CEA / IMT et région IDF. La plateforme (comprenant environ une trentaine de noeuds de calcul et une centaine de cartes) a aujourd'hui été complètement spécifiée, achetée et installée dans ses deux composantes ; elle est entrée en production au début 2019, prenant la suite d'un équipement mutualisé financé par le CNRS et dont j'ai supervisé la mise en place en 2017. Ces actions ont permis d'une part de maintenir nos capacités de calcul à un niveau opérationnel, en période de fort développement des activités en IA, et en préparation de l'installation d'une plateforme nationale (Jean Zay) à partir de 2020. Depuis 2021, je participe à la mise en place d'un Mésocentre au niveau Paris-Saclay, qui abrite en particulier cette plateforme de calcul.

La structuration progressive scientifique du site Paris-Saclay a été une opération longue, complexe et sinueuse, qui a consommé une énergie considérable pour des résultats encore incertains, même si les dernières évolutions laissent entrevoir des perspectives de stabilisation du périmère et de la définition d'un projet d'Université intégrée, qui ne se concrétisera réellement qu'après 2025, si jamais il parvient à son terme. Pour ce qui concerne les Sciences de l'Information, cette période aura été marquée par le début d'une réflexion sur le projet, impulsé par l'INS2I, d'un rapprochement des unités sous tutelle CNRS, et dont le débouché principal a été d'une part la fusion du LIMSI et d'une partie du LRI voisin pour donner lieu au LISN, alors que l'autre partie du LRI fusionnait avec le LSV arrivant sur le plateau de Saclay pour former le LMF.

L'ensemble de ce bilan et le projet d'unité associé ont été présentés au comité de visite de l'HCERES en décembre 2018.

## 6.2 En tant que Professeur de l'Université Paris-Sud

À l'Université Paris-Sud, ma principale responsabilité administrative a été celle de *vice-président du département d'informatique* de l'UFR des sciences de l'Université Paris-Sud, entre 2009 et 2012. Compte-tenu de sa taille (plus de 800 enseignants-chercheurs dans toutes les disciplines scientifiques), l'UFR des Sciences d'Orsay est structurée en 6 départements disciplinaires, parmi lesquels le département d'Informatique regroupe trois laboratoires (le LRI, le LIMSI et la Maison de la Simulation), soit environ 150 enseignants-chercheurs et

22. `http://labex-digicosme.fr/`
23. `http://www.datascience-paris-saclay.fr/`
24. `https://dataia.eu`
25. `https://www.universite-paris-saclay.fr/fr/polethis`

chercheurs. Le département est formellement reconnu au sein de l'UFR, est représenté dans toutes les instances de direction de la Faculté, et est systématiquement consulté pour toutes les affaires qui le concernent, qu'elles touchent à l'enseignement, à la recherche, aux questions de personnels permanents et temporaires, ou encore aux ressources (matériels, locaux, etc). En plus de participer, au sein du bureau de département, à la gestion au quotidien de nos activités, mes attributions principales concernaient la communication et les finances du département.

J'ai également été membre du conseil de l'Ecole Doctorale en Informatique de Paris-Sud (EDIPS) entre 2009 et 2013 ; membre du conseil de laboratoire du LIMSI, entre 2009 et 2013 ; membre (suppléant) du comité de programme de Digitéo entre 2011 et 2014.

## 6.3 Animation d'équipes de recherche

J'anime depuis septembre 2007 les activités en traduction automatique et en apprentissage statistique au sein du LIMSI, puis du LISN. Ce thème a fédéré pendant la période les activités de 5 chercheurs et enseignants-chercheurs permanents (2 chercheurs CNRS, 1 professeur et 2 maîtres de conférences), ainsi qu'un nombre fluctuant (entre 5 et 10 d'étudiants en doctorat, post-doctorat ou stage de Master).

Mon activité en tant que responsable du thème comprend l'animation scientifique (animation du groupe de lecture bi-mensuel, invitation de conférenciers), les recrutements de personnels temporaires, la rédaction des rapports d'activité, enfin la gestion et le prospection, le montage et le suivi des activités contractuelles touchant à la traduction automatique (voir la section 4.5).

## 6.4 Autres responsabilités collectives

Á l'extérieur du LIMSI/LISN mes principales responsabilités ont été les suivantes

- membre du Comité National du CNRS en Section 07 (2012-2013)
- membre du membre du Conseil exécutif du réseau européen META-NET (depuis 2014)
- création et présidence du Comité d'éthique de la recherche de l'Université Paris-Saclay (entre 2016 et 2019) ; à ce titre membre du conseil Polethis de l'Université ;
- membre du bureau du Réseau Francilien des Sciences de l'Information (de 2016 à 2021) ;
- portage du projet de financement (2017), puis animation de la plateforme de calcul DataIA, une plateforme mutualisée pour le calcul GPU sur le plateau de Saclay, maintenant intégrée au Mésocentre Paris-Saclay ;
- président de la commission 'contenus', membre du bureau exécutif du pôle de compétitivé Cap Digital (depuis 2019) ;
- membre du conseil scientifique du GDR TAL, co-responsable de l'axe 'Multilinguisme et Traduction' (depuis 2019) ;
- membre du comité des programmes de l'institut DataIA (Paris Saclay, jusqu'en 2020) ;
- membre du bureau exécutif de l'EACL (*European Association for Computational Linguistics*, depuis 2021)

## 6.5 Responsabilités et mandats nationaux, ou régionaux

*Contexte : L'ENST/Télécom ParisTech et l'établissement public (le GET / Institut Mines Télécom) qui exerce la tutelle des Écoles des Télécommunication ont un fonctionnement très différent de celui d'une université (pré-LRU). Le GET est un établissement public administratif (EPA) autonome, créé par la loi de privatisation de France Télécom en 1996 et dont les personnels ont un statut unique et spécifique à l'établissement. En particulier, les personnels peuvent choisir entre un contrat de droit public et un contrat de droit privé, ce qui est une singularité dans l'univers des établissements publics administratifs. L'ENST et le GET sont chacun dirigés par un Directeur, qui s'appuie sur le conseil d'administation de l'établissement. Des comités paritaires comprenant des élus du personnel sur listes syndicales sont consultés et émettent des avis sur les dossiers individuels (c'est le rôle de la Commission Consultative Paritaire (CCP)) et collectifs (c'est le rôle du Comité Technique Paritaire (CTP), qui est compétent pour examiner les politiques de rémunération, de primes, de congés, d'aide sociale etc.).*

Entre 1996 et 2007, j'ai exercé au sein de l'ENST/Télécom ParisTech, ainsi qu'au sein de l'Etablissement

public de tutelle (le Groupe des Ecoles de Télécommunication - GET), un certain nombre de mandats en tant qu'élu des personnels, dont en particulier :

— membre élu du Conseil d'École de l'ENST (2005-2007)
— membre élu du CTP et de la CCP de l'ENST (1996-2007)
— membre élu du CTP et de la CCP du GET/Institut Télécom (1996-2007)
— membre élu du Comité Hygiène et Sécurité de l'ENST (1996-2004)

Mon activité au sein de ces instances m'a conduit, en moyenne, à participer à plusieurs dizaines de réunions chaque année, en formation pleinière ou réduite. Ces responsabilités m'ont amené, en particulier, à prendre une part très active à la redéfinition des statuts des personnels de l'établissement, redéfinition qui s'est principalement opérée durant les années de transition 1997-2000.

# 7 Rayonnement

## 7.1 Diffusion du savoir scientifique

Le thème de la traduction automatique fait l'objet de débats animés, en particulier dans les milieux de la traduction professionelle, mais de plus en plus également auprès du grand public qui est de plus en plus amené à utiliser ces technologies. Pour aider à mieux faire comprendre les limitations des systèmes actuels et les recherches auxquels ils donnent lieu, j'ai été sollicité pour mon expertise scientifique à des titres divers.

J'ai d'une part collaboré sur ces questions avec des artistes contemporains, notamment avec Magali Debazeille, dans le cadre du projet C2m1 [26] et du metteur en scène Jean-François Peyret [27] dans le cadre de la création du pièce de théâtre (Re :Walden, jouée au théâtre de la Colinne en 2014).

Je suis également régulièrement sollicité pour des interviews dans la presse (généraliste ou professionnelle) concernant la traduction automatique ou plus généralement le traitement des langues, ou à la radio. J'ai ainsi récemment participé aux émissions de France Culture « Autour de la question » (en 2017), « la danse des mots » (en 2017), « la méthode scientifique » (en 2018 et en 2022).

Je participe à des tables rondes lors de conférences scientifiques ou professionnelles, par exemple *France is AI* en 2019 et 2020. J'ai enfin été sollicité à des conférences sur la traduction automatique pour d'étudiants de formations en traduction (à Bruxelles et à Paris en 2011, à Dijon en 2017, à Rome en 2018) ou auprès du grand public (à Avignon en 2010, et en Bourgogne en 2018).

L'ensemble de ces activités est documenté sur mon site personnel.

## 7.2 Expertise

**Évaluation d'équipes et de projets** Durant la période récente, j'ai mené deux missions d'expertise de grande ampleur : la présidence du comité de visite du GIPSA-LAB (Grenoble), rassemblant 12 experts (janvier 2020) ; la participation au comité de visite du centre d'excellence ADAPT (Dublin), qui rassemble 8 universités irlandaises autour des thématiques de l'analyse d'information multimedia multilingue (mars 2019). J'ai également expertisé au fil de l'eau des projets de moindre ampleur, au niveau européen (3 projets ERC), national (3 projets ANR), plus de nombreux projets locaux, dans cadre du DIM RFSI, de l'Institut DATA-IA ou du pôle de compétitivité Cap Digital.

Dans un passé plus lointain, j'ai réalisé de nombreuses expertises de projets pour le compte de l'ANR ou de l'ANRT (thèses CIFRE), ainsi que pour diverses agences européennes (Suisse, Allemagne, Belgique, Pays-Bas) ; j'ai également participé deux fois à l'évaluations des équipes projets Inria travaillant sur le traitement automatique des langues (en 2011 et en 2015).

**Évaluations individuelles, comités de recrutement** J'ai été membre de la CCSU 27e section de l'Université Paris Saclay sans discontinuer de 2013 à 2021. Je participe régulièrement à des comités de recrutement - 8 depuis 2018, 4 pour des postes de Professeurs, 3 pour des postes de Maîtres de Conférences, 1 pour des postes de CR Inria (au centre de Lille).

Le détail des missions d'évaluation est donné dans la section C.

26. http://www.desbazeille.fr/v2/index.php?/projects/c2m1/
27. http://www.theatrefeuilleton2.net/

**Jurys de thèse** Depuis mon HDR en 2007, j'ai participé à près de 80 jurys de thèses, dont plus de la moitié en tant que rapporteur. Depuis 2018, j'ai participé à 16 jurys, dont 7 en tant que rapporteur. Le détail est donné dans la section B.

## 7.3 Expertise pour des journaux et conférences

**Responsabilités éditoriales** Je suis *Action Editor* de la revue *Transactions of the ACL* (2018-) ; *Associate Editor* des revues *ACM Computing Surveys* (2019-2022), et *Computer Speech and Language* (2020-). Je suis également membre du Comité de Rédaction de la revue *Traitement Automatique des Langues* (2005-) dont j'ai été co-rédacteur en chef entre 2007 et 2011. Je suis enfin *Action Editor* au sein du système d'édition « au fil de l'eau » de l'Association for Computational Linguistics (*ACL Rolling Review*).

J'effectue des relectures ponctuelles pour les principales revues du domaine : *Computational Linguistics*, *Transaction of the ACL*, *IEEE transactions on Knowledge and Data Engineering*, *IEEE/ACM transactions on Speech and Audio*, *Journal of Natural Language Engineering*, *Journal of Artificial Intelligence Research* ; *Speech Communications*, *Machine Translation*, etc.

**Conférences** Depuis 2010, j'ai servi comme membre du comité de programme de grand nombre de conférences francophones et internationales dans les domaines du traitement des langues et de la traduction automatique (*ACL, COLING, EMNLP, AMTA, LREC, InterSpeech) ; et plus sporadiquement en apprentissage automatique et en intelligence artificielle et en Apprentissage Automatique (AAAI, NeurIPS, ICML, IJCAI). J'ai en particulier servi à plusieurs reprises comme responsable de thèmes (*Aera Chair*) dans des conférences *ACL : en 2021 '*Senior Area Chair*' pour thème « Traduction Automatique et Multilinguisme » de la principale conférence du domaine (ACL/IJCNLP 2021), en 2022 '*Senior Area Chair*' pour le thème « Phonologie, Morphologie et Segmentation en mots ».

J'ai présidé en 2010 et 2014 le comité scientifique de la conférence IWSLT (*International Workshop on Spoken Language Translation*) ; je suis membre du *Steering committee* de la conférence.

# A Direction d'étudiants : stages, doctorats et post-doctorats

Cette section présente les thèses de doctorat que j'ai encadrées ou co-encadrées à l'ENST, au sein de l'école doctorale EDITE puis à l'Université Paris-Sud / Paris-Saclay (au sein de l'EDIPS, puis de l'ED-STIC) en indiquant brièvement le devenir de ces docteurs. Je mentionne également les thèses actuellement en cours, ainsi que, plus sommairement, l'encadrement de stages de Master et d'étudiants en stage post-doctoral.

## A.1 Thèses de doctorat

### A.1.1 Thèse d'Ariane Halber (à 50%)

Ariane Halbert a soutenu sa thèse le 8 décembre 2000, devant un jury composé de Renato de Mori (Univ. Avignon, président), Anne Abeillé (Paris VII, rapporteur), Christophe Fouqueré (Paris XIII, rapporteur), Giorgio Satta (Univ. Bologne, rapporteur), Gérard Chollet (ENST et CNRS, Co-directeur de thèse). Elle a obtenu la mention « Très Honorable ».

Cette thèse porte sur le couplage reconnaissance vocale/TAL en particulier sur les grammaires d'arbres adjoints lexicalisées (LTAGs) et les algorithmes d'analyse tabulaire pour les LTAGs. Elle a fait l'objet d'une convention CIFRE avec Thomson/LCR.

Ariane Halber est aujourd'hui directrice technique à Vecsys, responsable des projets de développement de portails vocaux.

### A.1.2 Thèse de Florence Duclaye (à 50 %)

Florence Duclaye a soutenu sa thèse le 19 novembre 2003, devant un jury composé de Ludovic Lebart (ENST et CNRS, président), Béatrice Daille (IRIN, rapporteur), Benoît Habert (Paris X et LIMSI CNRS, rapporteur), Laurent Miclet (ENSSAT, examinateur), Olivier Collin (FT R&D, co-encadrant), François Yvon (ENST, directeur de thèse). Elle a obtenu la mention « Très Honorable ».

Cette thèse portait sur l'acquisition automatique de paraphrases à partir du Web. Elle a fait l'objet d'une convention avec FT R&D (Orange Labs).

Florence Duclaye est toujours ingénieur de recherche à Orange Labs, à Lannion.

### A.1.3 Thèse de Romain Vinot (à 100%)

Romain Vinot a soutenu sa thèse le 12 février 2004, devant un jury composé de Ludovic Lebart (ENST et CNRS, président), Florence d'Alché-Buc (LIP 6, rapporteur), Martin Rajman (EPFL, rapporteur), Yannick Toussaint (Loria Nancy, examinateur), Éric Gaussier (XeroX ERC, examinateur), François Yvon (ENST, directeur de thèse). Il a obtenu la mention « Très Honorable ».

Cette thèse porte sur la catégorisation automatique de textes.

Romain Vinot est décédé en 2021, après avoir effectuée l'essentiel de sa carrière chez Google, à Zurich.

### A.1.4 Thèse de Nicolas Stroppa (à 100%)

Nicolas Stroppa a soutenu sa thèse le 10 novembre 2005, devant un jury composé de Jacques Sakarovitch (ENST et CNRS, président), Pierre Zweigenbaum (LIMSI-CNRS), François Denis (Univ. Marseille, rapporteur), Laurent Miclet (ENSSAT Lannion, rapporteur), Vito Pirrelli (CNR/ILC Pise, examinateur), François Yvon (ENST, directeur de thèse). Il a obtenu la mention « Très Honorable ».

Cette thèse porte sur l'apprentissage par analogie pour le traitement des langues et a été financée par une bourse de l'École Doctorale (EDITE).

Nicolas Stroppa est ingénieur de recherche chez Google, à Zurich.

### A.1.5 Thèse de Loïs Rigouste (à 50 %)

Loïs Rigouste a soutenu sa thèse le 8 novembre 2006 devant un jury composé de Ludovic Lebart (ENST et CNRS), Michèle Sebag (LRI, Univ. Paris XI, rapporteur), Eric Gaussier (Univ. Grenoble, rapporteur),

Fabrice Clérot (FT R&D, examinateur), Olivier Cappé (ENST et CNRS, directeur de Thèse), François Yvon (ENST, co-directeur de Thèse). Il a obtenu la mention « Très Honorable ».

Cette thèse, co-encadrée avec Olivier Cappé, porte sur les modèles probabilistes pour la fouille de texte. Elle a fait l'objet d'un contrat de recherche avec FT R&D.

Loïs Rigouste est ingénieur de recherche chez Myscript, à Nantes, (après un passage chez Pertimm, à Paris).

#### A.1.6 Thèse de Lin Shuan-Sung (à 100%)

Shuan-Sung Lin a soutenu sa thèse le 8 novembre 2007 devant un jury composé de Jean-Paul Haton (LORIA), Kamel Smaïli (LORIA, rapporteur), Paul Deléglise (Univ. Le Mans, rapporteur), Gérard Gollet (ENST Paris, examinateur), François Yvon (ENST, directeur de Thèse). Il a obtenu la mention « Très Honorable ».

Cette thèse, portait sur la construction les techniques d'apprentissage discriminant pour la reconnaissance automatique de la parole.

Lin Shuan-Sung est aujourdh'hui ingénieur de Recherche à Taïwan.

#### A.1.7 Thèse de Alexandra Krul (à 30 %)

Alexandra Krul a soutenu sa thèse le 12 décembre 2008 devant un jury composé de André Salem (Univ. Paris III), Frédéric Béchet (LIA, Univ. Avignon, rapporteur), Olivier Boëffard (ENSSAT Lannion, rapporteur), Gaël Richard (ENST Paris, examinateur), Thierry Moudenc (FT R&D, examinateur), Géraldine Damnati (FT R & D, co-directrice de thèse), François Yvon (ENST, co-directeur de Thèse). Elle a obtenu la mention « Très Honorable ».

Cette thèse, co-encadrée avec Géraldine Damnati, porte sur la constitution de bases de données textuelles pour la synthèse de parole par sélection d'unités. Elle s'est déroulée pour sa plus grande partie dans les locaux de FT R&D/Orange Labs à Lannion.

Aleksandra Krul est ingénieur de recherche à Orange Labs, à Lannion.

#### A.1.8 Thèse de Thomas Lavergne (à 30%)

Thomas Lavergne a soutenu sa thèse le 3 avril 2009 devant un jury composé de Patrick Gallinari (Univ. Paris VI), Isabelle Tellier (Univ. Orléans, rapporteur), Mohand Boughanem (IRIT, Univ. Toulouse III, rapporteur), Romain Vinot (Yahoo !, examinateur), Jean-Louis Dessalles (ENST Paris, examinateur), Tanguy Urvoy (FT R & D, co-directeur de thèse), François Yvon (Univ. Paris-Sud, co-directeur de Thèse). Il a obtenu la mention « Très Honorable ».

Cette thèse, co-encadrée avec Thomas Urvoy, porte sur la détection automatique de langage faussement naturel dans le contexte de filtrage de spam sur le Web. Elle s'est déroulée pour sa plus grande partie dans les locaux de FT R&D/Orange Labs à Lannion.

Thomas Lavergne est Maitre de Conférences à l'Université Paris-Sud.

#### A.1.9 Thèse de Nataliya Sokolovska (à 50 %)

Natalyia Sokolovska a soutenu sa thèse le 25 février 2010 devant un jury composé de Thierry Artières (Univ. Paris VI), Marc Tommasi (Univ. Lille , rapporteur), Yves Granvallet (Univ. Technique de Compiègne, rapporteur), Francis Bach (INRIA, examinateur), Olivier Cappé (ENST Paris, co-directeur de Thèse), François Yvon (Univ. Paris-Sud, co-directeur de Thèse). Elle a obtenu la mention « Très Honorable ».

Cette thèse, co-encadrée avec Olivier Cappé, porte sur diverses extensions des champs aléatoires conditionnels : apprentissage semi-supervisé, sélection de caractéristiques.

Nataliya Sokolovska est Maitre de Conférences à Sorbonne Universités.

#### A.1.10 Thèse de Nadi Tomeh (à 50 %)

Nadi Tomeh a soutenu sa thèse à l'Université Paris Sud le 26 juin 2012 , devant un jury composé de Anne Vilnat (présidente), Éric Gaussier (Univ. Joseph Fourrier, rapporteur), Philippe Langlais (Univ. Montréal,

rapporteur), Hermann Ney (RWTH Aix la Chappelle), Nasredine Semmar (CEA, Invité), Alexandre Allauzen (LIMSI et Univ. Paris Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse).

Ce travail de thèse porte sur la traduction automatique, et plus précisément sur l'utilisation de méthodes d'apprentissage discriminant pour l'estimation de modèles de traduction à partir de corpus alignés ou partiellement alignés. Nadi Tomeh a obtenu une bourse de l'école doctorale en Informatique de Paris Sud.

Nadi Tomeh est Maitre de Conférences à l'Université Paris-Nord.

#### A.1.11 Thèse de Le Hai-Son (à 20 %)

Le Hai-Son a soutenu sa thèse à l'Université Paris Sud le 20 décembre 2012 , devant un jury composé de Holger Schwenk (Université du Maine, rapporteur)), Laurent Besacier (Université Joseph Fourrier, rapporteur), Yoshua Bengio (Univ. Montréal), Hermann Ney (RWTH Aix la Chappelle), Michelle Sebag (LRI Orsay, présidente), Alexandre Allauzen (LIMSI et Univ. Paris Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse).

Cette thèse porte sur l'utilisation de modèles neuronaux, en particulier de modèles neuronaux profonds pour la traduction automatique. Le Hai-Son a obtenu une bourse de l'école doctorale en informatique de Paris Sud.

Le Hai Son est chercheur à l'Académie des Sciences du Vietnam.

#### A.1.12 Thèse de Panagiota Karanasou (à 50 %)

Le Hai-Son a soutenu sa thèse à l'Université Paris Sud le 13 juin 2013 , devant un jury composé de Frédéric Béchet (Université de Marseille, rapporteur)), Eric Fosler-Lussier (Ohio State University, rapporteur), Lukáš Burget (Univ. Brno), Denis Jouvet (Loria, Nancy), Anne Vilnat (LIMSI et Univ. Paris-Sud, présidente), Lori Lamel (LIMSI, co-directrice de thèse) et François Yvon (co-directeur de thèse).

Cette thèse porte sur l'apprentissage de modèles pour les variantes de prononciation en reconnaissance vocale.

Panagiota Karanasou est ingénieure de recherche à Amazon research (Cambridge).

#### A.1.13 Thèse de Souhir Ghabiche (à 50 %)

Souhir Ghabiche a soutenu sa thèse, débutée en octobre 2009, en septembre 2013. Son jury était composé de Emmanuel Morin (Univ. Nantes, rapporteur), Kamel Smaïli (Univ. Nancy, rapporteur), Laurent Besacier (Univ. Grenoble, examinateur), Pierre Zweigenbaum (LIMSI, président), Hélène Bonneau-Meynard (LIMSI et Univ. Paris-Sud, co-directrice de thèse) et François Yvon (co-directeur de thèse).

Cette thèse de l'Université Paris-Sud, co-encadrée avec Hélène Bonneau-Meynard (LIMSI et Univ. Paris Sud), a été financée par le projet FUI/SAMAR. Elle portait sur l'utilisation de ressources linguistiques « riches » dans le cadre de la traduction automatique par des méthodes statistiques, dans le cadre d'un projet industriel visant à développer des outils de traduction automatique depuis l'arabe vers le français et l'anglais. S. Ghabiche est aujourd'hui ingénieure de Recherche à Harmann International (Paris).

#### A.1.14 Thèse de Li Gong (à 50 %)

Li Gong a été inscrit en thèse à l'Université Paris-Sud entre octobre 2011 et novembre 2014, dans le cadre d'un co-encadrement avec Aurélien Max (LIMSI et Univ. Paris-Sud). Le jury était composé de Marc Dymetman (XRCE Grenoble, rapporteur), Andy Way (Dublin College University, Dublin), Béatrice Daille (Univ. Nantes, examinatrice), Christian Jacquemin (LIMSI et Univ. Paris Sud, président), Aurélien Max (co-directeur de thèse) et François Yvon (co-directeur de thèse).

Cette thèse portait sur la conception et l'évaluation de systèmes de traduction automatique capables d'apprendre à la volée, et donc d'intégrer une partie des avantages des méthodes à base d'exemples. Elle était financée par une bourse de l'école doctorale en informatique de Paris-Sud.

Á l'issue de sa thèse, Li Gong a obtenu un poste d'ingénieur de recherche à Baidu, Beijin. Il est aujourd'hui chercheur à Systran (Paris).

### A.1.15 Thèse de Khahn Quoc Do (à 50 %)

Khahn Quoc Do a effectué sa thèse à l'Université Paris-Sud entre octobre 2012 et septembre 2016, en co-encadrement avec Alexandre Allauzen (LIMSI et Univ. Paris Sud). La soutenance a eu lieu le 31 mars 2016, devant un jury composé de Christof Monz (Univ. Amsterdam, rapporteur), Thierry Artières (Aix-Marseille University, rapporteur), Holger Schwenk (Facebook AI Research Paris, examinateur), Laurence Likforman-Sulem (Télécom Paristech, présidente), Alexandre Allauzen (LIMSI et Univ. Paris Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse). FIXME.

Cette thèse porte sur l'utilisation de modèles neuronaux profonds en traduction automatique, notamment sur l"étude de fonctions de pertes alternatives à la log-entropie. Khahn Quoc Do a été financé par une bourse de l'école doctorale en informatique de Paris-Sud. Il est aujourd'hui ingénieur de recherche chez Safran *Identity and Security*.

### A.1.16 Thèse de Yong Xu (à 100 %)

Yong Xu s'est inscrit en thèse à l'Université Paris-Sud en octobre 2012. La soutenance a eu lieu le 26 septembre 2016, devant un jury composé de Philippe Langlais (Univ. Montréal, rapporteur), Olivier Kraif (Univ. Grenoble, rapporteur), Yannick Estève (Univ. du Maine, examinateur), Stéphane Huet (Univ. Avignon, examinateur), Pierre Zweigenbaum (LIMSI, président) et François Yvon (co-directeur de thèse).

Cette thèse portait sur la production de mesures de confiance pour l'alignement phrastique et sous-phrastique de haute-qualité. Elle a été financée par le projet ANR/Transread.

Á l'issue de sa thèse, Yong Xu a obtenu un poste d'ingénieur de recherche à Baidu, Beijin.

### A.1.17 Thèse de Nicolas Pécheux (à 50 %)

Nicolas Pécheux a effectué sa thèse à l'Université Paris Sud entre octobre 2012 et septembre 2016, en co-encadrement avec Alexandre Allauzen (LIMSI et Univ. Paris Sud). La soutenance a eu lieu le 27 septembre 2016, devant un jury composé de Isabelle Tellier (Univ. Paris 3, rapporteur), de Frédéric Lefèvre (Univ. Avignon, rapporteur), Massih-Reza Amini (Univ. Grenoble, examinateur), Anne Vilnat (LIMSI et Univ. Paris-Sud, présidente) Alexandre Allauzen (LIMSI et Univ. Paris-Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse).

Cette thèse portait sur l'étude de la modélisation de la distortion en traduction automatique statistique fondée sur des modèles globalement conditionels. Nicolas Pécheux avait obtenu une bourse de l'école normale supérieure de Cachan. Il est aujourd'hui professeur d'informatique en classes préparatoires.

### A.1.18 Thèse de Julia Ive (à 20 %)

Julia Ive a effectué sa thèse à l'Université Paris Sud entre février 2014 et septembre 2017 ; son jury était composé de Pierrette Bouillon (Univ. Genève, rapporteur), Marco Turchi (FBK, Trento, rapporteur), Emmanuel Planas (Univ. Angers, examinateur), Philippe Ravaud (Univ. Paris Descartes), Nicolas Sabouret (LIMSI et Univ. Paris-Sud, président), Aurélien Max (LIMSI et Univ. Paris Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse) FIXME.

Cette thèse portait sur l'étude de dispositifs pour une meilleure coopération humain-machine en traduction automatique et s'est intéressée plus particulièrement aux méthodes de pré-édition et post-édition. Elle a bénéficié d'un financement CIFRE, en partenariat avec l'association Robert Debré et le centre Cochrane France.

Julia Ive est assistant professor à University College London (UK).

### A.1.19 Thèse de Lauriane Auffrant (à 20 %)

Lauriane Auffrant s'est inscrite en thèse à l'Université Paris-Sud en janvier 2015, en co-encadrement avec Guillaume Wisniewski (LIMSI et Univ. Paris-Sud), et a soutenu sa thèse le 5 avril 2018, devant un jury composé de Benoit Crabbé (Univ Paris Diderot, rapporteur), Anders Søggard (Univ. Copenhagen, rapporteur), Javier Carreras (dMetrics, examinateur), Pierre Zweigenbaum (LIMSI, président), Guillaume Wisniewski, (LIMSI et Univ. Paris Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse)..

Cette thèse portait sur l'apprentissage cross-lingues de modèles de parsing en dépendance. Lauriane Auffrant est maintenant ingénieure experte à Inria Paris, en charge avec des applications duales des technologies numériques.

### A.1.20 Thèse de Elena Knyazeva (à 50 %)

Elena Knyazeva s'est inscrite en thèse à l'Université Paris-Sud en septembre 2013, en co-encadrement avec Guillaume Wisniewski (LIMSI et Univ. Paris-Sud), et a soutenu sa thèse le 25 mai 2018, devant un jury composé de Thierry Artières (Aix-Marseille University, rapporteur), Pascal Denis (INRIA Lille, rapporteur), Artem Sokolov (Amazon Berlin, examinateur), Matthieu Geist (Univ. Metz, examinateur), Sophie Rosset (LIMSI, présidente), Guillaume Wisniewski, (LIMSI et Univ. Paris-Sud, co-directeur de thèse) et François Yvon (co-directeur de thèse).

Cette thèse portait sur l'apprentissage cross-lingue de modèles de parsing en dépendances. Elena Knyazeva est aujourd'hui post-doctorante au LISN.

### A.1.21 Thèse de Pierre Godard (à 50 %)

Pierre Godard a soutenu sa thèse à l'Université Paris-Sud le 16 avril 2019, devant un jury composé de Adam Lopez (Univ. Edimburg, rapporteur), Christophe Cerisara (LORIA, rapporteur), Emmanuel Dupoux (EHESS et Facebook, examinateur), Pierre Zweigenbaum (LIMSI, CNRS, examinateur), Laurent Besacier (LIG Grenoble, co-directeur de thèse) et François Yvon (co-directeur de thèse). Sa thèse portait sur les modèles de segmentation et d'alignement non-supervisés pour l'outillage des langues peu dotées ; elle était financée par le projet ANR/DFG BULB (*Breaking the Unwritten Language Barrier*).

Durant sa thèse, P. Godard a notamment étudié les modèles bayésiens non-paramétriques pour la segmentation en unités lexicales ou en morphèmes ; il s'est également intéressé au développement de méthodes neuronales inspirées des architectures encodeur-décodeur pour effectuer la même tâche.

Pierre Godard a repris à la fin de sa thèse à l'activité qu'il menait antérieurement, à sa voir la direction d'une compagnie de danse contemporaine.

### A.1.22 Thèse de Anh-Khoa Ngo-Ho (à 100 %)

Anh-Khoa Ngo-Ho a soutenu sa thèse le 8 février 2021 devant un un jury composé de Loïc Barrault (Univ. Sheffield, rapporteur), Yves Lepage (LORIA, rapporteur), Nadi Tomeh (Univ Paris Nord, examinateur), Pierre Zweigenbaum (LIMSI, CNRS, examinateur) et François Yvon (co-directeur de thèse). Cette thèse, débuté en octobre 2017 à l'Université Paris-Sud, a bénéficié d'une bourse doctorale de l'Université Paris-Saclay.

Le travail de Anh-Khoa Ngo-Ho a porté sur l'alignement sous-phrastique pour des langues morphologiquement complexes, et dans ce cadre il a notamment étudié des variantes neuronales des modèles d'alignement probabilistes classiques (IBM), analysé leur interaction avec les algorithmes de segmentation sous-lexicale, et développé un modèle d'alignement original s'appuyant sur le formalisme des auto-encodeurs variationnels. M. Ngo-Ho est actuellement *Data Scientist* dans la société Quantmetry à Paris.

### A.1.23 Thèse de Minh Quang Pham (à 100 %)

Minh Quang Pham a soutenu sa thèse le 11 décembre 2021 devant un un jury composé d'Alex Fraser (LMU Munich, rapporteur), Rico Sennrich (Univ. Zurich, rapporteur), Marine Carpuat (Univ. Maryland, examinatrice), Josep Crego (Systran, examinateur), Pierre Zweigenbaum (LISN, CNRS, examinateur) et François Yvon (directeur de thèse). Son travail portait sur l'adaptation automatique d'architectures neuronales pour la traduction multidomaine ; elle s'est déroulée dans le contexte d'un contrat CIFRE avec la société Systran (Paris), en collaboration étroite avec Josep Maria Crego.

Les principales contributions de M. Pham ont porté sur le développement et l'analyse de nouvelles méthodes d'adaptation multi-domaine pour la traduction automatique, ainsi que sur la construction d'un cadre générique pour évaluer ces méthodes. M. Pham est aujourd'hui ingénieur de recherche chez Zoom (Karlsruhe, Allemagne)

#### A.1.24 Thèse de François Buet (à 100 %)

François Buet est inscrit en thèse depuis octobre 2018 à l'Université Paris-Saclay. Son travail porte sur la simplification automatique de la parole et la génération automatique de sous-titres à partir de parole ; elle s'est principalement déroulée dans le contexte du projet Rosetta (Grands défis du Numérique, BPI). M. Buet a également bénéficié d'un contrat d'ATER en 2021-2022. F. Buet a soutenu sa thèse le 21 octobre 2022 devant un jury composé de Christophe Cerisara (LORIA, CNRS, rapporteur), Benoit Favre (LIS, AMU, rapporteur), Yannick Estève (LIA, Univ. Avignon), Thierry Etchegoyhen (Vicomtech), Annelies Braffort (LISN, CNRS, examinatrice) et François Yvon (directeur de thèse).

F. Buet s'est principalement intéressé à étendre les méthodes neuronales de simplification automatique au cadre du sous-titrage d'émissions télévisuelles, qui se caractérisent par des contraintes multiples (longueur maximale, durée miminale d'exposition) ainsi que par leur grande variété.

#### A.1.25 Thèse de Jitao Xu (à 100 %)

Jitao Xu est inscrit en thèse depuis décembre 2019 à l'Université Paris-Saclay. Son travail porte sur la génération automatique de textes simultanément en deux langues ; elle est financée par la Région Ile-de-France, dans le cadre du programme *PhD Talent* et elle est co-financée par la société Systran. La soutenance se tiendra le 2 décembre 2022, le jury étant composé de Qun Liu (Noah Ark Lab, Huawei, rapporteur), Philippe Langlais (DIRO, Univ. Montréal, rapporteur), Jan Niehues (KIT, rapporteur), Rachel Bawden (Inria Paris, examinatrice), Pierre Zweigenbaum (LISN, CNRS, examinateur), Josep Crego (Systran, invité), et François Yvon (LISN, CNRS, directeur de thèse).

#### A.1.26 Thèse de Shu Okabé (à 100 %)

Shu Okabé est inscrit en thèse depuis octobre 2020 à l'Université Paris-Saclay et était co-encadré jusqu'en septembre 2021 par L. Besacier (LIG, Univ. Grenoble Alpes). Son travail porte sur l'apport des méthodes de traitement automatique à la documentation des langues, en particulier pour les étapes de segmentation en mots et de génération de gloses ; elle se déroule dans le contexte du contrat ANR-DFG "Computational Language Documentation". La soutenance de thèse est prévue pour la fin 2023.

#### A.1.27 Thèse d'Alban Petit (à 25 %)

Alban Petit est inscrit en thèse depuis septembre 2020 à l'Université Paris-Saclay en co-encadrement avec Caio Corro (LISN, Université Paris-Saclay). Son travail porte sur l'analyse sémantique profonde ; elle bénéficie d'un contrat doctoral dans le cadre du programme UDOPIA de l'Université Paris-Saclay. La soutenance est prévue pour la fin 2023.

#### A.1.28 Thèse de Maxime Bouthors (à 100 %)

Maxime Bouthors est inscrit en thèse depuis avril 2022, financé par une convention CIFRE avec la société Systran. Sa thèse porte sur l'étude des méthodes non-paramétriques pour améliorer la traduction neuronale avec des mémoires de traduction dans un contexte industriel. Elle est menée en collaboration étroite avec Josep Maria Crego (Systran).

### A.2 Post-doctorants

| Date | Nom | Source de financement |
|---|---|---|
| 03/20 - 09/21 | Sadaf Abdul Rauf | SOULT - Contextes étendus en traduction automatique |
| *Maintenant Assistant Professor à Fatima Jinnah Women University, Rawalpindi, Pakistan* | | |
| 03/15 - 03/18 | Franck Burlot | QT21 - Traduction automatique pour les langues à morphologie riche |
| *Maintenant Data Scientist à Cornerstone On Demand, Paris* | | |
| 03/15 - 09/16 | Ophélie Lacroix | Papyrus - Transfert cross-lingue pour l'analyse syntaxique |
| *O. Lacroix est Data Scientist à Wunderman Thompson MAP, à Copenhague* | | |

| | | |
|---|---|---|
| 03/15 - 08/15 | Raphaël Bailly | Papyrus - Méthodes spectrales pour l'analyse en dépendances |
| *Maintenant Maitre de Conférences au SAMOS, Univ. Paris Panthéon Sorbonne* | | |
| 01/14 - 01/15 | Natalia Segal | DGA Rapid RapMat - Traduction automatique de la parole |
| *N. Segal est ingénieure de recherche chez Systran, Paris* | | |
| 05/12 - 04/13 | Anil Kumar Singh | ANR Trace - Mesures de qualité en traduction automatique |
| *A. Kumar Singh est associate Professor à IIT (BHU) Varanasi, Uttar Pradesh, India* | | |
| 12/12 - 10/13 | Marco Dinarelli | Quaero - Traduction automatique Statistique |
| *M. Dinarelli est chercheur CNRS au LIG, Grenoble* | | |
| 10/10 - 03/12 | Adrien Lardilleux | SAMAR - Alignement automatique sous phrastique |
| *A. Lardilleux est aujourd'hui Consultant NLP à De Cronos Groep, Luxembourg* | | |
| 2010 - 2012 | Artem Sokolov | Quaero - Traduction automatique statistique |
| *Artem Sokolov est ingénieur de recherche chez Google, Berlin et checheur associé à Université d'Heidelberg* | | |
| 2009 - 2012 | Thomas Lavergne | ANR CROTAL - Implémantation des champs aléatoires conditionnels (CRF) |
| *Thomas Lavergne est Maitre de conférences à l'Université Paris Saclay.* | | |
| 2009 - 2010 | Ilknur Durgar | Quaero - Analyse morphologique en traduction automatique. |
| *Ilknur Durgar est aujourd'hui Machine Learning Lead, à la Turkish Radio and Television Corporation, Istanbul* | | |
| 02/08 - 08/11 | Josep M. Crego | Quaero - Traduction automatique statistique |
| *J. M. Crego est directeur de la recherche chez Systran, Paris* | | |
| 2006 - 2008 | Hemant Misra | Modèles à données latentes pour la transcription automatique et l'indexation audio |
| *Hemant Misra est Senior Vice President, Global Decision Management Group, Citicorp (Inde)* | | |
| 2006-2007 | Erwan Moreau | Infomagic - Variations orthographiques dans les noms propres |
| *Erwan Moreau est aujourd'hui Research Fellow à Trinity College, à Dublin.* | | |

## A.3 Stagiaires de DEA et Master 2

| | | |
|---|---|---|
| 2000-2001 | Jean-Philippe Demoulin | DEA Ingéniérie des Langues, Univ. Marne La Vallée |
| *Synchronisation parole / transparents pour la reconnaissance vocale de cours* | | |
| 2000-2001 | Romain Vinot | DEA IARFA, Univ. Paris VI |
| *Filtrage et Routage de Mails* | | |
| 2003-2004 | Robert Wettinger | Masters Thesis, KTH |
| *Semi-Hidden Markov Models for Sequence Labelling* | | |
| 2004-2005 | Mehdi Guemmar | Univ. Paris-Dauphine |
| 2006-2007 | Amine Lajmi, | Univ. Paris-Dauphine |
| *Méthodes bayésiennes pour l'apprentissage de modèles de classification* | | |
| 2008-2009 | Raphael Payen | Master Traduction, Université Paris Diderot |
| *Analyse des erreurs de traduction automatique* | | |
| 2009-2010 | Mohamed Sehili | Master SETI, Univ. Paris-Sud |
| *Méthodes d'ensemble en traduction automatique* | | |
| 2009-2010 | Lucie Martinet | Master 1, Université Claude Bernard Lyon |
| *Graph-based methods in Dependency Parsing* | | |
| 2010-2011 | Fan Zhang | Master SETI, Univ. Paris-Sud |
| *Modèles globalement conditionnels pour la traduction automatique* | | |
| 2010-2011 | Qian Yu | Master SETI, Univ. Paris-Sud |
| *Modèles pour l'alignement de phrases* | | |
| 2012-2013 | Nicolas Pécheux | Master MVA, ENS Cachan |
| *Modèles conditionnels pour la traduction automatique* | | |
| 2012-2013 | Yong Xu | Master AIC, Univ. Paris-Sud |
| *Alignement de livres bilingues* | | |
| 2012-2013 | Khahn Quoc Do | Stage Ingénieur, Télécom ParisTech |

| |
|---|
| *Modèles neuronaux bayésiens* |
| 2015-2013 Lauriane Auffrant Master IAC, Univ. Paris-Saclay<br>*Apprentissage cross-lingue de la syntaxe - Roumain et langues romanes* |
| 2017-2018 Minh Quang Pham M2 Datascience, Univ. Paris-Saclay<br>*Apprentissage dual pour la traduction automatique* |
| 2017-2018 Trong Bach Vu Master AIC, Univ. Paris-Saclay<br>*Méthodes neuronales pour l'alignement de mots* |
| 2019-2020 Sooyong Park Master PluriTAL, Univ. Paris Censier<br>*Ressources pour le calcul automatique de gloses* |
| 2020-2021 Xinneng Xu Master IA, Univ. Paris Saclay<br>*Improving word alignments by exploiting constraints of parsimony in machine translation* |
| 2021-2022 Félix Herron Master IA, Univ. Paris Saclay<br>*Génération automatique de néonymes* |

# B Participation à des jurys de thèse et de HDR

1. Philippe Boula de Mareüil, Université Paris Sud, soutenue le 19 décembre 1998
2. Boris Cormons, Thèse de l'Université de Rennes 1, soutenue le 22 mars 1999
3. Antoine Rozenknop, Thèse de l'École Polytechnique Fédérale de Lausanne, soutenue le 9 décembre 2002 (avec rédaction d'un rapport)
4. Laurent Blin, Thèse de l'Université de Rennes 1, soutenue le 19 décembre 2002
5. Pierre Alain, Thèse de l'Université de Rennes 1, soutenue le 23 janvier 2007
6. Benoît Favre, Thèse de l'Université d'Avignon et des pays de Vaucluse, soutenue le 19 mars 2007 (Rapporteur)
7. Sabri Bayoudh, Thèse de l'Université de Rennes 1, soutenue le 14 novembre 2007 (Rapporteur)
8. Florent Jousse, Thèse de l'Université de Lille 1, soutenue le 31 octobre 2007 (Rapporteur)
9. Guillaume Wisniewski, Thèse de l'Université Pierre et Marie Curie, soutenue le 29 novembre 2007 (Rapporteur)
10. Stéphane Huet, Thèse de l'Université de Rennes 1, soutenue le 11 décembre 2007 (Rapporteur)
11. Huyen Trang Vu, Thèse de l'Université Pierre et Marie Curie, soutenue le 27 octobre 2008 (Rapporteur)
12. Benjamin Lecouteux, Thèse de l'Université d'Avignon et des pays de Vaucluse, soutenue le vendredi 5 décembre 2008 (Rapporteur)
13. Jean-Baptiste Bordes, Thèse de l'École Nationale Supérieure des Télécommunications, soutenue le 3 mars 2009 (Rapporteur)
14. Nicolas Hervé, Thèse de l'Université Paris Sud, soutenue le 8 juin 2009
15. Carmela Ignat, Thèse de l'Université de Strasbourg, soutenue le 16 juin 2009, (Rapporteur)
16. Tuong Vinh Truong, Thèse de l'Université Pierre et Marie Curie, soutenue le 8 octobre 2009 (Rapporteur)
17. Vassilina Nikoulina, Thèse de l'Université Joseph Fourrier (Grenoble), soutenue le 19 mars 2010
18. Adrien Lardilleux : Thèse de l'Université de Caen, soutenue le 14 septembre 2010. (Rapporteur).
19. Olivier Hamon, Thèse de l'Université Paris Nord, soutenue le 8 décembre 2010, (Rapporteur).
20. Romaric Gaudel, Thèse de l'Université Paris Sud, soutenue le 14 décembre 2010.
21. Stéphane Clinchant, Thèse de l'Université Joseph Fourrier (Grenoble) soutenue le 5 octobre 2011 (Rapporteur)
22. Sasa Hasan, Phd de l'Université Aix-la-Chapelle, soutenue le 25 novembre 2011, (Rapporteur)
23. Cyril Joder, Thèse de Télécom ParisTech, soutenue le 29 septembre 2011, (Rapporteur)
24. Anne-Laure Bianne-Bernard, Thèse de Télécom ParisTech, soutenue le 21 novembre 2011 (Président)
25. Isam Rebai, Thèse de l'Université Paris-Sud, soutenue le 18 mai 2011, (Président)

26. Fabien Poulard, Thèse de l'Université de Nantes, soutenue le 24 mars 2011 (Rapporteur)
27. Arnaud Grappy, Thèse de l'Université Paris-Sud, soutenue le 8 novembre 2011 (Président)
28. Camille Guinaudeau, Thèse de l'Université Rennes 1, soutenue le 7 décembre 2011 (Rapporteur)
29. Julien Gosmes, Thèse de l'Université de Caen, soutenue le 13 février 2012
30. Mathieu Dubois, Thèse de l'Université Paris-Sud, soutenue le 20 février 2012 (Président)
31. Yuqi Zhang, Phd Université Aix-la-Chapelle, soutenue le 12 mai 2012 (Rapporteur)
32. Amel Hamzaoui, Thèse de l'Université Paris-Sud, soutenue le 10 mai 2012 (Président)
33. Houda Bouamor, Thèse de l'Université Paris-Sud, soutenue le 11 juin 2012 (Président)
34. Alexander Pak, Thèse de l'Université Paris-Sud, soutenue le 13 juin 2012 (Président)
35. Sheng Gao, Thèse de l'Université Pierre et Marie Curie, soutenue le 18 juin 2012 (Rapporteur)
36. Pratyush Banerjee, Phd, Dublin City University, soutenue le 21 décembre 2012 (External examiner)
37. Nadir Durrani, Phd University of Stuttgart, soutenue le 19 novembre 2012
38. Basam Jabain, Thèse de l'Université d'Avignon et des pays de Vaucluse, soutenue le 4 décembre 2012 (Rapporteur)
39. Sylvain Raybaud, Thèse de l'Université de Lorraine, soutenue le 5 décembre 2012 (Rapporteur)
40. Ilya Loshchilov, Thèse de l'Université Paris-Sud, soutenue le 5 janvier 2013 (Président)
41. Zeeshan Ahmed, Phd University College Dublin, soutenue le vendredi 13 septembre 2013 (External examiner)
42. Jan Niehues, Phd Karlsruhe Institute of Technology, soutenue le 17 janvier 2014 (rapporteur)
43. Sokol Koco, Thèse de l'Université d'Aix-Marseille, soutenue le 16 décembre 2013 (Rapporteur)
44. Édouard Grave, Thèse de l'Université Pierre et Marie Curie, soutenue le 20 janvier 2014 (Rapporteur)
45. Dhouha Bouamor Thèse de l'Université Paris-Sud, soutenue le 21 février 2014 (Président)
46. Mousa Amr El-Desoky, Phd Université Aix-la-Chapelle, soutenue le 18 juin 2014 (Rapporteur)
47. Gaëtan Marceau-Caron, Thèse de l'Université Paris-Sud, soutenue le 22 septembre 2014 (Président)
48. Mohammed Morchid Thèse de l'Université d'Avignon et des pays de Vaucluse, soutenue le 25 novembre 2014 (Rapporteur)
49. Alexandre Chotard, Thèse de l'Université Paris-Sud, soutenue le 24 septembre 2015 (Président)
50. Marc Evrard, Thèse de l'Université Paris-Sud, soutenue le 30 septembre 2015 (Président)
51. Seyedabolghasem Mirroshandel, Thèse de l'Université d'Aix-Marseille, soutenue le 10 décembre 2015 (Rapporteur)
52. Amel Hamzaoui, Thèse de l'Université Paris-Sud, soutenue le 10 mai 2012 (Président)
53. Moussa El Desocky, Phd Université Aix-la-Chapelle, soutenue le xxxx, (Rapporteur)
54. Martin Gleize, Thèse de l'Université Paris-Saclay, soutenue le 7 janvier 2016 (Président)
55. Marcus Freitag, Phd Université Aix-la-Chapelle, soutenue le mercredi 6 avril 2016, (Rapporteur)
56. Benjamin Marie, Thèse de l'Université Paris-Saclay, soutenue le 23 mars 2016 (Président)
57. Jérémie Sublime, Thèse de l'Université Paris-Saclay, soutenue le 09 novembre 2016 (Président)
58. Melissa Ailem, Thèse de l'Université Paris Descartes, soutenue le 18 novembre 2016,
59. Patrick Lehnen, Phd Université Aix-la-Chapelle, soutenue le 17 mai 2017, (Rapporteur)
60. Mohamed Hadjadj, Thèse de l'Université Paris-Saclay, soutenue le 17 novembre 2017 (Président)
61. Yohann Dupont, Thèse de l'Université Paris 3, soutenue le 23 novembre 2017 (Rapporteur)
62. Ekatarina Garmasch, Thèse de l'Université d'Amsterdam, soutenue le 12 décembre 2017 (Rapporteur)
63. Alexandre Bérard, Thèse de l'Université de Grenoble, soutenue le 15 juin 2018 (Examinateur)
64. Gabriele Marzinotto, Thèse de Aix-Marseille Université, soutenue le 13 décembre 2019 (Rapporteur)
65. Alexis Conneau, Thèse de l'Université du Mans, soutenue le 20 mai 2019 (Président, Rapporteur)
66. Anissa Hamza, Thèse de l'Université de Strasbourg, soutenue le 20 septembre 2019 (Examinateur)
67. Guillaume Lample, Thèse de Sorbonne Université, soutenue le 17 octobre 2019 (Rapporteur)

68. Zheng Zhang, Thèse de l'Université Paris Saclay, soutenue le 18 octobre 2020 (Président)
69. Maha Elbayad, Thèse de l'Université de Grenoble Alpes, soutenue le 22 juin 2020 (Président)
70. Valentin Belissen, Thèse de l'Université Paris Saclay, soutenue le 12 novembre 2020 (Président)
71. Hicham El Boukkouri, Thèse de l'Université Paris Saclay, soutenue le 18 novembre 2021 (Président)
72. Philip Düfter, Thèse de l'Université Ludvig-Maximilian de Munich, soutenue le 28 avril 2021 (Rapporteur)
73. Émile Chapuis ,Thèse de l'Institut Polytechnique de Paris, soutenue le 15 décembre 2021 (Rapporteur)
74. Franck Dary, Thèse de Aix-Marseille Université, soutenue le 12 juillet 2022 (Rapporteur)
75. Betty Fabre, Thèse de l'Université de Rennes 1, soutenue le 16 septembre 2022 (Examinateur)
76. Marion Kaczmarek, Thèse de l'Université Paris-Saclay, soutenue le 26 septembre 2022 (Président)
77. Élie Azeraf ,Thèse de l'Institut Polytechnique de Paris, soutenue le 2 octobre 2022 (Examinateur)
78. Léo Laugier, Thèse de l'Institut Polytechnique de Paris, soutenue le 8 novembre 2022 (Rapporteur)

**Jurys d'Habilitation à Diriger des Recherches**

1. Hélène Bonnaud-Maynard, Habilitation de l'Université Paris-Sud, soutenue le 9 décembre 2009 (Garant)
2. Matthieu Constant, Habilitation à Diriger des Recherches de l'Université Paris-Est, soutenue le 3 décembre 2012 (Rapporteur)
3. Ludovic Tanguy, Habilitation à Diriger des Recherches de l'Université de Toulouse 3, soutenue le 11 septembre 2012 (Rapporteur)
4. Alexandre Allauzen Habilitation à Diriger des Recherches de l'Université de Paris-Sud, soutenue le 30 janvier 2014 (Garant)
5. Benoît Crabbé, Habilitation à Diriger des Recherches de l'Université Paris Diderot, soutenue le 1er décembre 2017 (Rapporteur)
6. Natalia Grabar, Habilitation de l'Université Paris-Sud, soutenue le 9 décembre 2019 (Garant)

# C Missions d'expertise