Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM).

Rachel Panckhurst

UNIVERSITÉ PARIS-EST

HABILITATION À DIRIGER DES RECHERCHES

présentée par

Rachel Panckhurst

soutenue le : 30 mai 2017

Discipline/Spécialité : linguistique-informatique

Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM).

Volume I. — Synthèse

Jury

ANTONIADIS Georges (Examinateur), Professeur, Université Grenoble-Alpes
FAIRON Cédrick (Rapporteur), Professeur, Université catholique de Louvain (Belgique)
KRSTEV Cvetana (Examinatrice), Professeure, Université de Belgrade (Serbie)
KYRIACOPOULOU Panayota (Directrice), Professeure, Université Paris-Est Marne-la-Vallée
LAPORTE Éric (Rapporteur), Professeur, Université Paris-Est Marne-la-Vallée
MOÏSE Claudine (Examinatrice), Professeure, Université Grenoble-Alpes
ROCHE Mathieu (Examinateur), Chercheur HDR, Cirad, Montpellier
SEGOND Frédérique (Rapporteure), Directrice du centre de R&D Viseo, Grenoble ; PAST à l'INALCO, Paris

For Saji,
Whatever path you choose to follow,
it will be the right one,
for you will have chosen it yourself.

Sommaire

Remerciements

Prendre la décision de rédiger mon habilitation à diriger des recherches a duré de très nombreuses années. Pendant assez longtemps, j'étais même certaine que je ne la ferais jamais ; il y avait toujours d'autres priorités. Je ne sais pas exactement pour quelle raison le déclic a eu lieu. Une accumulation, et un moment propice, sans doute. Le besoin de se donner un nouveau projet d'écriture, aussi, entre autres facteurs. En tout état de cause, je remercie toutes les personnes, nombreuses, dont la liste est trop longue pour que je les énumère de manière exhaustive : amis, collègues, étudiants, qui ont insisté jusqu'au jour où j'ai enfin décidé de tenter l'aventure, plus de vingt-cinq ans après avoir terminé mon doctorat.

Sur le plan institutionnel, je suis très reconnaissante envers le président de la COMUE Université Paris-Est, Philippe Tchamitchian, de m'avoir autorisée à présenter mon habilitation à diriger des recherches au sein de son établissement.

J'ai été très émue que Panayota Kyriacopoulou, professeur de linguistique-informatique à l'université Paris-Est Marne-la-Vallée, accepte d'être ma directrice. Nous avons commencé nos études ensemble, à Clermont-Ferrand, dans les années 1980, puis nos chemins ont bifurqué, pour un temps. Tita est « monté à Paris » faire sa thèse sous la direction de Maurice Gross. Nous nous sommes retrouvées au colloque de l'*Acfas* à Sherbrooke en mai 2011, autour des SMS, et ensuite à Montpellier, puis à Alicante. Elle m'a d'ailleurs tout de suite demandé pourquoi je n'avais pas encore soutenu mon HDR. *Ευχαριστώ πολύ* Tita, pour

cette interrogation, et, surtout, d'avoir accepté de me guider. Tes conseils avisés et ton accompagnement assidu tout au long de cette aventure — alors que je te savais très occupée — ont été déterminants pour moi.

Je remercie très chaleureusement Georges Antoniadis, Cédrick Fairon, Cvetana Krstev, Panayota Kyriacopoulou, Éric Laporte, Claudine Moïse, Mathieu Roche, Frédérique Segond, de m'avoir honorée par leur acceptation enthousiaste de siéger au sein de mon jury.

Je remercie aussi les personnes qui ont hésité ou qui hésitent à leur tour à se lancer dans cette rédaction — qui se reconnaîtront — et avec lesquelles j'ai eu de nombreuses discussions fécondes.

Des doctorants et leurs directeurs de recherche en France, en Belgique et en Suisse, m'ont sollicitée ces dernières années pour les co-encadrer en thèse ou pour être membre de leur jury de soutenance. Je leur en suis reconnaissante, car leur détermination et leur confiance m'ont (à nouveau) fait réfléchir et peut-être infléchir. L'exigence envers moi-même m'a toujours amenée à refuser tant que je n'avais pas soutenu mon habilitation. C'était une question de principe et un respect des règles nationales. Cependant, en 2016-2017 j'ai commencé à assouplir ce principe, puisque j'avais démarré l'aventure de la réflexion et de la rédaction. J'ai donc accepté de participer au jury de soutenance de doctorat de Frédéric André à l'université Paris-Sorbonne, le 24 avril 2017.

Puis, il y a celles et ceux que je voudrais remercier nominativement — qui doucement, tranquillement, inlassablement, revenaient de temps à autre « à la charge ».

Je pense en premier lieu à Claudine Moïse, qui me racontait, devant l'école primaire de nos enfants, il y a fort longtemps, le bonheur de son congé de recherches pendant lequel elle a commencé le projet d'écriture de son habilitation. Sans être pesante, Claudine n'a jamais renoncé à me convaincre. La réflexion menée pendant notre brunch et la promenade au bord de la mer à Palavas, le lundi de Pâques 2016, a constitué un moment clef.

Myriam Maréchal, lorsque je me rendais dans son bureau de secrétariat des sciences du langage, à l'université Paul-Valéry Montpellier 3, insistait de manière doucement appuyée, ces dernières années ; le point de vue des étudiants, qui

pourraient, si je poursuivais, me solliciter pour leur encadrement doctoral. Tout en s'excusant de se répéter, elle m'incitait tranquillement à le faire.

Je pense aussi à Ksenija Djordjević Léonard, qui m'a annoncé un jour — lorsqu'on s'est donné rendez-vous dans un café de Montpellier où ils font un excellent cheesecake — qu'elle avait commencé à rédiger son HDR. Elle m'a fait réfléchir à nouveau sur ma position. Depuis ce jour, nous nous retrouvons régulièrement pour échanger autour de nos états d'avancement et, en parallèle, déguster les différents cheesecakes, en quête du meilleur de la ville.

Mathieu Roche a su laisser de côté les courriels, à des moments propices, pour m'inviter à prendre un café et discuter de nos projets scientifiques, avec beaucoup de discrétion et de finesse. Cela tombait toujours à pic. Je me souviens, notamment, d'un café pris en face de *France Inter*, fin mars 2016, lorsque nous attendions notre direct dans l'émission de Mathieu Vidard, *La tête au carré*. Cette courte discussion a été également déterminante pour moi.

À la même période, j'ai voyagé en train avec Laurence Vincent-Durroux, en direction de Lyon pour une conférence sur les SMS à l'ENS. De manière très posée et lucide, elle m'a suggéré de faire mon habilitation, tout en réfléchissant pour m'aider à tirer les fils de mes projets scientifiques successifs.

Je remercie de tout cœur *ma dream team* du projet SMS (*sud4science/88milSMS*). Depuis 2011, je partage des moments de recherche privilégiés avec Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche et Bertrand Verine. C'est un pur bonheur, toujours renouvelé, et sans conteste la collaboration scientifique la plus épanouissante de ma carrière — en tout cas, jusqu'à présent. Dès ma prise de décision concernant l'habilitation, ils m'ont fortement encouragée et soutenue. *Long live nos échanges*.

Cédrick Fairon et son équipe du CENTAL (Centre du traitement automatique du langage, université catholique de Louvain, UCL), notamment Louise-Amélie Cougnon — avec l'aide technique de Hubert Naets, également — ont constitué une source d'inspiration cruciale, car sans leur projet *SMS4science*, le nôtre n'aurait sans doute jamais vu le jour.

J'ai effectué mon doctorat sous la direction de Gabriel G. Bès (malheureusement décédé en octobre 2013), dont je retiens — entre autres — des enseignements inestimables sur le plan de la rigueur scientifique.

Mes séjours de recherche à Montréal, Paris et Sydney, m'ont permis de rencontrer d'autres personnes particulièrement importantes pour mon cheminement scientifique, parmi lesquelles, par ordre chronologique des rencontres : Jacqueline Léon, Jean-Marie Marandin, Bernard Fradin, Pierre Plante, Claude Ricciardi Rigault, Sophie David, Muriel Amar, Shirley Alexander, entre autres. Les échanges fructueux ont été déterminants et ont fortement contribué à façonner le regard que je porte sur les sciences du langage, le traitement automatique du langage naturel, les technologies de l'information et de la communication éducatives et la formation ouverte et à distance.

Mon université de rattachement, Paul-Valéry Montpellier 3, mon laboratoire de recherche, Praxiling, UMR 5267 CNRS, dirigé par Agnès Steuckardt, et d'autres établissements (*Maison des Sciences de l'Homme de Montpellier*, MSH-M, la *Délégation générale à la langue française et aux langues de France*, DGLFLF, le *Laboratoire d'informatique, de robotique et de microélectronique de Montpellier*, LIRMM, le *Centre de coopération internationale en recherche agronomique pour le développement*, CIRAD) m'ont aidée par leur soutien. Je remercie aussi Sabine Cotreaux (ingénieur et responsable du service Partenariat Recherche), d'avoir eu l'excellente idée de me suggérer de répondre à l'appel de financement MSH-M, fin 2010, à un moment où j'ai commencé à douter de la possibilité d'un financement pour démarrer notre projet SMS.

À Gilles Pérez, Myriam Rivoire, Florian Pascual, et les autres camarades, pour les moments de partage syndical, toujours dans la bonne humeur. Ils m'ont appris à regarder l'université d'un autre point de vue, celui des personnels techniques et administratifs.

À celles et à ceux avec qui j'ai partagé des moments forts de coordination ou d'écriture d'ouvrages en tant que co-auteurs ou co-éditeurs ces deux dernières décennies. Je pense notamment à Sophie David, Gilles Pérez, Laurence Vincent-Durroux, Lisa Whistlecroft, ainsi qu'à Daniel Savey, qui a écrit la préface pour l'ouvrage dont Gilles Pérez et moi-même sommes co-auteurs : *Introduction aux technologies et de la communication éducatives*, en 2000. Mes co-auteurs d'articles,

plus nombreux ces dernières années, m'ont également fait avancer dans mes réflexions scientifiques, et je leur en suis reconnaissante.

Pour les conseils et la relecture, je remercie ma directrice, Tita Kyriacopoulou, ainsi que Ksenija Djordjević Léonard et Gilles Pérez. Gilles, qui *m'espante* toujours, a apporté ses connaissances plus qu'utiles en typographie et en LaTeX tout au long de l'élaboration et de la finalisation de ce document. *Kiel mi iam povas sufiĉe danki vin ?*

I would like to thank my parents, Dr Fay and John Panckhurst, for their everlasting encouragement, enthusiasm and support, notwithstanding! My sister, Helen, her partner Kevin, my brother, Michael, his partner Clare, and other family members have always been there for me. *Kia ora*. Many thanks to my son Saji, who actively participated in my previous (non-academic) project, a cookbook: *Fait Maison. Recipes from a Kiwi in France*, (2014). He is always a wonderful source of inspiration and has tolerated my many hours at the desk writing up this most recent challenge while he was preparing his international-option French *baccalauréat*.

Préface. Destination *La France*

This is a record of my professional research life — spanning over 25 years. It is most likely unusual to start one's *habilitation* by a personal note, but as (socio)linguists would remark, the co(n)text is meaningful for any research. In this instance, I think it important to briefly retrace my steps, the path leading from New Zealand via London to France. After all, "Home is where one starts from" (T.S. Eliot, *East Coker*). I have written this preface in English — *avec quelques incursions en français, parsemées ici et là*, since it will hence convey, I hope, my transition from the English language to the French language, and my permanent everyday journey/quest as a "false bilingual":

> Il y a bilingues et bilingues. Les vrais et les faux. Les vrais sont ceux qui [...] apprennent dès l'enfance à maîtriser deux langues à la perfection et passent de l'une à l'autre sans état d'âme particulier. [...] Les mots le disent bien: la première langue, la "maternelle", acquise dès la prime enfance, vous enveloppe et vous fait sienne, alors que pour la deuxième, l'"adoptive", c'est vous qui devez la materner, la maîtriser, vous l'approprier. (Huston 1999, p. 53, p. 61).

Life is full of encounters. My first taste of France came in 1975 when my parents obtained a six-month sabbatical to London. The highlight of our stay was a six-week campervan tour of Europe, including France. I was 14. My diary of that time contains a lot of detail on food... Back in Wellington, French was covered in my secondary school curriculum at Onslow College, but even though the teachers did their best to enthuse us, our oral French remained dismal. In my sixth-form

report, my French teacher, Bernadette Rundle, wrote, "She should try to get more practice with spoken French." Having dreamed for several years about living in France, I had no qualms about taking her up on full French immersion a few years later.

After having obtained the New Zealand *University Entrance* (official equivalent to the French *baccalauréat*) in 1977, at age 16, and working for almost a year at *Pacific and Orient* (*P&O*) as a "container controller", in central Wellington, I found myself back in London working in the City (at *Overseas Containers Ltd*.) and living with my parents, Fay and John. There I met Patricia, from Cognac in south-west France, who had been in London for five years. We often had lunch together. One day, Patricia asked me, "Do you speak French?" followed by a few questions just to check. Her conclusion was *sans appel:* "you don't speak French yet." That and the school report were enough to push me over the channel armed for new experiences. I was ready to go to France and take a French language course. Patricia helped me choose the destination: "Go to the south. They have a wonderful accent there!" The month's intensive course was to take place at the *Céravum* (*Centre d'études et de recherches audio-visuelles d'universitaires à Montpellier*), linked to one of Montpellier's three universities, *université Paul-Valéry Montpellier 3*. The plan was vague: go for the month of September (1979) for my *stage d'apprentissage de la langue française par les méthodes audio-visuelles*, then see what I wanted to do: move back in with my parents in London and find a job there, or stay in France. For me, of course, having dreamed of living in France since adolescence, the idea of extending my time there was more enticing.

I went to Montpellier. After a month of intensive French, I was fairly fluent — at least for basic conversation (mastering a language takes decades!); don't forget that most French people didn't speak English in those days — and decided I needed to stay longer. I moved to Clermont-Ferrand and continued to learn French at the *Service Interuniversitaire des Étudiants Étrangers* (S.I.E.E.), which was linked to the *université Blaise-Pascal Clermont 2*. I would be there for two years — during which time I obtained the *Diplôme d'études françaises, Certificat Pratique de Langue française*, then the *Diplôme supérieur d'études françaises* (equivalent to the first year at University). For the 3-hour civilisation final exam, there was a choice between "Le gaullisme: essai de définition" and "La civilisation médiévale: montrez ses apports durables à la civilisation française". One may not realise,

but for a foreigner who has only lived in France for 1 ½ years, without having grown up immersed in French language, culture and civilisation, these sorts of typically *franco-français* exams were very difficult indeed, even if they were supposedly designed for foreigners.

I then moved into a university course in French literature, and on to a fully-fledged French degree (*DEUG*, then *Licence*) in *Lettres Modernes, mention FLE*, two years later.

In the meantime, I had learnt quite a bit about French culture and *soirées*, usually gathering together 15 or 20 different nationalities. So began my own melting-pot version of French and international culture.

(Partial extracts from Panckhurst, 2014, *Fait Maison. Recipes from a Kiwi in France*, p. 5-9).

I had originally come to France for one month, and I have been here now for over 35 years.

1

Présentation de mon parcours

> I want to be all that I am capable of becoming.
>
> *Katherine Mansfield*

1.1 Parcours initial. *De Vive Voix* jusqu'au doctorat

Après mon *stage d'apprentissage de la langue française par les méthodes audio-visuelles*, au Céravum de Montpellier, puis deux ans d'apprentissage plus approfondi de langue française au *Service Interuniversitaire des Étudiants Étrangers* (S.I.E.E.), rattaché à l'université Blaise-Pascal Clermont 2, j'ai opté pour une *immersion totale* en français, avec les français, *à la fac*. Je me suis inscrite en DEUG de *Lettres Modernes*. Grâce à l'obtention de mon *Diplôme supérieur d'études françaises*, l'accès en deuxième année du cursus était (presque) de droit ; il fallait néanmoins passer un examen d'entrée à l'université réservée exclusivement aux étudiants étrangers n'ayant pas en leur possession l'équivalent du baccalauréat français. Peu de temps après cette réussite, j'ai appris — en lisant le courrier à l'envers sur le bureau de la secrétaire du S.I.E.E. — que mon « University Entrance » néo-zélandais était *de facto* officiellement reconnu équivalent au baccalauréat français [1], et ainsi j'aurais pu m'éviter cette épreuve supplémentaire… C'était le

1. Depuis le 28 novembre 2013, cette équivalence du baccalauréat s'est étendue, grâce à un accord de coopération universitaire signé entre la conférence des présidents d'univer-

début d'un long parcours universitaire en tant qu'étrangère, dans un système où j'avais tout à apprendre, *to say the least*.

Le DEUG en poche, je me suis attaquée à la licence de *Lettres Modernes*. Cette année-là fut la plus difficile de ma carrière estudiantine. En deux ans, j'étais passée de la méthode audio-visuelle par excellence, *De Vive Voix*, avec Pierre et Mireille (prénom excessivement difficile à prononcer pour des anglophones débutants), avec laquelle j'avais appris le français à Montpellier durant le mois de septembre 1979, à des enseignements un peu plus austères. La *phonétique historique* (j'ai gardé le manuel de l'époque en guise de souvenir), la *littérature française* en commençant par l'ancien français ; Béroul, Chrétien de Troyes, puis Du Bellay, Ronsard, Pascal, La Fontaine, Marivaux, Diderot, Michelet, Flaubert, Sand, Baudelaire, Mallarmé, Mauriac, Giono, Michaux, Ionesco, Beckett, Robbe-Grillet — pour ne citer que ceux dont je me souviens — étaient mes compagnons de chevet au quotidien. Si je connaissais bien quelques auteurs français contemporains avant d'arriver en France, grâce à mes lectures d'adolescente (Camus, Sartre, de Beauvoir, etc.) j'ai dû, tant bien que mal, tenter de rattraper mes lacunes pour les siècles précédents — je révisais mes examens avec une amie, étrangère comme moi, avec des pauses pour regarder les matchs de tennis de Roland Garros à la télévision ! Le seul examen que j'ai dû repasser pendant mes études supérieures à l'université française, était le certificat de littérature française. Il me manquait 1,5 point sur un total de 160. Déterminée, obstinée, butée, l'année universitaire suivante j'ai écrit 27 dissertations afin de m'entraîner convenablement, avant de le repasser, tout en continuant en maîtrise. Je me souviens surtout d'une de ces dissertations, « La folie dans *Les Pensées* de Pascal », que j'ai refaite à deux reprises. Mon entourage m'a beaucoup aidée dans cette démarche d'écriture. Une de mes amies françaises — d'ailleurs, nous sommes toujours d'excellentes amies ; après tout, les déboires avec la littérature française, cela crée des liens ! — m'a avoué quelques années plus tard, qu'elle s'était énormément inquiétée pour moi, à l'époque, tant j'avais des lacunes en la matière. Par conséquent, j'ai été

sité (CPU) et « Universites New Zealand » (l'équivalent de la CPU en Nouvelle-Zélande. Il existe désormais une reconnaissance mutuelle entre la France et la Nouvelle-Zélande pour les diplômes de licence/bachelor, master/master, doctorat/PhD : http://www.cpu.fr/actualite/reconnaissance-des-diplomes-entre-la-france-et-la-nouvelle-zelande/ (consulté le 8 janvier 2017).

très fière (et soulagée !) d'obtenir mon certificat de littérature française à la fin de l'année universitaire.

À 20 ans, en 1981, dans le cadre du DEUG, j'avais suivi un cours optionnel d'initiation à la linguistique. J'étais séduite. C'était le début d'un choix de parcours, d'une véritable vocation, sans que je ne m'en rende compte sur le moment. C'est en licence, puis en maîtrise, que j'ai compris que les Sciences du Langage m'attiraient vraiment. L'université Blaise-Pascal à Clermont-Ferrand ne disposait pas encore d'un département de linguistique à part entière, donc le diplôme était rattaché au département de Lettres Modernes, mais la spécialisation en linguistique était possible. Je me suis épanouie en faisant mon premier travail de recherche, un mémoire de maîtrise intitulé « La place de la communication non verbale dans l'enseignement et l'apprentissage d'une langue étrangère ». Mon expérience estivale d'enseignante d'anglais au CAVILAM [2] m'avait permis de rencontrer l'équipe pédagogique et de l'associer à mon travail de recherche. J'ai obtenu l'autorisation de filmer et d'observer trois enseignants et leurs apprenants en situation de classe d'anglais langue étrangère pour adultes francophones. Par la suite, j'ai segmenté et décortiqué les séquences vidéo afin d'analyser la communication non verbale (CNV) utilisée : la gestualité — avec les termes de l'époque — *illustrative* : emblèmes, co-verbaux, déictiques, *interactionnelle*, *quasi-linguistique*, *expressive*, *extra-communicative* ; les expressions faciales ; le regard ; la posture ; la proximité ; les aspects para/non-verbaux du discours, etc. Cette analyse et le questionnaire associé m'ont permis de mieux cerner les fonctionnements de la CNV en situation pédagogique et par la suite de sensibiliser les enseignants et les apprenants à ces pratiques. Nous étions en pleine prise de conscience de la « compétence de communication » en didactique des langues étrangères. Ce mémoire avait été encadré par Max Dany, auteur de méthodes didactiques et directeur du CAVILAM, mais officiellement dirigé par un professeur de l'université Blaise-Pascal Clermont 2, Gabriel G. Bès.

Ma rencontre avec Gabriel G. Bès fut tout à fait déterminante pour la suite de mes études supérieures et pour l'avenir de ma vie professionnelle universitaire. Un véritable mentor.

2. Initialement, « Centre Audio-Visuel de Langues Modernes », désormais « Centre d'Approches Vivantes des Langues et des Médias ». Le CAVILAM, situé à Vichy, a rejoint le réseau international de l'Alliance Française en 2012.

> Dans un domaine des sciences du langage souvent cloisonné, on rencontre avec Bès un homme de science qui, en quarante ans de carrière, a à la fois côtoyé la linguistique fonctionnelle de Martinet et les nouvelles technologies, tout en restant toujours indépendant des doctrines. En la matière, la seule affiliation qu'on peut lui reconnaître est celle de la méthode : la linguistique est une science empirique, où l'on se doit d'expliciter clairement et rigoureusement les observations faites sur le réel, au même titre que les hypothèses. (Trouilleux 2015, p. 14).

La combinaison de recherche fondamentale théorique et de recherche appliquée à l'aide de *données authentiques* est une dimension récurrente, que j'estime très importante, depuis mon premier mémoire universitaire.

Mon mémoire de *D.E.A. de Linguistique et Informatique*, soutenu en octobre 1985, intitulé : *Les formes interrogatives en français* constituait un changement thématique fondamental par rapport à mon mémoire de maîtrise. Le sujet me plaisait énormément, car j'ai toujours aimé la grammaire, la description, les explicitations, les typologies, mais il était quasi imposé pour une raison très précise : Gabriel G. Bès venait d'obtenir un financement pour un projet européen (ESPRIT Project 393), le projet ACORD, *Construction and Interrogation of Knowledge Bases Using Natural Language Text and Graphics*. Celui-ci rassemblait des équipes d'universitaires et d'industriels de trois pays européens et l'objectif principal était d'élaborer une interface homme-machine permettant de construire et d'interroger une base de connaissances commune en français/anglais/allemand. Très novateur pour l'époque. Pour mener à bien la tâche de l'équipe clermontoise, à savoir développer une grammaire du français, à utiliser dans le cadre du projet ACORD, Gabriel G. Bès avait besoin d'étudiants. Karine Baschung (1990) et moi-même avons été ses premières doctorantes, au sein de la formation doctorale *Linguistique et informatique*, de l'université Blaise-Pascal Clermont 2. Plus tard, son équipe est devenue le *Groupe de recherche dans les industries de la langue (GRIL)*.

Le mémoire de D.E.A. consistait en une description et un « codage » de l'épineux problème des structures interrogatives en français. L'idée était de poursuivre en thèse afin d'aboutir à « une indexation assistée par ordinateur en vue d'élaborer un fichier informatique de structures interrogatives linguistiques » (Panckhurst,

1985, « Les formes interrogatives du français », Mémoire de DEA en Linguistique en Informatique, université Blaise-Pascal Clermont 2, p. 93.).

Aussitôt inscrite en thèse, j'ai bénéficié d'une allocation de recherche D.G.R.S.T. (*Direction Générale de la Recherche Scientifique et Technologique*), dans la Formation Doctorale *Linguistique et Informatique* de l'université Blaise-Pascal Clermont 2, sous la direction de Gabriel G. Bès. Il m'a mise en situation professionnelle très exigeante dès le début, en me confiant diverses responsabilités administratives-techniques, pédagogiques et de recherche : j'ai été chargée de cours dans le cadre de la *Maîtrise de français langue étrangère* (FLE), de 1985 à 1988 ; j'ai participé aux réunions du projet européen ACORD, et j'ai écrit 51 pages à propos des structures interrogatives en français, dès juillet 1986, pour un rapport officiel du projet, intitulé « Contextual phenomena in dialogue » (Baschung et al. 1986), alors que j'étais en première année de thèse.

Formée à la programmation en PROLOG depuis le D.E.A., mon horizon s'est élargi lorsque j'ai rencontré l'équipe parisienne du LISH (*Laboratoire d'Informatique pour les Sciences Humaines*) et de l'INALF (*Institut national de la langue française*), notamment, Jacqueline Léon, Jean-Marie Marandin et Bernard Fradin, alors chercheurs au CNRS. Ils travaillaient en coopération avec le *Centre d'Analyse de Textes par Ordinateur* (Centre d'ATO) de l'*université du Québec à Montréal* (UQÀM), avec Pierre Plante, l'auteur du langage DEREDEC (Plante 1979), écrit en LISP. Jacqueline Léon nous a appris à programmer en DEREDEC, à la MSH, Boulevard Raspail, à Paris, en 1986-1987. Ma curiosité était piquée ; je voulais en savoir plus. Je souhaitais approfondir ces enseignements, afin d'envisager une implémentation informatique sous forme d'analyse, d'indexation et de classification des structures interrogatives du français, en y ajoutant une dimension de consultation de bases de données.

En 1987-1988, j'ai décidé de me porter candidate étudiante au Conseil scientifique de l'université. J'ai été élue pour un mandat de deux ans. En début de troisième année de thèse, mon allocation de recherche DGRST était désormais terminée. J'aurais pu bénéficier d'une prolongation pour une année, mais j'ai préféré accepter un poste *d'assistante associée* (l'équivalent de l'ATER en ces années-là) à plein temps — peut-être pour sortir un petit peu du laboratoire de recherche aussi, où je travaillais au moins 40 heures par semaine. Parallèlement, j'ai effectué une

demande de bourse dans le cadre du projet de coopération franco-québécoise : *Conception et application d'un analyseur lexico-syntaxique du français*, afin de me rendre au Centre d'ATO, à l'UQÀM. Le stage a pu s'organiser grâce aux liens établis précédemment avec l'équipe du LISH/INALF. J'ai obtenu une bourse pour six mois, et je me suis envolée pour le Québec en septembre 1988 — juste après avoir déposé ma demande de naturalisation française, *but that is another story*. Finalement, j'ai prolongé mon séjour montréalais jusqu'en mai 1989.

À la fin des années 1980, le Centre d'ATO était l'un des lieux à la pointe en linguistique computationnelle, tout comme « le pôle high-tech » à Clermont-Ferrand :

> Le 10 octobre 1990, *Le Monde* publie dans son supplément *Campus* un article sur le DEA « Linguistique et informatique » de Clermont-Ferrand. Un chercheur des Laboratoires de Marcoussis y déclare « l'idée d'un pôle high-tech à Clermont-Ferrand, cela m'a paru surprenant » ; mais c'était bien réel : l'équipe du GRIL était à la pointe de la recherche en linguistique computationnelle.
>
> (Trouilleux 2015, p. 25).

Arrivée à Montréal, le langage Dérédec avait cédé sa place au langage FX (la programmation modulaire en « faisceaux » Plante, 1988-1996 (Plante 1990, 1991, 1993, 1996), écrit en Common Lisp, puis Le_Lisp [3]. Le principe central de ce langage était « l'autonomie conceptuelle », « basé sur le constat que le strict contrôle hiérarchique (ascendant ou descendant) impose au dispositif computationnel un canevas trop rigide qui vient contraindre la conceptualisation linguistique » (lettre de P. Plante à G.G. Bès, 7/11/1990). Afin de bien profiter de ma période de formation au Québec, je me suis rapidement familiarisée avec FX (et ses versions évolutives), au Centre d'ATO, et j'ai suivi un cours d'approfondissement de LISP avec des étudiants en informatique à l'UQÀM. Cette période a été extrêmement enrichissante ; non seulement j'ai indéniablement progressé en programmation — au point où j'étais confiante que j'allais pouvoir élaborer mon progiciel informatisé pour le traitement de mes données de thèse — mais les rencontres avec d'autres stagiaires et collègues de l'équipe de *Recherche et Développement en Linguistique Computationnelle* (RDLC), au Centre d'ATO, ont été très fructueuses. J'ai

3. Le_Lisp (Chailloux, Inria).

eu de nombreuses discussions scientifiques fécondes avec Pierre Plante, Sophie David et Claude Ricciardi-Rigault, également professeur à la Télé-université du Québec. C'est grâce à Claude Ricciardi-Rigault qu'après mon retour à Clermont-Ferrand j'ai pu à nouveau bénéficier d'une mission de recherche en août 1989, permettant un approfondissement des aspects informatiques de ma thèse. Cette quatrième année de thèse outre-mer m'avait permis d'élaborer mon répertoire informatisé des structures interrogatives du français (RISIF) et une consultation de bases de données (CBD) en interface optionnelle avec RISIF.

Pendant ma dernière année de doctorat, j'ai rédigé la partie théorique et j'ai finalisé la programmation informatique, tout en étant chercheur au sein de l'École doctorale et en travaillant comme enseignante d'anglais au *Service Commun des Langues Vivantes* de l'université. J'ai également participé à un projet de recherche ponctuel au sein de *l'Association pour le développement de l'enseignement et la recherche* (ADER-Auvergne). J'ai été enfin *naturalisée française* en mai 1990, devenant ainsi binationale franco-néo-zélandaise — *but again, that is another long story*…

En 1990, mon répertoire descriptif et formel généralisé des structures interrogatives directes du français, relié à une consultation automatisée de base de données, constituant ainsi une recherche pluridisciplinaire (linguistique, informatique et documentation), était terminé. À présent, je pouvais songer à la soutenance. Celle-ci a eu lieu le 15 décembre 1990 (Panckhurst 1990). Outre la présentation et la discussion scientifiques avec les quatre membres du jury (Gabriel G. Bès, Francis Corblin, Geneviève Lallich-Boidin et Claude Ricciardi-Rigault), le tout ayant duré environ 3 heures, deux souvenirs demeurent.

Premièrement, à l'issue de mon exposé oral de soutenance, j'ai fait une démonstration informatique (chose très rare dans une université de Lettres et Sciences humaines à l'époque) de mon répertoire informatisé (RISIF) et la consultation de bases de données (CBD) sur un Macintosh SE 30. Au moment du chargement en mémoire vive, j'ai eu un « bug informatique » que je n'avais jamais eu auparavant, qui a empêché le bon fonctionnement de mon progiciel. J'ai aussitôt modifié une ligne de code — devant le jury et le public présents — et, par miracle, en « rebootant » le tout, j'ai pu mener à bien ma démonstration sans soucis supplémentaires. J'ai souvent repensé à ce moment, et aux conséquences éventuelles

(désastreuses ?) si je n'avais pas pu montrer mon travail informatique de manière convenable aux membres du jury.

Deuxièmement, l'une des questions du jury : « Comment envisageriez-vous la réfutation de vos propres hypothèses ? » J'y ai répondu, selon mon vague souvenir tant d'années après, *grosso modo*, dans un premier temps, en précisant que dans la mesure où je proposais une description volontairement en dehors d'un cadre grammatical ou d'un formalisme il me semblait que je n'avais pas réellement besoin de réfuter. Après tout, une autre description proposée par autrui, pourrait tout aussi bien convenir.

Cela étant, on peut, me semble-t-il, approfondir quelque peu la réponse à cette question concernant *la réfutation*, ou, précisément *l'absence de besoin de réfutation*, en reprenant les points évoqués au début de ma thèse, dont voici un extrait (Panckhurst 1990, p. 13-14) :

> Notre but en développant un répertoire dans le domaine de l'interrogation est analogue aux propos de Milner [...] : « La science du langage doit se proposer au minimum de construire une littéralisation qui permette la taxinomie la plus complète, la plus fine et la plus économique possible » (Milner 1989, p. 108). Le choix du mot *répertoire* est délibéré ; il signifie pour nous l'élaboration d'une *classification descriptive systématique* des faits caractérisant les interrogatives. Ce mot est, nous l'espérons, suffisamment neutre pour ne pas être rapproché d'une grammaire et par extension d'un *formalisme*. En effet, le répertoire que nous proposons de bâtir ne doit pas être confondu avec ces deux autres notions. Notre but est de fournir une *description* et non pas de construire une grammaire. Dans cette optique nous défendons l'idée qu'il est souhaitable d'envisager une exhibition de phénomènes sans avoir recours à l'utilisation d'un cadre de grammaire et donc de se situer en dehors d'un quelconque dispositif de reconnaissance prédéfini. L'idée en soi, au sens large n'est pas nouvelle. Gross 1975, p. 9, insistait sur le fait qu'un « long travail d'accumulation systématique de données » est nécessaire avant de procéder à l'étape de construction théorique. [...] Nous souhaitons encoder un certain savoir sur les structures interrogatives. Notre description en un répertoire formalisé doit « être un outil permettant de formuler des observations d'une manière uniforme » :
>
> We call descriptive metalanguage this formalised knowledge about the object language. [...] It is intended to present in an orderly and explicit manner observa-

> tions about natural language. Because it is formalised, it is a mathematised object on which inferences are possible (Bès et Jurie 1989, p. 10, p. 208).
>
> (Panckhurst 1990, p. 13-14).

Cette question autour de la réfutation m'est restée solidement ancrée en mémoire, depuis la soutenance de thèse, ainsi que celle concernant les descriptions, les typologies, etc. Un des enseignements que je retiens de Gabriel G. Bès est le suivant : *on peut faire des observations, des descriptions, et les formaliser de façon indépendante des théories* :

> L'œuvre de Bès alterne donc des exégèses de travaux majeurs, à travers le prisme de l'empiricité et de la formalisation, et des développements d'outils ou de descriptions linguistiques particulières qui ne prétendent jamais au statut de « théorie » mais visent à une meilleure compréhension des problèmes.
>
> (Trouilleux 2015, p. 18).

De même, avec l'apprentissage de la programmation, notamment en FX, j'avais poursuivi ma découverte de la notion clef d'autonomie, voire *d'autonomie conceptuelle*.

> Après l'analyse des *données authentiques* de mes premiers travaux universitaires, viennent en doctorat, les *observations*, les *descriptions*, le *questionnement* autour de la *réfutation, l'autonomie conceptuelle*. Ces thématiques constituent les premiers fils de ma « pelote scientifique » [4]

Ici s'achève mon parcours initial qui avait débuté avec *De Vive Voix* à Montpellier 11 ans plus tôt.

4. (Moïse 2012, p. 321-336).

1.2 L'habilitation. *Ready, set, go !*

> You must do the things you think you cannot do.
>
> *Eleanor Roosevelt*

Avant de me lancer dans l'aventure de cette réflexion-rédaction, j'ai voulu savoir quelles étaient les *règles* de l'exercice de l'habilitation. J'ai donc consulté les documents officiels, puis j'ai commencé à sonder mon entourage professionnel. Plusieurs collègues ont eu la gentillesse de m'envoyer leurs mémoires de HDR [5] pour que je m'en fasse une idée personnelle.

Arrêté du 23 novembre 1988 relatif à l'habilitation à diriger des recherches [6]

J'ai rapidement constaté qu'en dehors du cadre général (relativement vague, mais néanmoins ouvert) de l'article 4 de l'arrêté de 1988, rappelé en note, il y avait une multitude, une diversité immense concernant les types de dossiers de candidature, même au sein d'une même discipline. Cela n'est guère étonnant, dans la mesure où l'habilitation correspond finalement à un cheminement de réflexion personnelle, de ses propres travaux de recherche. Selon la direction que l'on souhaite emprunter, chaque dossier d'habilitation sera *de facto* unique, originale, tant par sa forme que par son contenu.

5. Les termes utilisés pour évoquer le mémoire par les uns et les autres m'ont amusée : « la chose », « la merveille »…

6. https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=9144E2A4220DC00E049ACF479488515D.tpdjo10v_3?cidTexte=JORFTEXT000000298904&dateTexte (consulté le 8 janvier 2017).

Article 1

L'habilitation à diriger des recherches sanctionne la reconnaissance du haut niveau scientifique du candidat, du caractère original de sa démarche dans un domaine de la science, de son aptitude à maîtriser une stratégie de recherche dans un domaine scientifique ou technologique suffisamment large et de sa capacité à encadrer de jeunes chercheurs.

Elle permet notamment d'être candidat à l'accès au corps des professeurs des universités.

[…]

Article 4

Le dossier de candidature comprend soit un ou plusieurs ouvrages publiés ou dactylographiés, soit un dossier de travaux, accompagnés d'une synthèse de l'activité scientifique du candidat permettant de faire apparaître son expérience dans l'animation d'une recherche.

Pourquoi avoir attendu plus de 25 ans entre le doctorat et l'habilitation à diriger des recherches ? La question se pose, obligatoirement, à cette étape, déjà bien avancée, de ma carrière. Comme je l'ai précisé supra, j'ai hésité pendant de longues années, car j'avais d'autres projets, qui me paraissaient plus urgents. Puis, je ne fais les choses que lorsque je me sens intellectuellement prête, lorsque cela s'impose à moi, quel que soit le délai. J'ai toujours fait ainsi [7]. Après la soutenance de ma thèse, l'un des membres du jury, Francis Corblin, m'a demandé si j'allais postuler à des postes de maîtres de conférence dès janvier 1991, et m'a suggéré fortement de le faire. Alors que j'ai effectivement immédiatement obtenu la qualification, j'ai pourtant décidé de ne pas emprunter ce chemin tout de suite. J'avais d'autres projets. Je voulais retourner au Centre d'ATO à Montréal. J'avais peut-être besoin d'une année de *césure*, aussi. En revanche, l'année suivante, j'ai souhaité le faire, et, en juillet 1992, à 30 ans, j'ai eu la chance d'obtenir le poste de maître de conférences en linguistique-informatique que j'occupe actuellement à l'université Paul-Valéry Montpellier 3.

Le lieu de l'habilitation est également un choix délibéré. L'université Paul-Valéry Montpellier 3 n'est dotée d'une structure d'encadrement professoral en linguistique-informatique au département des sciences du langage que depuis la rentrée 2016-2017, suite à un recrutement de PU [8]. Mon projet d'habilitation était déjà bien avancé, et j'avais pris contact avec ma directrice à l'université Paris-Est Marne-la-Vallée (UPEM) l'année universitaire précédente. J'ai souhaité présenter mon habilitation à la COMUE Université Paris-Est (UPE) pour deux raisons : 1) étant donné mon rattachement à la section CNU 7, le suivi par un professeur en linguistique-informatique de cette même section me paraissait primordial, et le rattachement à un département/institut d'informatique me semblait être un atout, voire un défi, intéressant ; 2) les recherches menées entre le CENTAL (Centre de Traitement Automatique des Langues) à l'UCL (université catholique de Louvain) en Belgique et l'UPEM autour de la constitution informatisée d'une base de données de SMS authentiques depuis 2004 ((Fairon et al. 2006b,c), (Fairon et Paumier 2006), Unitex/GramLab,

7. Peut-être est-ce héréditaire. Ma mère a soutenu son doctorat, quelques mois avant moi, en 1990.

8. La collègue recrutée, Agata Jackiewicz, informaticienne-linguiste, a néanmoins été recrutée sur un poste ayant un profil de sémantique et d'analyse de discours.

http://www-igm.univ-mlv.fr/~unitex/, (Paumier 2003)) sont en lien direct avec mes recherches actuelles.

Enfin, si *l'habilitation* doit permettre de vérifier ses capacités *à animer des recherches*, je ne saurais faire abstraction, dans le cadre de cet écrit, de toute la dimension *pédagogique*, que ce soit en formation initiale ou en formation continue, car je suis *enseignant-chercheur* [9]. La recherche est (ou devrait être, selon moi) inexorablement liée à l'enseignement. La recherche nourrit l'enseignement et inversement. Ce positionnement n'est pas obligatoirement partagé par tous, bien évidemment. Pour ma part, je ne peux faire abstraction de la façon dont la présentation de mes travaux de recherche est ressentie par les étudiants : tel ou tel aspect peut être alors approfondi, réorienté, remis en question, etc. suite à des moments de partage avec les étudiants [10]. Ceux que j'ai encadrés en recherche ont d'abord suivi mes enseignements (ou ont collaboré à des projets de recherche sous forme de stages). Nous nous sommes mutuellement choisis dans le positionnement d'encadrant/encadré à partir d'une première rencontre et d'une volonté d'aborder ou d'approfondir des sujets de recherche. Par ailleurs, la vie universitaire et les activités qui s'y déroulent sont cruciales ; je me suis souvent intéressée aux dimensions *administrative* et *technique* de mon établissement. Je voudrais donc faire part, ici, de l'ensemble des aspects qui composent ma vie professionnelle en tant qu'enseignant-chercheur, et qui sont indissociables, inextricables.

1.2.1 Volets de recherche

Depuis mon doctorat (Panckhurst 1990), et ma nomination en tant que maître de conférences à l'université Paul-Valéry Montpellier 3 (en octobre 1992), mes activités d'enseignement, d'administration et de recherche s'inscrivent dans le domaine du traitement automatique du langage/des langues (TAL), et, plus précisément, du traitement automatique du langage naturel écrit (TALNE). Trois

9. Dans le cadre de ce manuscrit, j'ai décidé d'utiliser le masculin afin d'éviter la surcharge du document.

10. L'exemple en date le plus récent est la modification de la typologie de l'écriture SMS (Panckhurst 2009), (Roche et al. 2016), suite à des discussions avec Frédéric André en 2014, alors qu'il était étudiant en M2 à l'université Paul-Valéry Montpellier 3 préparant son mémoire sous la double direction de mon collègue Fabrice Hirsch et moi-même (André 2014).

cheminements ou volets de recherche, traversent et s'imbriquent tout au long de mes 25 années de carrière universitaire, jusqu'à présent :

1. *prototypes et outils* (1991-2003) : interrogatives, verbes, gloses ;
2. *formation, (auto)évaluation, réseaux pédagogiques* (technologies de l'information et de la communication éducatives (TICE), eLearning/formation ouverte et à distance (FOAD)) (1996-2012) ;
3. communication médiée par ordinateur (CMO), discours électronique médié (DEM), discours numérique médié (DNM) : analyse de courriels, forums, chats, SMS (1996-2017).

Ceux-là seront explorés tout au long de ce manuscrit. La façon dont la recherche s'imprègne et s'enrichit de mes activités d'enseignement et d'administration, est je crois, cruciale. De ce fait, je me dois de présenter au lecteur l'ensemble de mes activités tripartites en tant qu'enseignant-chercheur, afin qu'il puisse mieux entrevoir mon parcours global.

1.2.2 Organisation du manuscrit

À travers mon parcours, certes atypique — dans la mesure où je n'ai découvert l'espace francophone au quotidien qu'à partir de 18 ans — j'espère pouvoir montrer comment j'ai contribué en recherche (mais également en pédagogie et en administration), au domaine de la linguistique-informatique, et, par conséquent, de quelle manière j'estime être en mesure d'animer des recherches doctorales. Dans cette première section, j'ai brièvement dessiné mon parcours initial jusqu'au doctorat — afin de guider le lecteur à travers mes tout premiers pas de jeune chercheur. La deuxième section (*cf.* § 2) est consacrée à l'évocation de mes activités d'enseignement, de formation et mes responsabilités pédagogiques et administratives. Cela n'est peut-être pas habituel dans le cadre d'une habilitation, mais je m'octroie le droit de le faire, dans la mesure où je souhaite mettre en lumière leur importance pour moi et indiquer comment elles ont nourri ma réflexion en recherche. La troisième section (*cf.* § 3) constitue le « noyau dur » du manuscrit, la recherche. J'expliquerai comment j'ai tissé les fils des volets, des thématiques, comment j'ai tâtonné, bifurqué au gré des rencontres scientifiques. Puis, je montrerai aussi comment l'enseignement, l'administration et

la recherche s'imbriquent, de manière plus approfondie. Dans la quatrième et dernière section (*cf.* § 4), je mentionnerai les horizons et les perspectives à venir, avant de proposer une sélection globale de mes publications.

Cette habilitation est pour moi l'occasion de me poser un temps, pour réfléchir au quart de siècle précédent, pour apporter quelques touches personnelles, afin de montrer de quelle façon j'envisage la recherche. J'ai toujours prôné une recherche appliquée, et, j'espère, vivante, conduite avec, si possible, gaieté, et accessible. Le lecteur l'aura compris : j'accorde une importance suprême à la transmission. Étudiants, collègues, amis, membres de ma famille, le grand public, tous doivent être en mesure de comprendre l'utilité finale de la recherche conduite dans un contexte de service public. Tel est mon combat, mon défi, sans cesse renouvelé.

2

Enseignement, formation, responsabilités pédagogiques et administratives

> Whatever you decide to do for your future career, don't become a teacher.
>
> *Dr Fay and John Panckhurst*

Depuis l'âge de 18 ans, depuis mon arrivée en France, j'ai toujours enseigné, malgré les conseils teintés d'ironie rieuse de deux ex-universitaires spécialisés en psychologie éducative et en sciences de l'éducation, mes parents. C'était moins un choix délibéré, au départ, qu'un strict besoin, pour aider à financer mes études. J'ai longtemps enseigné l'anglais, à toutes les catégories d'âge — enfants en primaire le mercredi après-midi, adolescents pour des cours particuliers le soir, étudiants et stagiaires apprenants dans des centres de langues (CAVILAM à Vichy http://www.cavilam.com/, METAFORM à Clermont-Ferrand, http://www.metaform-langues.fr/) et à l'université Blaise-Pascal Clermont 2. Mon statut d'enseignant a également évolué d'un statut privé à un statut de contractuel, puis titulaire, dans la fonction publique : *enseignante vacataire*, *chargée de cours*, *assistante associée*, avant de devenir *maître de langue* (obligatoirement diplômé d'un(e) maîtrise/master, à la différence de *lectrice*, statut pour lequel la licence

suffisait, à l'époque) et enfin, *maître de conférences en linguistique-informatique*. Dès 1984 — année fatidique ? — huit ans avant ma nomination à mon poste actuel, et ayant un statut d'étudiant de 2e puis de 3e cycle, j'ai pratiqué le mixte enseignement-recherche, avec quelques ajouts administratifs, *into the bargain*.

Évoquer les aspects d'enseignement, de formation et d'administration dans le cadre de cette habilitation n'est pas sans intérêt pour mon cheminement en recherche, me semble-t-il. Cela me permettra d'évoquer, dans la troisième section de cette habilitation (*cf.* § 3), les liens, les réflexions, qui ont été nécessaires et qui ont pu être tissés et menés avec une recherche fondamentalement appliquée.

2.1 Enseignements

En tant qu'enseignant-chercheur au département de Sciences du Langage, je suis toujours intervenue en licence et en Master et ce, à tous les niveaux des diplômes (*cf. Activités d'enseignement et de formation* de mon curriculum vitae et la figure 2.1), avec une répartition globale de 70 % d'enseignement en licence, et 30 % en Master.

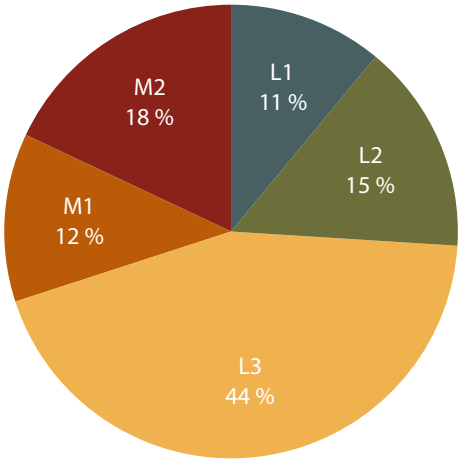


Figure 2.1 – Répartition des enseignements par niveau d'études *

* *J'ai harmonisé avec les appellations actuelles 3-5-8 : L1, L2, L3, M1, M2.*

Par ailleurs, j'interviens régulièrement dans d'autres départements de l'université. En début de carrière, dans le cadre de la maîtrise de documentation, pour un cours d'initiation au traitement automatique pour documentalistes, puis, au département mathématiques et informatique appliquées (MIAP), anciennement MASS, pour les enseignements spécifiques de compétences informatiques (compétences numériques (C2i)), et au sein d'un nouveau Master (ouvert récemment, 2016-2017) de mathématiques et informatique pour les sciences humaines et sociales (MIASHS), qui a pour vocation de former des *data scientists* ou ingénieurs *big data/data miner*. Je partage un cours [1] centré sur l'analyse de données textuelles. Par ailleurs, lorsque nos capacités d'accueil le permettent, nous rendons accessibles nos cours de TAL offerts au département des sciences du langage aux étudiants en provenance du département MIAP afin de créer un espace de dialogue et d'échange fructueux entre étudiants inscrits dans des diplômes distincts (actuellement, j'accueille quelques étudiants en L3 MIASHS dans le cadre de mon cours d'initiation à la linguistique-informatique, en licence de sciences du langage). En parallèle, et afin d'assurer une visibilité extérieure dans d'autres établissements, j'ai également enseigné en tant que chargée de cours au CRIM (*Centre de Recherche en Ingénierie Multilingue*, devenu plus récemment *Équipe de Recherche Textes, Informatique, Multilinguisme*), à l'INALCO (*Institut National des Langues et Civilisations Orientales*), à Paris, pendant les cinq premières années de ma carrière universitaire montpelliéraine (1992-1997). Il s'agissait d'effectuer une initiation au traitement automatique des langues, à raison de 21 heures annuelles, dans le cadre du D.E.S.S *Ingénierie multilingue* (responsable, Monique Slodzian).

Le ratio département/hors département est indiqué dans la figure 2.2.

Il me paraît fondamental que dès les premières années de licence, les étudiants puissent manier les outils informatiques élémentaires et comprendre les concepts de base en informatique. Mes enseignements en *Sciences du Langage*, en *Médiation Culturelle et Communication* (appellation devenue ensuite, *Communication, Médias, Médiations numériques*, CMM), de notre département, et en MIAP (C2i) s'inscrivent dans cette optique, et ils permettent notamment la préparation des matières qui ont été enseignées en licence de Sciences du Langage, mention

1. Avec mes collègues récemment nommés : Sasha Diwersy et Francesca Frontini.

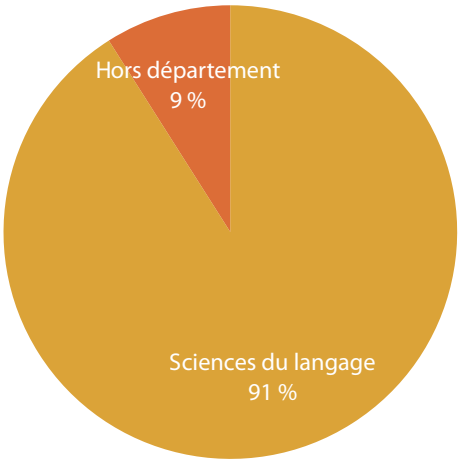


Figure 2.2 – Enseignements au département des Sciences du Langage/hors département (MIAP, Documentation, Inalco-Paris)

traitement automatique des langues (TAL), puis en maîtrise de Sciences du langage, mention *(industries de la langue (IDL)*) (*cf.* § 2.2) à la fois d'un point de vue théorique et pratique en s'inspirant de la notion de « texte ».

Le cours « Langue naturelle et informatique » (que j'ai assuré pendant 7 ans entre 1992 et 1999 en M2, avant l'obtention de nos mentions *Traitement automatique des langues* et *Industries de la langue* (L3 et M1), (*cf.* § 2.2) cherchait à sensibiliser les étudiants aux problématiques de recherche en linguistique-informatique et en *Industries de la langue* (reconnaissance de la parole/synthèse vocale, traduction assistée, sémantique lexicale, terminologie, interrogation de bases de données, correcteurs grammaticaux, etc.).

Le tableau 2.1 permet de visualiser les thématiques de l'ensemble de mes cours effectués depuis 1992 : Traitement automatique des langues (TAL), communication médiée par ordinateur (CMO), réseaux pédagogiques en eLearning/formation ouverte et à distance (FOAD), publication assistée par ordinateur (PAO), technologies de l'information et de la communication éducatives (TICE), compétences informatiques (C2i), (*cf.* § *Enseignement et formation* de mon Curriculum Vitæ (Volume III) pour les intitulés exacts).

Tableau 2.1 – Thématiques d'enseignement (1992-2017)*

Filières/départements : SL, CMM, MIAP **Thématiques : CMO/TAL, FOAD, PAO, C2i**	
Analyse de données textuelles (M1, à partir de 2016), MIASHS (MIAP) Données et corpus (CMO) (M1, à partir de 2015) Outils d'analyse des corpus écrits et oraux (M1, à partir de 2015) Langage, technologies et corpus (M1, jusqu'en 2015) Communication médiée par ordinateur (CMO/TAL) (M1, jusqu'en 2015) Linguistique-informatique (L3, SL et MIASHS) PAO et autopublication pour la médiation numérique (L2) Compétences informatiques (C2i) (L1, jusqu'en 2015)	2014-2017
Filières/départements : SL, CMM, MIAP **Thématiques : CMO/TAL, FOAD, PAO, C2i**	
Formation ouverte et à distance (FOAD), M2 Communication médiée par ordinateur (CMO/TAL), M1 Ingénierie éditoriale (PAO) (L3) Compétences informatiques (C2i) (L1, L2, jusqu'en 2012)	2011-2014
Filières/départements : SL, CMM, MIAP **Thématiques : CMO/TAL, FOAD, PAO, C2i** **Autre : Directrice d'études CMM, Enseignant référent, chargée de mission COMUE (anciennement PRES) (2009-2011)**	
Formation ouverte et à distance (FOAD) (M2) Communication médiée par ordinateur (CMO/TAL) (L3, M1) PAO (L3) Compétences informatiques (C2i) (L1, L2)	2008-2011
Filières/département : SL, CMM, MIAP **Thématiques : CMO, TAL, FOAD, PAO, C2i**	
Formation ouverte et à distance (FOAD) (M2) Communication médiée par ordinateur (CMO/TAL) (L3, M1) Linguistique-informatique (TAL) (L2) Édition, typographie, PAO (L3) Compétences informatiques (C2i) (L1, L2)	2004-2008
Filières : SL (Mentions TAL, IDL), CMM **Thématiques : TAL, IDL, TICE, PAO, bureautique** **Autre : formation en TICE pour les personnels de l'université (jusqu'en 2003)**	
Industries de la langue : théories et pratiques (M1, mention IDL) TAL et programmation pour linguistes (Prolog, Perl) (L3, mention TAL) Communication, nouvelles technologies éducatives, techniques documentaires (TICE) (L2, L3) Publication assistée par ordinateur (PAO) (L2) Bureautique (L1)	1999-2004
Filières/départements/institutions : SL, Ingénierie multilingue (Inalco) **Thématiques : TAL, TICE, bureautique** **Autre : formation en TICE pour les personnels de l'université (à partir de 2006)**	
Langue naturelle et informatique (programmation : FX, LISP) (M2) Initiation au traitement automatique du langage naturel (TALN) (L3, M2) Informatique et nouvelles technologies (TICE) (L2)	1995-1999

Bureautique (L1)	
Filières/départements/institutions : SL, Documentation, Ingénierie multilingue (Inalco) **Thématiques : TAL, bureautique**	
Langue naturelle et informatique (programmation : FX, LISP) (M2) Initiation au traitement automatique des langues (TAL) pour documentalistes (M1) Initiation au TAL(L3) Analyse du discours (L3) Bureautique (L1, L2)	1992-1995

*Une harmonisation des appellations a été effectuée : *Technologies de l'information et de la communication éducatives* (TICE) au lieu de *Nouvelles technologies éducatives* (NTÉ) ; *Communication, médias, médiations numériques* (CMM), correspond au plan quadriennal actuel — au lieu des appellations des plans précédents (MCC, communication, etc.). Pour les appellations exactes des cours, *cf. Volume III, Curriculum Vitæ*, rubrique « Enseignement et formation ». Le contenu des cours y est détaillé.

La figure 2.3 indique la répartition de ces enseignements.

Les thématiques indiquées ont bien sûr évolué en fonction, d'une part, des besoins de notre département et des autres départements/institutions, et d'autre part, vis-à-vis de mon propre cheminement en recherche.

Les enseignements en TAL ont toujours constitué le pourcentage le plus important, même si l'évolution de la matière enseignée peut être constatée. Entre 1992 et 2004, mon cours de TAL incluait toujours une part de programmation (d'abord en FX/Lisp, puis en Prolog et en Perl, en collaboration avec ma collègue Augusta Mela). Par la suite, s'est graduellement mis en place un enseignement d'« outils pour linguistes », plus accessible aux étudiants en sciences du langage.

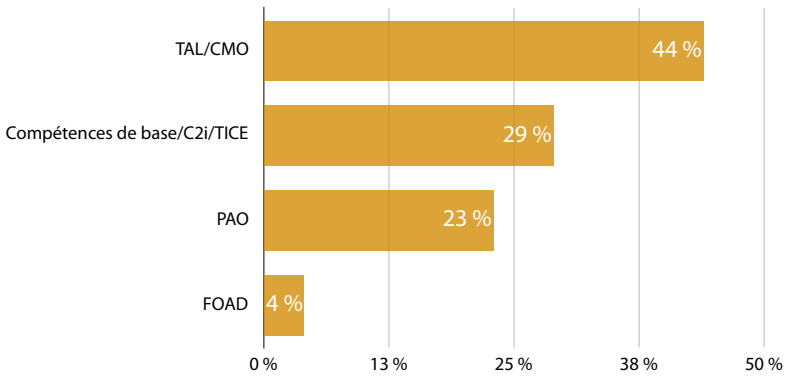


Figure 2.3 – Pourcentages de répartition des thématiques

L'objectif recherché était qu'ils deviennent des « utilisateurs avertis » de logiciels de TAL, puis de linguistique de corpus, afin d'être en mesure d'échanger avec des informaticiens. Mes enseignements TAL « pur » ont donc muté vers une dimension « usager critique » beaucoup plus affirmée, avec un apport analytique de la communication médiée par ordinateur (CMO) puis du discours électronique médié, discours numérique médié (DEM/DNM) dans les années qui ont suivi.

Les compétences « machinales de base » (bureautique/TICE, etc.) ont été directement assurées par ma collègue et moi-même jusqu'en 2004, puis un accord entre départements, et au niveau de la direction de l'université, a permis de centraliser l'ensemble des enseignements de « compétences informatiques » au seul département de MIAP. Néanmoins, notre intervention en tant que « prestataire de service » a toujours été encouragée (*cf.* tableau 2.1). À l'heure actuelle, tous les étudiants inscrits à l'université Paul-Valéry Montpellier 3 suivent un (ou plusieurs) enseignement(s) de C2i. Les groupes pour ces cours sont constitués de manière hétérogène, car les étudiants s'inscrivent pêle-mêle, quelle que soit leur discipline de rattachement. Dans ce cadre, la plupart des cours (d'une durée hebdomadaire de deux heures) se déroulent en première et en deuxième années. Après une brève présentation par l'enseignant-chercheur en début de séance, les étudiants, qui travaillent en binôme, effectuent leurs travaux dirigés en « autonomie guidée » [2] (Vincent-Durroux et Panckhurst 2002). L'enseignant-chercheur peut ainsi être au plus près des étudiants car il se déplace dans la salle et intervient ponctuellement directement auprès des étudiants en éventuelle difficulté. Ce public et ces cours m'ont toujours intéressée car cela constitue une richesse, voire un défi, pour l'enseignant-chercheur qui doit constamment s'adapter en fonction des besoins/des disciplines des étudiants. L'expérience de la formation pour les personnels de l'université (*cf.* § 2.3) m'a permis, peut-être, de mieux appréhender les inquiétudes de certains étudiants pour cette matière.

À partir du moment où je suis devenue auteur ou coordinatrice d'ouvrages, la chaîne de traitement éditoriale m'a fascinée. Sans doute à cause de la dimension applicative de mes recherches en linguistique-informatique, j'ai tout de suite voulu mieux comprendre le processus de l'intérieur. J'ai toujours participé à la mise en pages de mes livres, et cela a constitué une source d'inspiration pour

2. L'explication de ces concepts pédagogiques sera détaillée dans la troisième partie de l'habilitation. *Cf.* § 3.3.2.4.

élaborer des cours en publication assistée par ordinateur (PAO) à partir de 1999-2000, d'abord autour de *QuarkXpress*, puis à l'aide d'*Adobe InDesign*. Ce cours d'initiation — en pré-professionnalisation — que je partage actuellement avec un professionnel de l'édition, Gilles Pérez, est toujours très apprécié des étudiants, car très empreint d'une dimension pratique. Il cible actuellement une autre de mes découvertes (non-académiques) : l'auto-publication (Panckhurst, 2014).

Depuis 2004, j'assure un cours en Master *Gefnum* (*Gestion des connaissances, apprentissages et formation ouverte et à distance*, devenu ensuite, *Gestion des connaissances, formations et médiations numérique,* puis *Huma-Num, Humanités Numériques*), pour des étudiants inscrits eux-mêmes entièrement à distance. Pendant plusieurs années, je me suis associée avec une professionnelle du secteur eLearning, Debra Marsh, et nous avons partagé un enseignement qui constitue une réflexion sur la formation ouverte et à distance, avec mise en situation professionnelle immédiate des étudiants.

> Tous ces enseignements seront mis en relation avec mon parcours en recherche dans la troisième section de ce manuscrit, *cf.* § 3.

2.2 Maquettes ministérielles

Revenons un temps au début de ma carrière. Je crois que cela est important pour comprendre les mutations qui ont suivi. Dans un contexte national, au milieu des années 1990, le *traitement automatique des langues* prenait de l'importance, dans deux sections : sciences du langage (CNU 7) et informatique (CNU 27). Dans un article pédagogique, Panckhurst (1996a, « Formation en linguistique-informatique : une expérience Montpelliéraine », revue *Traitement automatique des langues* (T.A.L.), 37, 1, p. 51-64), j'ai présenté la mise en place d'un cursus d'informatique pour linguistes dans mon université. Cela incluait, en première année, un enseignement des « connaissances machinales de base » : initiation à la bureautique et à quelques concepts fondamentaux en informatique théorique (aperçu historique, architecture et structuration interne d'un ordinateur, hiérarchie et organisation, données et codage, mémoire, etc.). En deuxième année, j'avais opté pour un enseignement théorique général (qu'est-ce qu'un découpage

de problème, un algorithme, un organigramme, etc.) et une formation pratique à un logiciel favorisant l'apprentissage de la programmation orientée objet [3] :

> L'objectif était de passer de « l'informatique-bureautique » (la manipulation), à « l'informatique-programmation légère » (familiarisation avec le domaine, création et élaboration de petits programmes. (PANCKHURST 1996a, p. 56).

Dans les années supérieures, j'ai proposé une initiation au TAL. En troisième année :

Introduction

- rappel de notions de linguistique générale, et application au traitement automatique ; grammaires formelles et automates correspondants ;
- industries de la langue, génie linguistique, etc. : domaines, produits, marché actuel… ;
- linguistique appliquée au TAL : niveaux ou modules de traitement (prétraitement, morphologie, syntaxe, sémantique, pragmatique).

Approfondissement

- démonstrations et utilisation de systèmes existants ;
- premières réflexions sur l'évaluation de logiciels.

Puis, en cinquième année, la programmation pour linguistes, en FX/Lisp.

J'ai demandé une habilitation ministérielle pour la licence Sciences du Langage mention TAL pour la rentrée 1995-1996. Dans un esprit d'ouverture, après avoir élaboré la maquette, j'ai proposé aux collègues d'autres départements de rattacher la mention à leurs licences, notamment en langues. Plusieurs départements ont souhaité nous accompagner dans cette démarche. Conclusion ministérielle : un *avis défavorable* pour les Sciences du Langage (motif : « nombre faible d'inscrits » dans la maquette), mais un *avis favorable* pour les licences d'anglais et de portugais ! Le département de Sciences du Langage a décidé de ne pas accepter de fonctionner uniquement en « prestataire de service » pour les spécialistes d'autres disciplines. Peut-être était-ce une erreur. Il a fallu donc attendre le plan

3. À l'époque, (lointaine !), il s'agissait de *Hypercard/Hypertalk*.

quadriennal suivant, à la rentrée 1999, pour l'ouverture officielle d'une mention TAL, que nous avons ensuite obtenue.

Pour le secteur sciences humaines et sociales, le positionnement du TAL a toujours posé question : quel type de linguiste fallait-il former ? Un vrai linguiste-informaticien, taliste ? Un « utilisateur averti » sachant manipuler les outils-logiciels existant pour le TAL ? Ces questions étaient déjà présentes dans mon article pédagogique de 1996, paru dans la revue *Traitement automatique des langues* (T.A.L.) :

> À l'heure actuelle, notre université en est à ses premiers balbutiements en linguistique-informatique. Ainsi, maintes questions demeurent sans réponse adéquate. [...] Comment proposer l'insertion professionnelle d'étudiants (trop peu) formés en linguistique-informatique ? Faut-il renoncer à une formation « stricte » en TAL, en sachant qu'un employeur préférera embaucher un ingénieur informatique apprenti linguiste qu'un linguiste apprenti informaticien ? Ne serait-il pas plus prudent de former des linguistes-documentalistes-chercheurs de l'information qui sachent manipuler un ordinateur et qui comprennent les enjeux de la linguistique-informatique mais sans avoir réellement besoin de programmer ?
>
> (Panckhurst 1996a, p. 62).

Parallèlement, (Segond et Zaenen 1995) exprimaient leurs soucis à recruter dans le milieu industriel. Voici un extrait du résumé de l'article, intitulé « Cherche linguiste-informaticien désespérément », paru dans T.A.L :

> Les auteurs expriment leurs difficultés à trouver les spécialistes dont ils ont besoin au Centre de recherche Rank Xerox à Grenoble. Les étudiants français en effet sont à l'heure actuelle généralement assez mal préparés à la recherche en traitement automatique du langage naturel (TALN) en milieu industriel. Les auteurs montrent que, compte tenu de la diversité des domaines que rencontre le TALN et des faibles moyens des universités, il est souhaitable de mettre en place des cursus spécialisés, formant des experts en linguistique capables de collaborer avec des experts en informatique, et vice-versa. Ils comparent ainsi le système français aux cursus américains qui privilégient à juste titre le travail collectif en équipe. (Segond et Zaenen 1995).

En 1996, la tension montait au niveau national. Les recrutements en linguistique-

informatique en section 7 du CNU étaient rares, voire non-existants. Dans PANCKHURST 1996b (« Linguistique-informatique : la crise », revue *T.A.L.*, 37, 2, 176-177), j'ai « pété les plombs » :

> Que l'on m'excuse le ton amer ; lorsque je constate que nous continuons à former des étudiants pour des « métiers » de linguistes-informaticiens, que par ailleurs, dans les maquettes ministérielles on insiste sur l'importance de l'apprentissage informatique/bureautique (que j'ai intitulé « machinal de base ») en premier cycle et de la formation aux « nouvelles technologies », je m'insurge ; je ne peux qu'être très en colère de constater qu'aucun poste de linguistique-informatique n'est publié en France pour la rentrée 1997-1998 (un seul poste est proposé en « intelligence artificielle », mais celui-ci ne dépend pas de la 7e section).
> Que faire face à cette crise ? Abandonner tous les enseignements en TAL en France ? Proposer un certain « vernis culturel » aux étudiants sans jamais dépasser le stade superficiel d'un enseignement introductif ? Muter en 71e section et former des « info-graphistes » ? Je n'ai pas de solutions effectives à proposer mais l'avenir des étudiants dans ce secteur m'inquiète fortement.
>
> (PANCKHURST 1996b, p. 177).

Ai-je été entendue ? L'année suivante, nous avons obtenu un poste de MCF en linguistique-informatique [4]. Grâce à ce recrutement, nous avons pu assumer la mise en place d'enseignements disciplinaires plus intenses, car ma nouvelle collègue avait un triple profil mathématiques/informatique/linguistique.

L'habilitation de la mention *traitement automatique des langues* assortie à la licence Sciences du Langage a été (enfin) obtenue en 1999. Cette demande ministérielle avait été faite sous ma responsabilité, dans le cadre du plan quadriennal. Notre spécificité résidait dans l'orientation suivante : « traitement automatique des langues et nouvelles technologies éducatives : conceptualisation et évaluation de ressources multilingues ».

Les enseignements spécifiques de la mention TAL (126 heures) s'organisaient de la manière suivante :

4. Augusta Mela a été recrutée sur ce poste qu'elle a occupé jusqu'à son décès en octobre 2014.

— *TAL et Programmation pour linguistes* (84 heures, consacrées à la formalisation et à l'exploitation des données linguistiques en analyse syntaxique automatique et à deux langages de programmation (Prolog et Perl)).
— *Initiation au TAL et aux nouvelles technologies éducatives* (42 heures. Utilisation et étude critique de logiciels. Recherche, recueil et exploitation de ressources linguistiques. Évaluation de logiciels et de sites Web conçus pour l'enseignement supérieur).

Nous avions également obtenu l'habilitation de la mention *Industries de la langue*, assortie à la maîtrise de Sciences du Langage pour cette même période, qui consistait en un approfondissement (56 heures) des enseignements de la mention TAL en licence, ainsi qu'un stage en entreprise.

Parallèlement, grâce à ces mentions assorties à nos diplômes, et suite à des discussions nourries avec le département MIAP, nous avons décidé de déléguer la mission de l'enseignement généraliste (bureautique, compétences informatiques, C2i) au département MIAP afin de nous concentrer sur nos spécialités.

Le contexte français continuait néanmoins à être complexe pour les mentions TAL et IDL. En 2001, Jean Véronis, alors président de l'ATALA (*Association pour le traitement automatique des langues*), a impulsé une « Réflexion sur l'Enseignement et la Pédagogie des Techniques d'Informatique Linguistique (REPTIL) », qui visait à répertorier les formations nationales et à envisager le positionnement des formations TAL. J'ai été invitée à une table ronde du colloque annuel TALN (2002) à Nancy, intitulée : « Réflexion sur la recherche et l'enseignement de l'informatique linguistique » afin d'exposer la situation de notre université.

Au niveau local, comme sur le plan national, le positionnement du TAL/IDL au sein des licences de Sciences du Langage était constamment remis en question. Plusieurs raisons expliquent cela :

1. En 2003-2004, nos mentions officielles TAL et IDL n'ont pas été renouvelées, en tant que telles, dans le plan quadriennal suivant, pour cause d'« effectif trop réduit ». En effet, si le parcours TAL tentait certains étudiants, son côté formel, d'une part, et la nécessité de se former assez longtemps avant de devenir réellement performant — notamment en programmation — d'autre part, pouvaient provoquer une certaine frilosité pour certains étudiants ;

2. Nous avons néanmoins continué à assumer des enseignements d'initiation théorique et pratique au TAL, mais le ratio enseignement/ECTS était très faible. Nous avions 117h de TAL pour 6 ECTS seulement (alors que 3 ECTS correspondaient, à l'époque, à 19h30), d'où un désintérêt compréhensible pour certains étudiants ;
3. Le droit de cumul entre les parcours FLE et TAL a été supprimé. Devant ce choix, de nombreux étudiants ont préféré le parcours FLE, qui leur paraissait plus immédiatement professionnellement exploitable ;
4. Peu de linguistique appliquée existait dans le Master SL recherche à l'époque et les étudiants souhaitant une orientation informatique pour leurs travaux de recherche en doctorat avaient moins d'options d'encadrement.

À partir de ces années-là, les questions posées une décennie auparavant (Panckhurst 1996a, 1996b) ont été définitivement tranchées — tout au moins à l'université Paul-Valéry Montpellier 3. En sciences du langage, il fallait désormais former des étudiants « utilisateurs avertis » de logiciels, capables de dialoguer avec des informaticiens, mais les enseignements plus poussés en programmation allaient être réservés aux informaticiens formés en informatique-linguistique. Les linguistiques de corpus, entre autres, commençaient également à prendre de l'importance dans les formations en sciences du langage *Cf.* (Cori et al. 2008a).

Dans d'autres universités, les réflexions continuaient également. Antoniadis 2008 évoque sa position pour l'université Stendhal (qui est devenue maintenant Grenoble-Alpes) :

> Même si actuellement l'enseignement d'informatique pour les non-spécialistes commence à trouver ses marques, les mutations technologiques et leur impact sur la société doit nous amener à continuer à se poser la question de son curriculum. À notre avis, quelles que soient ces mutations, la compréhension des bases de l'informatique « science » permet de les cerner et d'y poser un regard avisé.
>
> (Antoniadis 2008, p. 45).

Pour le poste d'enseignant-chercheur mis au concours récemment (MCF, n° 0554, 2016) au sein de l'université Paul-Valéry Montpellier 3, nous avons souhaité profiler le poste en fonction de ces évolutions, avec une assise en linguistique-informatique :

> *Natural language processing/computational linguistics.*
> A linguist with a computational linguistics profile is required for this position. The current teaching commitments also cover computational lexicography and corpus linguistics. Research requirements include natural language processing (NLP) with a special interest in processing authentic corpora data in relation to discourse analysis. Programming skills in NLP would be a valuable asset but are not mandatory; in the latter case, the linguist must be capable of elaborating methodological proposals to be implemented by computer scientists.

Francesca Frontini — spécialiste en linguistique computationnelle, en provenance du CNRS italien, le CNR (*Consiglio Nazionale delle Ricerche*), plus précisément de *Istituto di Linguistica Computazionale « Antonio Zampolli »*, CNR Pisa, et ayant un large rayonnement international en recherche — a été recrutée en tant que maître de conférences sur ce poste à la rentrée 2016.

2.3 Formation

Retour en arrière. En 1996, après avoir été élue directrice adjointe de l'UFR, j'ai été nommée coordinatrice de la formation pour les enseignants-chercheurs dans le cadre de l'initiation aux TICE. J'ai assuré entre 40 et 70 heures en plus de mon service statutaire annuel pour répondre à ce besoin, jusqu'en 2003. De prime abord, ont été incluses, dans mes heures complémentaires, des séances de formation des formateurs aux enseignants-chercheurs, puis cela s'est élargi [5] aux personnels administratifs et techniques, et aux doctorants : familiarisation à l'utilisation de la plateforme institutionnelle pour la mise à disposition des cours (*cf.* également la section *Encadrement de séminaires de formation* de mon curriculum vitae).

Après trois ans de formation, j'ai entrepris de rédiger, en co-auteur, une introduction aux technologies de l'information et de la communication dans un contexte universitaire. Le livre se limite volontairement aux formations pour les

5. Trop souvent, à mon goût, un clivage demeure entre les catégories de personnels : les enseignants-chercheurs d'un côté, les personnels administratifs et techniques de l'autre. Pendant les 7 années que j'ai assuré ces formations, j'ai essayé de réduire cette division, en organisant des formations pour un public mixte.

personnels menées pendant l'année universitaire 1998-1999 : notions de base (Mac OS, Windows), accès à Internet (paramétrage de l'ordinateur), courrier électronique, formats de fichiers, navigation/moteurs de recherche. (*Cf.* § 3.3.2 pour une réflexion sur le rapport formation-recherche).

Notre souci, dans le cadre de ces formations, était de faire en sorte d'aider les personnels universitaires à se former aux outils et aux ressources technologiques en évolution constante (*cf.* § 3.3.2.1, § 3.3.2.2).

> Comme les formes particulières d'enseignement subissent une évolution importante à l'heure actuelle (individualisation de l'enseignement — enseignement sur mesure — diversité des formations non-présentielles : enseignement distributif, interactif, collaboratif...), il est naturel que les enseignants-chercheurs ressentent le besoin, l'urgence de se former aux outils via lesquels les savoirs et les savoirs-faire seront véhiculés à l'avenir. Bien entendu, cette envie d'apprendre peut être accompagnée d'une crainte importante (voire un refus) liée à la remise en question de soi, à la dépersonnalisation de la relation avec autrui, etc.
> Par ailleurs, cette évolution implique une révolution déjà en gestation depuis fort longtemps, et préconisée par les didacticiens de la première heure : centrer réellement l'enseignement sur l'apprenant. Celui-ci, en participant activement dans un groupe, peut apporter, à son tour, des informations nouvelles à l'enseignant, qui, en quelque sorte, abandonne sa casquette du tout-puissant-détenteur- du-savoir, pour devenir un guide, un tuteur, au sein d'un système d'apprentissage collectif, de travail en équipe.
> Cela peut paraître parfaitement utopique dans une université de près de 20 000 étudiants, au sein de laquelle perdurent des problèmes de locaux, d'encadrement, de manque de financement. Mais, pour ma part, je veux y croire ; je sais que nous avons à notre disposition, au sein de notre université, une masse énorme de compétences et de savoir-faire. Saurons-nous en tirer parti, mettre nos connaissances en commun, les exploiter, les offrir aux autres via les nouvelles ressources ? Et ainsi, permettre aux personnels et aux étudiants de débuter dans le troisième millénaire de manière confiante ? Je l'espère !
>
> (Panckhurst et Pérez 2000), p. 23-24.

2.4 Direction de service commun

Outre les formations, j'ai été nommée responsable, pendant la période (1996-1998) qui a précédé le nouveau plan quadriennal (démarrage 1999), d'une commission sur les nouvelles technologies, notamment pour veiller aux besoins des enseignants-chercheurs en matière de formation aux outils informatiques. En février 1999, la présidente de l'université, via son conseil d'administration, m'a nommée directrice du service commun SEAM (dont j'ai ensuite modifié le nom en multimédia, enseignement, technologies de l'information et de la communication éducatives (METICE)), afin de mettre en place un « campus virtuel » et de développer notre offre en matière de formation à distance via les nouvelles technologies. En tant que directrice, j'ai impulsé une politique de numérisation et de mise en ligne des cours qui existaient sous forme de papier depuis 25 ans. Nous avons choisi *WebCT*, un outil canadien, doté d'une interface en français — car cela nous paraissait un aspect primordial, et toutes les plateformes de l'époque ne le proposaient pas — pour constituer la plateforme institutionnelle de diffusion des cours. Nous l'avons utilisé une dizaine d'années, avant de changer pour *Moodle*.

Dans le cadre de mes enseignements propres, j'ai conçu et élaboré une dizaine de cours/TD spécifiques (cours en L1, L2, L3, M1 et M2) mis en ligne sur notre première plateforme, WebCT, puis une autre dizaine, ensuite, sur Moodle.

La responsabilité passionnante, mais très prenante, de direction du METICE, a pris fin en juin 2001, date à laquelle j'ai souhaité me réinvestir dans mes activités de recherche. J'ai néanmoins continué les actions de formation pour les personnels jusqu'en 2003.

Plusieurs publications font suite à cette période de direction de service et n'auraient sans doute pas vu le jour sans ce lien avec une responsabilité administrative (*cf.* § 3.3.2, Panckhurst 2001a, Vincent-Durroux et Panckhurst 2002, § 3.3.3, Panckhurst 2003a).

2.5 Missions pédagogiques et administratives

J'ai toujours veillé à ce que le lien se fasse entre l'enseignement et l'administration, d'où mon implication au niveau de missions pédagogiques et administratives aux niveaux local et national :

— Présentation du cursus « traitement automatique des langues » (TAL) de l'université Paul-Valéry Montpellier 3 à la journée METIL (Métiers du TAL et des Industries de la langue), Paris, 7/3/02.

— « Salon de la formation ouverte et à distance », Paris, Porte de Versailles, 28 février–1er mars 2001.

— Journée au Ministère « Enseigner, apprendre et communiquer les TIC », Paris, Ministère, 3 avril 2000.

— Responsable enseignant dans le cadre de la formation « Apogée », Rennes, mai 1997, 2 stages de 2 jours.

— Responsable du groupe *Nouvelles technologies de l'information et de la communication* (NTIC), préparation du contrat quadriennal (1998) et chargée de mission TICE (avant de devenir directrice du METICE en février 1999).

— Correspondant de l'université Paul-Valéry aux groupes de travail « enseignement à distance » (en collaboration avec un autre collègue) et « technologies de l'information et de la communication éducatives », sous la coordination du Conseil scientifique du pôle européen (février 1999-juin 2001).

2.6 Commissions, conseils, comités, jurys, diplômes [6]

J'ai assuré la fonction de membre élu du Conseil d'Administration de 2001 à 2002, suite au décès d'une collègue mathématicienne, Dany Serrato.

Au sein du département de Sciences du Langage, j'ai été membre (1994-2002) puis suppléante (2002-2004) puis à nouveau membre (2005-2008) de la commission des spécialistes pour le recrutement d'enseignants-chercheurs en CNU 7, ainsi que membre du jury de la licence de Sciences du Langage, mention *TAL*, puis membre du jury de la licence Médiation Culturelle et Communication (MCC) (1996-2008), avant d'assurer la présidence du jury de la licence Sciences du Langage parcours MCC. Depuis 2008, la commission des spécialistes a été remplacée par un comité d'experts locaux. J'ai participé à ces comités jusqu'en 2009.

En 2015-2016, j'ai été élue présidente du comité de sélection pour le recrutement d'un maître de conférences en linguistique-informatique (MCF, n° 0554).

Ayant toujours été impliquée dans les activités pédagogiques et administratives dans d'autres disciplines et dans d'autres établissements également, j'ai assumé la fonction de membre suppléant de la commission des spécialistes de la 27e section du CNU (informatique, université Paul-Valéry Montpellier 3) de 1998 à 2002, et de membre extérieur de la commission des spécialistes de la 7e section du CNU, (université Aix-Marseille, 2004-2006, université d'Avignon, 2009-2010). Par ailleurs, comprenant l'importance de faire le lien entre le secondaire et le supérieur, j'ai accepté la charge de présidence du jury de baccalauréat à deux reprises (Lycée Henri IV, Béziers, juillet 1996 ; Lycée Clémenceau, Montpellier, septembre 1996).

J'ai été responsable de la licence de Sciences du Langage, mention *TAL* (1999-2002) et directrice d'études de la licence Sciences du Langage, *parcours MCC*

6. D'autres responsabilités administratives et pédagogiques en plus de celles mentionnées, dont les détails sont indiqués dans le Curriculum Vitæ (Volume III), n'ont pas de rapport direct avec mes activités de recherche et ne sont pas développées ici (parmi celles-ci : chargée de mission pour l'international (mise à disposition pour un tiers de mon temps de service d'enseignement auprès du PRES (devenu COMUE), membre de la commission des équipements ; membre suppléant puis membre du comité d'hygiène et sécurité ; responsable de planning pour notre département ; membre de commissions de validations d'acquis professionnels, commissions de choix/ad hoc, etc.)

(2008-2010). Par ailleurs, notre département, sous la responsabilité d'un comité de pilotage que je coordonne, a mis en place (entre 2009 et 2011) un projet d'enseignement de la LSF) ainsi que des cours portant sur la surdité et les implications linguistiques l'enseignement adapté et les médiations sémantiques [7]. Ce projet a permis de relier l'université, le Rectorat-DAFPEN et différentes structures liées à la surdité en région (CROP : *Centre de Rééducation de l'Ouïe et de la Parole*, Arieda : *Association Régionale pour l'Intégration et l'Éducation des Déficients Auditifs*, Visuel-LSF : *Visuel-Langue des signes française*, CESDA : *Centre d'éducation spécialisée pour déficients auditifs*). Je suis actuellement responsable du parcours « surdité et LSF » de la licence Sciences du Langage, depuis sa mise en place en 2011.

2.7 Évaluation (cursus universitaire)

En 1993, Claude Ricciardi-Rigault, alors professeur à la Télé-université du Québec, m'a confié l'évaluation officielle de la mise en place d'un cursus universitaire de baccalauréat (équivalent québécois de la licence), spécialisé en linguistique-informatique, au sein de l'université du Québec à Laval. Mon rapport de 19 pages a exigé une évaluation approfondie des aspects à la fois théoriques et pratiques du cursus, avec une proposition de réorganisation de la progression des cours à mettre en place ainsi qu'une modification, voire suppression, de certains contenus. Cela a constitué une opportunité très intéressante qui m'a permis de porter un regard comparatif entre les programmes mis en place dans les universités françaises et les universités québécoises.

L'évaluation de mes propres cours à l'université par les étudiants a toujours fait partie de mes priorités. À l'issue de chaque cours, depuis 25 ans, les étudiants remplissent un questionnaire anonyme et répondent à une dizaine de questions portant sur le contenu, l'organisation, la structuration du cours et du contrôle contenu. J'estime que cette étape est primordiale : premièrement, les étudiants se sentent responsabilisés dans leur appréciation du cours ; deuxièmement, les commentaires permettent parfois de réorganiser ou d'approfondir certains points pour le cours de l'année suivante.

7. Pour une introduction au domaine, j'invite le lecteur à consulter un mémoire professionnel portant sur l'accessibilite scolaire et pedagogique des jeunes déficients auditifs (Lamrani 2006) ainsi qu'un ouvrage sur la langue orale des jeunes sourds profonds (Vincent-Durroux 2014).

2.8 Conclusion

Pourquoi longuement évoquer les aspects pédagogiques et administratifs dans le cadre d'une habilitation ? Précisément, car mon enseignement et mes responsabilités administratives ont constamment nourri mon cheminement en recherche, et *vice versa*. La question de cette imbrication sera approfondie dans la troisième section de l'habilitation, à travers la présentation des volets de recherche (*cf.* 3).

Cette section, § 2, concernant les aspects d'enseignement, de formation, et les responsabilités diverses, peut être naturellement liée au volet 2 (*cf.* § 3.3.2) de mes recherches : *Formation, (auto)évaluation, réseaux pédagogiques*, avec la publication, entre autres, de trois ouvrages en co-auteur ou en co-coordination (Panckhurst et Pérez 2000), (Vincent-Durroux et Panckhurst 2002), (Panckhurst et al. 2004a).

Les publications liées aux trois volets (*cf.* § 3.3.1, § 3.3.2, § 3.3.3), seront explicitées et approfondies dans la troisième section de ce manuscrit (*cf.* § 3).

> Le parcours initial jusqu'au doctorat, puis les activités d'enseignement et d'administration étant contextualisés, il est désormais temps de se tourner vers la recherche.

Te mutunga – ranei te take

Maori : « La fin – ou le début » (Hulme 1983).

3

Recherche

> If we knew what it was we were doing, it would not be called research, would it ?
>
> *Albert Einstein*

Mes premiers travaux de recherche ont commencé en 1984, aidés par deux assistants de recherche — mes parents ! À l'époque, ils faisaient un tour du monde. Après 5 mois de voyage à travers un grand nombre de pays qui séparent la Nouvelle-Zélande de la France, il se sont heureusement arrêtés un temps à Clermont-Ferrand. Tous deux universitaires, ils m'ont aidée pour la mise en place du travail méthodologique et le dépouillement de vidéos pour mon premier mémoire. C'était une première initiation (*cf*. Volume III, Curriculum Vitæ).

Les trois décennies qui ont suivi ont été imprégnées de nombreux questionnements, croisements, incertitudes, micro-avancements. Si je devais situer des périodes, j'indiquerais sans doute les deux suivantes, de 13 ans pour chacune : 1) 1991-2004 ; 2) 2004-2017. Au début de la première période, j'ai bénéficié d'une bourse post-doctorale d'une année à l'université du Québec à Montréal (UQÀM). Entre les deux périodes, j'ai effectué un congé de recherche de 6 mois à *l'University of Technology* à Sydney (*cf.* Volume III, Curriculum Vitæ), ce qui m'a donné un espace-temps pour prendre du recul et réfléchir à l'orientation que je souhaitais donner pour la suite de mes recherches, et comment j'allais équilibrer l'enseigne-

ment et la recherche. La troisième période est celle qui débute maintenant et qui me permettra, je l'espère, de rebondir dans d'autres directions (*cf.* § 4).

Mais je préfère organiser ma réflexion autour de trois volets, qui aideront à comprendre plus précisément comment mes recherches ont parfois été menées en parallèle avec quelques enlacements. Cela me paraît plus approprié que d'évoquer une périodicité séquentielle. Comme indiqué précédemment (*cf.* § 1.2.1), trois volets essentiels ont traversé mon cheminement :

1. *Prototypes et outils* (1991-2003) : interrogatives, verbes, gloses ;
2. *Formation, (auto)évaluation, réseaux pédagogiques* (Technologies de l'information et de la communication éducatives (TICE), eLearning/ formation ouverte et à distance (FOAD)) (1996-2012) ;
3. *Communication médiée par ordinateur* (CMO), *discours électronique médié* (DEM), *discours numérique médié* (DNM) : analyse de courriels, forums, chats, SMS (1996-2017).

Certes, ces volets se situent sur des plans très différents.

En intitulant le **premier volet** « Prototypes et outils », je ne veux aucunement insinuer que l'outil est une fin en soi. Après mon doctorat, ma recherche fondamentale a porté — un temps — sur la structuration du groupe verbal, sur la portée de la sous-catégorisation verbale, sur les unités verbales polylexicales, et, plus tard, sur la glose. Simplement, à partir d'une recherche initialement fondamentale, j'ai élaboré des **prototypes** ou des **outils informatisés**, utilisables par autrui. De ce fait, **la recherche fondamentale devenait réellement appliquée**, (*cf.* § 3.3.1).

Le **deuxième volet** se concentre sur mes expériences et réflexions qui mêlent enseignement, administration et recherche. Il s'agit de se tourner davantage vers les autres, d'analyser leurs besoins, de prendre en considération leurs attentes, et d'apporter une réflexion pouvant à son tour leur permettre d'avancer dans leur propre(s) cheminement(s). Ici, il ne s'agit pas de créer des outils informatisés, mais d'**aider les autres à utiliser des outils informatisés** à bon escient. Cela ne s'arrête pas à une visée purement applicative. Il s'agit d'**apporter une dimension réflexive aux apprenants/formateurs sur leurs propres usages**, ce qui peut

aider à envisager la façon dont ils structurent leurs propres visées pédagogique et de recherche. Encore une fois, l'outil ne doit pas être une fin en soi. Ici, la **recherche appliquée est dotée d'un enlacement avec sa dimension réflexive**, (*cf.* § 3.3.2).

Enfin, le **troisième volet** est le plus long, en matière de périodicité, traversant, pour l'heure, un peu plus de deux décennies. C'est le début du recueil et de l'exploitation de **données authentiques**, d'abord exclusivement dans un contexte d'enseignement supérieur et de recherche (courriels, forums, chats), puis, plus récemment, en élargissant au grand public, par une collecte importante de SMS. À partir de 1996, j'avais commencé à constater des variations de comportement lorsque l'ordinateur (puis, plus tard, le téléphone) est utilisé dans un cadre de communication interpersonnelle. Est né, en 1997, mon néologisme pour le français, la **communication médiée par ordinateur**, suite à (Herring 1996), et non la communication médiatisée par ordinateur. J'ai modifié ce terme en **discours électronique médié**, puis **discours numérique médié**, afin de prendre en considération, notamment, le téléphone portable. Il s'agit ici d'effectuer des observations et des **analyses de discours**, à l'aide (ou non) de logiciels appropriés (des analyseurs morpho-syntaxiques ou des logiciels de type concordanciers, etc. typiquement utilisés en linguistique de corpus), et en élaborant certains prototypes pour des implémentations informatiques éventuelles. D'une part, la **recherche** est **descriptive, analytique** et, d'autre part elle est **appliquée dans une dimension utilisatrice**. L'**implémentation informatique** est désormais envisagée **en collaboration avec des informaticiens**, (*cf.* § 3.3.3) .

À l'heure actuelle, je continue à effectuer des recherches dans le cadre du troisième volet, en essayant de réunir l'ensemble de mes compétences acquises au cours de ces dernières décennies. Mes recherches actuelles tendent peut-être vers de la recherche empirique. J'essaierai d'expliquer pourquoi.

Avant d'aborder de manière détaillée la « synthèse de mes travaux scientifiques », par le biais des trois volets, j'évoque brièvement ci-dessous les aspects administratifs de la recherche ainsi que l'encadrement des étudiants et des stagiaires.

3.1 Administration de la recherche

3.1.1 Activités post-doctorales

J'avais obtenu une bourse d'études et effectué un premier **stage doctoral** de linguistique-informatique dans le cadre du projet de coopération franco-québécois « Conception et application d'un analyseur lexico-syntaxique du français » (ALSF), au Centre d'Analyse de Textes (ATO) à l'université du Québec à Montréal (UQÀM), d'une durée de 9 mois, de septembre 1988 à mai 1989. Quelques semaines après avoir soutenu ma thèse en décembre 1990, j'ai commencé un nouveau sujet de recherche — « Validation du modèle retenu pour la représentation du groupe verbal et description de la sous-catégorisation verbale » — dans le même cadre de coopération franco-québécoise, et ce pendant 6 mois, de janvier à juin 1991. À partir de septembre 1991, pendant un an, j'ai bénéficié d'une **bourse d'excellence post-doctorale** (délivrée par l'*Association des universités partiellement ou entièrement de langue française* (AUPELF-UREF), devenu en 1998, *Agence universitaire de la francophonie* (AUF)), Domaine : lexicologie, terminologie, traduction (TAO). Projet : « Repérage et extraction automatique des unités polylexicales verbales ». J'ai pu regagner Montréal et travailler à nouveau au Centre d'ATO, à l'UQÀM. En parallèle, j'ai déclaré ma société au Québec (le nom de la raison sociale était PLICO (*Panckhurst LInguistique COmputationnelle*). Cela m'a permis de réaliser deux **contrats de services** avec l'*Office de la langue française* (OLF), pendant mon année québécoise : 1) *Réalisation de travaux portant sur la description du groupe verbal, sur la modélisation des hypothèses linguistiques de résolution des différents problèmes qui y sont relatifs et sur la sous-catégorisation verbale ;* 2) *La production d'un rapport sur un dispositif automatisé de classification lexicale pour la sous-catégorisation verbale (Scatlex), la finalisation d'un manuel d'utilisation et de l'implémentation informatique* .

3.1.2 Laboratoire, conseils, comités, jurys

Mon rattachement au **laboratoire** de recherche est resté stable ; je suis membre de Praxiling (UMR 5267 CNRS université Paul-Valéry Montpellier 3) depuis 1992. Étant donné ma spécialité disciplinaire, j'ai été sollicitée pendant plusieurs années (1992-1998) pour être **responsable** du parc informatique du laboratoire. J'ai

ensuite créé le **site web** du laboratoire (http://www.praxiling.fr/), et, à partir de 2004, j'ai réalisé un nouveau site web sous SPIP à l'aide du kit SPIP CNRS (http://www.harmoweb.cnrs.fr), en collaboration avec le service informatique de l'université Paul-Valéry Montpellier 3. Ma fonction de responsable du site et **Webmaster** a pris fin en 2012 lorsque le directeur du laboratoire a décidé de modifier les responsabilités des membres. L'hébergement du site est désormais assuré au CNRS.

En 1994, j'ai été sollicitée pour participer au **jury** européen pour l'évaluation de logiciels universitaires, European Academic Software Award (EASA). J'ai été membre du jury à trois reprises pour la finale du concours biannuel : Heidelberg (Allemagne), 1994 ; Klagenfurt (Autriche), 1996 ; Oxford (Angleterre), 1998. En 1997, j'ai rejoint le **comité d'organisation** du concours, *European Knowledge Media Association*, jusqu'à sa dissolution en 2006. Huit réunions ont eu lieu après cette date :

— Montpellier (France), 20 novembre 2004.
— Le Lôcle et Neuchâtel (Suisse), Finale, 25-27 septembre, 2004.
— Ronneby (Suède), Finale EASA, 22-25 novembre, 2002.
— Édimbourg (Écosse), 8-9 septembre, 2001.
— Rotterdam (Pays-Bas), Finale EASA, 25-28 novembre, 2000.
— Amsterdam (Pays-Bas), 24 -25 avril, 1999.
— Vienne (Autriche), 31 janvier-1er février, 1998.
— Oxford (Angleterre), 12-13 avril 1997.

Convaincue de l'importance des colloques de jeunes chercheurs, j'ai accepté de participer au **comité scientifique** du colloque international, NEDEP (*Numérique(s), enjeux défis et perspectives)*, organisé par les jeunes chercheurs de Praxiling à Montpellier en 2009. Par ailleurs, j'ai été sollicitée pour être membre du **comité scientifique** du colloque international IMPEC (*Interactions Multimodales Par ÉCran*), qui s'est tenu à Lyon, 2014.

En 2013, j'ai été invitée à rejoindre le **Conseil scientifique** de *l'Agora des Savoirs*, de la Ville de Montpellier (http://www.montpellier.fr/3806-savoirs.htm). Le

but de l'Agora est de diffuser des connaissances culturelles, scientifiques et techniques par le biais de conférences hebdomadaires, gratuites et ouvertes à tous. Cette responsabilité a pris fin en 2016.

3.1.3 Responsabilités éditoriales

Mon implication éditoriale a été constante. J'ai participé au **comité de rédaction** de la revue *Cahiers de Praxématique* de 1992 à 2006. J'ai **coordonné** le numéro 22 des *Cahiers* en 1994, (Panckhurst 1994b), (http://praxematique.revues.org/1887). Lorsque j'ai dirigé le service METICE (*cf.* § 2.4) j'ai impulsé la mise en place d'une **collection**, intitulée *MédiaTic*, publiée par le service des publications. Cette collection proposait des ouvrages consacrés à l'application pédagogique des nouvelles technologies dans le cadre de l'enseignement présentiel, aménagé et à distance. Trois ouvrages sont parus, (Panckhurst et Pérez 2000), (Vincent-Durroux et Panckhurst 2002), (Panckhurst et al. 2004a) avant que le service ne soit restructuré en les *Presses Universitaires de La Méditerranée* (PULM) en 2007 et que la totalité des collections ne soit réorganisée.

3.1.4 Séminaires, évaluations, tables rondes, journées d'études et colloques

Pendant mon année post-doctorale en 1991-1992, j'ai organisé des **séminaires** de recherche internes et intra-équipes entre l'équipe *Recherche et développement en linguistique computationnelle* (RDLC), et le département de linguistique à l'université du Québec à Montréal (UQÀM).

Pendant la période 1994-1998, dans le cadre du concours EASA (*cf.* § 3.3.2), j'ai assumé la fonction de **coordinateur de discipline**, qui consistait à trouver des collègues universitaires européens experts qui étaient susceptibles d'effectuer des **évaluations** de logiciels.

En 1995, j'ai coordonné une **table ronde** à l'Espace République du Conseil Général de Montpellier, intitulé « Qui pousse les hypermédia, la technique ou la société ? », dont les participants étaient : Jean Sallantin (LIRMM, CNRS), Alain Cazes (CNAM, Paris), Corinne Chuat (ITEM, CNRS), Jean-Christophe Mielnik (Sextant Avionique), Pierre Lévy (Paris, par vidéo-conférence).

FIGURE 3.1 – Table ronde : Hypermédia (1995)

J'ai assuré la coordination locale pour Praxiling du **colloque** *Association for French Language Studies* (AFLS : « Description du français : discours, corpus, analyses, applications »), qui s'est tenu au sein de notre université en septembre 1997.

Entre 2000 et 2011, j'ai organisé les journées d'études nationales, européennes et internationales suivantes :

Organisation d'une journée d'étude internationale (deux jours) :
Harmonisation/standardisation des méthodes de traitement de corpus écrits de type SMS. Anonymisation, transcodage, annotation, 15 participants, Maison des Sciences de l'Homme de Montpellier (MSH-M), 14-15 novembre 2011.
Organisation d'une journée d'étude européenne :
Formation ouverte et à distance : perspectives européennes — European e-learning perspectives, 15 participants, université Paul-Valéry Montpellier 3, 14 octobre 2005.
Organisation d'une journée d'étude européenne :
Évaluation en formation ouverte et à distance : bilan et perspectives — Evaluation in e-learning : review & future directions, 15 participants, université Paul-Valéry Montpellier 3, 19 novembre 2004.
Organisation d'une journée nationale sur l'évaluation de logiciels, université Paul-Valéry Montpellier 3, 15 participants, 21 juin 2002.
Organisation d'une journée internationale sur l'autoformation et l'autoévaluation : METICE, université Paul-Valéry Montpellier 3, en collaboration avec le pôle scientifique du Languedoc-Roussillon, mai 2000.

Entre 2011 et 2013, j'ai organisé 14 séminaires internationaux en présence et par visio-conférence dans le cadre du projet *Sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée* (2011-2012), puis du projet D.G.L.F.L.F. (Délégation générale à la langue française et aux langues de France) : *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS* (2012-2013) à la Maison des Sciences de l'Homme de Montpellier (*cf.* § 3.4 Réseaux, diffusion et valorisation pour le détail). Les 10 séminaires filmés sont en ligne : http://msh-m.tv/spip.php?rubrique138.

3.1.5 Évaluation/expertise

Mon expérience d'évaluation et d'expertise a commencé lorsque j'ai évalué la mise en place d'un cursus pédagogique de linguistique-informatique (*cf*. Volume III, Curriculum Vitæ).

Côté recherche, j'ai été sollicitée à plusieurs reprises pour être évaluateur d'articles soumis à des colloques, à des revues dans des spécialités variées, et également d'un projet de recherche international et d'un manuscrit d'ouvrage collectif en sciences du langage.

Colloques *:*
Évaluateur de 2 articles soumis à IMPEC 2014, http://impec.ens-lyon.fr
Évaluateur de 4 articles soumis à ascilite, Sydney, 2006, http://ascilite.org
Revues *:*
Évaluateur d'un article soumis à *Lingvisticæ Investigationes*, https://benjamins.com/catalog/li, mars 2017.
Évaluateur d'un article soumis à la revue *Applied Psycholinguistics*, http://journals.cambridge.org/APS, juillet 2013.
Évaluateur d'un article soumis à la revue *ALSIC https://alsic.revues.org*, avril 2013.
Évaluateur d'un article soumis à la *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, http://ees.elsevier.com/erap/, décembre 2010.
Évaluateur de 3 articles soumis à la revue de sociolinguistique en ligne, *Glottopol*, http://glottopol.univ-rouen.fr/, mars 2007.
Projet de recherche international *:*
Rapport d'expertise confidentielle d'un projet de recherche universitaire anglais-

australien, pour la Nuffield Foundation, www.nuffieldfoundation.org, Londres, Royaume-Uni, juin 2010.
Ouvrage collectif *:*
Rapport d'expertise confidentielle d'un manuscrit d'un ouvrage collectif pour ENS Éditions, Lyon, décembre 2009, http://www.ens-lyon.fr/editions/catalogue.

3.1.6 Projets de recherche et CRCT

En début de carrière universitaire (1992-1994), j'ai participé, au sein de notre laboratoire, au projet international MED-Campus, 170, intitulé « Nouvelles techniques de traitement de l'information », Réseau Transméditerranéan, Dialogos « Communication et situations d'interculturalité ». À ce titre, j'ai été responsable du secteur « Traitement automatique du langage naturel et hypertexte » dans le cadre de missions de coordination, pédagogiques et de recherche à l'étranger : Florence (19-22 juin, 1993), Le Caire (23-30 septembre, 1993), Florence, (1-7 mai 1994).

Ma période de présidence de l'association « Montpellier-Cognition » (1995-1996) a été l'occasion d'organiser des projets de recherche informels et des lieux de débats multidisciplinaires entre linguistes, spécialistes des sciences cognitives, informaticiens (Lirmm, Montpellier) (*cf*. entre autres, § 3.1.4).

En 1997-1999, j'ai participé au projet « Lexique : catégorisations et représentations » avec le soutien du Fonds International de Coopération Universitaire (AUPELF-UREF), sous la direction de Paul Siblot. J'ai mené une activité internationale en ayant été conférencière invitée à l'université du Québec à Chicoutimi (en septembre, 1997). J'ai également participé à deux autres projets de recherche de notre laboratoire : « De l'actualisation » et « L'autre en discours ». Pendant cette période, j'ai été responsable des deux projets suivants, dans le cadre des activités de notre laboratoire : *Dynamique et fonctionnements discursifs ; Formes discursives et textualité* : « Le discours médié », et *Sujet, praxis et production de sens ; Lexique et discours* : « Le verbe et son environnement ».

Entre 2000 et 2010, j'ai assuré les responsabilités de deux projets de recherche au sein de Praxiling : « La communication médiée par ordinateur » (2000-2006) et « Discours électronique médié » (2007-2010).

J'ai fait ma première demande de congé de recherche (congé pour recherches ou conversions thématiques (CRCT)) en 2003. Celui-ci m'a été immédiatement accordé et je l'ai effectué sous la direction du professeur Shirley Alexander, à *l'Institute for Interactive Media and Learning* (IML), à l'University of Technology (UTS), Sydney, Australie, pendant le deuxième semestre de l'année universitaire 2003-2004, de janvier à juin. Depuis cette date, je n'ai pas fait d'autres demandes de congés de recherche, à cause de mes autres implications et projets de recherche. Je n'ai pas non plus sollicité de congé de recherche pour la rédaction de cette habilitation — même si cela fut parfois difficile de trouver des moments prolongés pour écrire — car je préfère en demander un prochainement, pour un autre projet d'écriture (*cf.* § 4).

Entre 2010 et 2013, j'ai été responsable de deux projets de recherche autour des SMS : *Sud4science Languedoc Roussillon. Mutation des pratiques scripturales en communication électronique médiée*, Programme Maison des Sciences de l'Homme de Montpellier (MSH-M), (2011-2012), et *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS*, Délégation générale à la langue française et aux langues de France (D.G.L.F.L.F.), (2012-2013) (*cf.* § 3.3.3 pour les détails de ces projets).

En 2013-2014, j'ai participé au projet PEPS-HuMaIn (Humanités - Mathématiques - sciences de l'Information), ECOMESS (Analyse contrastive des émotions contenues dans les messages courts), porté par Mathieu Roche (Tétis, Cirad).

Depuis 2015, je continue à faire de la recherche avec mes collègues ayant participé aux projets SMS, mais sans financement extérieur accordé, pour l'heure.

3.2 Encadrement global des travaux de recherche

L'encadrement de la recherche me paraît crucial ; je participe au comité de suivi des thèses, mis en place en 2014-2015. Jusqu'à présent, j'ai assuré la (co-)direction de 34 mémoires de M1 & de M2 ainsi que l'encadrement de 2 stagiaires affectés au METICE lorsque je dirigeais ce service commun. Pendant le déroulement du projet SMS *Sud4science Languedoc Rousillon*, j'ai (co-)encadré 8 stagiaires étudiants en M1/M2. J'ai participé à maints jurys de soutenance de M1 & M2, voire de doctorat, également, dans deux départements de l'université Paul-Valéry Montpellier 3, à l'INALCO et à la Sorbonne.

Tableau 3.1 – Encadrement de la recherche

Participation à jury de doctorat	
André Frédéric, « Pratiques scripturales et écriture SMS : analyse linguistique d'un corpus de langue française », Doctorat de Sciences du langage, université Paris-Sorbonne, Jury : C. Fairon, E. Stark, R. Panckhurst, S. Plane, G. Siouffi (directeur). Soutenance le 24 avril 2017.	2016-2017
Comités de suivi de thèses	
Hang Gao : « Discours de sport entre France et Chine : le cas des jeux paralympiques » ; Camille Lagarde-Belleville : « De la médiatisation du rugby à travers le rugbyman » ; Michel Otell : « De quelques processus de production du sens dans les SMS conversationnels des sourds signants ».	2015-2016
Camille Lagarde-Belleville : idem ; Michel Otell : idem. Jérémy Perroux : « Archéologie discursive des dynamiques de l'altérité à l'œuvre dans les stéréotypisations de la paysannerie africaine. Du discours institutionnel colonial à celui du développement ».	2014-2015
26 Directions de mémoires, rapports de stage	
M2, Sciences du langage, Llorach Carole, « Comment comprendre l'acquisition et l'évolution du langage chez l'individu utilisant les nouveaux moyens de communication (sms, internet) ».	2016-2017
M1, Sciences du langage, Discours médiatiques, institutionnels et politiques : « Le métier de journaliste reporter d'images dans une chaîne de télévision du service public » (S. Bort), Rapport de stage.	2015-2016
1) M1, Sciences du langage, Discours médiatiques, institutionnels et politiques : « L'inscription de la complicité dans les SMS », (C. Luong) ; 2) M1, Sciences du langage, Spécialité Gestion des connaissances, formations et médiations numériques : « La conception des recettes de cuisine en ligne : une approche ergonomique et linguistique », (C. El Khouri).	2014-2015
1) M2, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « Écriture SMS et Néographie chez les jeunes de 11-30 ans », (G. Porto). 2) M1, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « Caractéristiques propres au scripteur et au Smartphone impactant l'écriture SMS », (J. Laboureau) ; M1, Gestion des connaissances, formations et médiations numériques : 3) « Camfranglais : Pratiques et usages plurilingues dans les SMS au Cameroun » (A. Moussa) ; 4) « La reconnaissance vocale sur mobile : lécritur d sms, 2 lécri a Loral » (A.Kaba).	2013-2014

M1 Gestion des connaissances, formations et médiations numériques : « Plurilinguisme et SMS » (N. Dos Santos).	2012-2013
1) M2 : « Jeux sérieux dans un contexte pedagogique. Memoire de recherche sur les potentialites educatives des jeux video. », (S. Reverte). 2) M1 : « L'expression des émotions et des sentiments dans Twitter et les SMS : analyse comparée des usages, des formes et des objectifs » (E. Orlando).	2011-2012
M1 : « Téléphones intelligents et tablettes : usages pédagogiques des logiciels d'apprentissage mobile (Mobile learning software). » (L. Gobert).	2010-2011
M1 : 1) « L'utilisation des TICE dans la rééducation orthophonique : les troubles du langage écrit » (E. Pennes) ; 2) « Tice et dyspraxie/dysgraphie : Les TICE permettent-elles une meilleure inclusion des élèves dyspraxiques/dysgraphiques en milieu scolaire ordinaire ? » (D. Bartholemé).	2009-2010
M1 : 1) « TICE et surdité en milieu scolaire : de la théorie à la réalité » (S. Pollet) ; 2) « Conversations asynchrones en communication médiée : comparaison entre courriels, forums et SMS. » (D. Fontaine)	2008-2009
M2 : « Réseaux sociaux et enjeux pédagogiques » (B. Naoul) ; M1 : 1) « SMS et déficiences visuelles » (A. Gaussuron) ; 2) « Jeux en ligne et stratégies pédagogiques » (V. Lucas).	2007-2008
M1 : « Le langage SMS des jeunes. Approche lexicale et morpho-syntaxique » (M. Fayada).	2006-2007
M1 : 1) « Nouveaux usages de communication électronique en russe. » (N. Svarinska) ; 2) « Étude lingustico-communicationnelle des SMS en France et en Bulgarie. » (K. Filipov).	2005-2006
Mémoire/rapport de stage de M2, Gestion, apprentissages et formation ouverte et à distance « Conception et réalisation de la mise en ligne d'un cours dans le cadre d'une formation au sein du département de français de l'université Pédagogique de Maputo » (N. Bernard).	2004-2005
Maîtrise en Sciences du Langage, mention Industries de la langue : 1) « La glose et une application informatique du phénomène. Création d'une rubrique consacrée à la glose à l'aide d'un système de publication par Internet » (C. Porta) ; 2) « Enquête, logiciel, tutoriel : conception, évaluation, modification » (S. Lafaye).	2001-2002
Maîtrise : « Étude linguistique sur les correcteurs grammaticaux » (C. Tresallet).	1998-1999
8 Co-directions de mémoires	
M2, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « L'interjection dans les SMS : usages et tendances scripturales » (J. Laboureau, co-directeur : L. Fauré).	2014-2015
M2, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « Écriture SMS et Phénomènes Phonétiques. Évolution des pratiques scripturales entre 2004 et 2011 » (F. André, co-directeur : F. Hirsch).	2013-2014
M1 : « L'identité dans une communauté en ligne, l'exemple du forum rock6070.com » (D. Blind, co-directrice : C. Béal).	2009-2010
M1 : « Conversations et SMS » (H. Catapano, co-directeur : L. Fauré.	2008-2009
M1 : « Discours électronique médié et SMS : étude phonologique de quelques morphèmes verbaux » (O. Caumont, co-directeur : L. Fauré).	2006-2007
DEA : « L'accès à l'information sur Internet » (F. Pascual, co-directeur : P. Siblot).	1998-1999

DEA : « Introduction à une étude des difficultés du français et de leur traitement par ordinateur » (M. da Conceição, co-directeur : A. Coïaniz).	1996-1997
DEA : « La préposition "à" au sein des constructions locatives en français » (S. Lescure, co-directeur : J. Bres).	1994-1995
Participations à jurys de soutenance	
M1 Recherche, « Interdiscours et créativité dans les slogans de manifestations » (Y. Ghliss, Département de Sciences du langage, directeur : B. Verine). M1, « La salade russe était française…Français/Russe, de la compétence linguistique à la compétence communicative » (L. Montagne, Département de Sciences du langage, directrice : C. Béal)	2012-2013
M1 : Participation à jurys, journées pour l'ensemble de la promotion « Gestion des connaissances et formation ouverte et à distance », MSH, Montpellier.	2006-2010
M2 Gestion, apprentissages et formation ouverte et à distance : 1) « Stage de production d'un site Intranet », Sufco, université Paul-Valéry Montpellier 3, (A. Diallo) ; 2) « Mise en place du site Internet Coaching-lombalgie.com » (C. Sesboué).	2005-2006
M2 : Participation à jurys, journées pour l'ensemble de la promotion « Gestion des connaissances et formation ouverte et à distance », MSH, Montpellier.	2004-2005
M1 Recherche : « Outils pour l'analyse automatique du discours » (S. Riou, Département de Sciences du Langage, directeur : P. Siblot).	2003-2004
Maîtrise en Sciences du langage, mention Industries de la langue : « La traduction automatique : quelques pistes de réflexion à partir d'une première expérience pratique sur le système de T.A. du LIRMM » (V. Vedel, direction : A. Mela).	2001-2002
Maîtrise en Sciences du langage, mention Industries de la langue : « Repérage automatique des sources d'énonciation » (F. Ruas, direction : A. Mela).	2000-2001
DEA : « La phrase clivée c'est...qui/que et ses équivalents en polonais » (A. Nowakowska, directeur : J. Bres). Maîtrise : « Phonétique Expérimentale : Traitement et analyse numérique de la parole. Essai d'une description acoustique des voyelles nasales du français assisté du logiciel Signalyze 3.12. » (C. Pujol, direction : M. Lanvin).	1997-1998
Maîtrise : « La rupture publicitaire » (N. Llinares, direction : C. Charnet).	1995-1996
D.E.S.S : « Terminologie du traitement automatique du langage naturel » (Mourad Amine, Ingéniérie multilingue, CRIM, INALCO, direction : M. Slodzian). DEA : « Quand la langue devient un produit industriel : le projet Translearn » (K. Wagner, Département de Grec Moderne, université Paul-Valéry Montpellier 3, direction : M.-P. Masson).	1994-1995
8 stagiaires ayant travaillé dans le cadre du projet de recherche SMS *sud4science/DGLFLF*	
Co-encadrement de cinq stagiaires de M1/M2, programme *sud4science* et *DGLFLF* : Camille Lagarde-Belleville, Michel Otell, Frédéric André, Yosra Ghliss, Reda Bestandji.	2012-2013
Co-encadrement de deux stagiaires de M2 (LIRMM, Namrata Patel et Pierre Accorsi), juin-juillet, 2012. Encadrement d'un stagiaire de M1 (A. Stifani), septembre-décembre 2011.	2011-2012
2 stagiaires affectés au METICE, sous ma direction	
Encadrement de deux stagiaires SUFCO (Formation Continue) en DU (diplôme d'université) de concepteur médiatique : 1. Marilyne Martin : « Médiatisation de cours en ligne », septembre et octobre 2001 ; 2. Sabine Cotreaux : « Présences unies vers cités distantes », avril et mai 2001.	2001

3.3 Synthèse de mes travaux scientifiques

Dans les trois paragraphes qui suivent, je détaille les moments marquants de chacun des trois volets mentionnés supra (*cf.* § 1.2.1) :

1. *Prototypes et outils* (1991-2003) : interrogatives, verbes, gloses ;
2. *Formation, (auto)évaluation, réseaux pédagogiques* (Technologies de l'information et de la communication éducatives (TICE), eLearning/ formation ouverte et à distance (FOAD)) (1996-2012) ;
3. *Communication médiée par ordinateur* (CMO), *discours électronique médié* (DEM), *discours numérique médié* (DNM) : analyse de courriels, forums, chats, SMS (1996-2017).

Je n'évoquerai pas l'ensemble de mes travaux, mais seulement une sélection de ceux-ci. Je mettrai davantage l'accent sur le troisième volet que sur les deux premiers, car c'est celui qui m'a vraiment permis de m'épanouir en recherche ces dernières années. J'essaierai de définir aussi ce qu'est la recherche pour moi, comment je l'envisage, avec ses moments de bonheur, ses imperfections, ses tâtonnements aussi. Surtout, je voudrais tout simplement apporter un regard sur ma façon de travailler — qui pourrait paraître à première vue éclectique, car j'essaie de privilégier une ouverture d'horizons relativement divers — depuis de nombreuses années. Mais plutôt que de retenir l'idée d'éclectisme, je préférerais penser que j'ai une certaine dose de liberté, d'ouverture, de souplesse, tout en tentant de me maintenir hors d'un cadre disciplinaire strict. Sans doute est-ce quelque peu utopique…

3.3.1 Volet 1 : Prototypes et outils (1991-2003)

3.3.1.1 Formation clermontoise

Grâce à Gabriel G. Bès, et les travaux scientifiques et industriels menés dans le cadre du projet européen ESPRIT-ACORD, à Clermont-Ferrand, (ESPRIT Project 393, ACORD, *Construction and Interrogation of Knowledge Bases Using Natural Language Text and Graphics*), j'avais baigné dans un réseau important européen de chercheurs utilisant et proposant différents modèles et formalismes (*cf.* infra.). À l'origine, mon directeur de thèse est lui-même arrivé en France (de son Argentine natale) pour faire son doctorat (soutenu en 1972) sous la direction du fonctionnaliste, André Martinet. Il était également vivement intéressé par la grammaire générative de Chomsky et avait mis au programme les grammaires formelles dans ses enseignements en Argentine, dès 1969 :

> Bès, intéressé par la formalisation et le calcul, se disait admiratif du « premier Chomsky » et a dû être séduit par les propositions concrètes de ses premiers textes (Trouilleux 2015, p. 22).

Il s'en est éloigné par la suite — à l'époque où Chomsky prônait une relation forte entre structures syntaxiques et structures cérébrales. Bès s'est alors tourné vers d'autres façons de formaliser : la grammaire de Montague [1], les modèles *Lexical-Functional Grammar* (LFG, (Kaplan et Bresnan 1982)) et *Generalized Phrase Structure Grammar* (GPSG, (Gazdar et al. 1985)), et les implémentations possibles en traitement automatique des langues. Tout au long de sa vie de linguiste il a prôné le triplet 1) observations, 2) hypothèses formalisées, 3) test des hypothèses par rapport aux observations. De ce fait, il m'a transmis l'importance du lien nécessaire entre recherche théorique et recherche appliquée, que j'avais déjà commencé à privilégier dans mes premiers travaux de recherche en maîtrise. En continuité avec ces choix théoriques, en thèse, **la formalisation et**

1. À l'issue de mon année de DEA, Michel Chambreuil, qui nous avait enseigné la grammaire de Montague, m'a demandé de traduire en anglais le livre qu'il avait écrit sur la sémantique formelle du logicien, afin que celui-ci puisse paraître en édition bilingue. J'ai signé un contrat avec la maison d'édition et j'ai complété la traduction — ce qui m'a financé un voyage en Nouvelle-Zélande — mais en définitive seule l'édition française est parue (Chambreuil 1990), à cause d'une question de surcoût et de complexité éditoriale.

l'abstraction, en dehors d'un formalisme prédéfini, constituaient mes maîtres mots :

> [Les] propriétés [pertinentes concernant l'interrogation] doivent être suffisamment abstraites pour que n'importe quel formalisme puisse les incorporer au besoin. (Panckhurst 1992b).

Après la recherche doctorale, dont l'objectif était de bâtir un *répertoire descriptif généralisé* des interrogatives directes du français, et dont la démarche était pluridisciplinaire, se situant à la fois en linguistique, en informatique et en documentation — avec une double dimension théorique et appliquée — j'ai effectué une recherche de 6 mois, dans le même cadre franco-québécois que précédemment. Dans un premier temps, il s'agissait de recenser les courants, les modèles, les grammaires, les différentes publications sur le groupe verbal et la sous-catégorisation verbale de l'époque. J'ai étudié, entre autres, les courants suivants : GPSG (Grammaire syntagmatique généralisée, (Gazdar et al. 1985), (Bès et Baschung 1985), pour le français), FUG (Grammaire d'unification fonctionnelle, (Dik 1978 ; Kay 1979)), HPSG (Grammaire syntagmatique généralisée guidée par les têtes, (Pollard et Sag 1994)), LFG (Grammaire lexicale fonctionnelle, (Kaplan et Bresnan 1982)), LG (Lexiques-grammaires, (Gross 1975)), TAG (Grammaire d'arbres adjoints, (Joshi et al. 1975), (Abeillé 1991, 1993)), UCG (Grammaire catégorielle d'unification, (Calder et al. 1986), (Baschung 1991) pour le français), GraCE (Grammaire Catégorielle Étendue (Segond 1990) [2]. Après le structuralisme, notamment le distributionnalisme et le transformationalisme [3] de mes études précédentes, je m'étais initiée aux grammaires formelles

2. Un colloque s'est tenu à Clermont-Ferrand au sein de notre laboratoire de recherche, le GRIL, en 1990, à propos des grammaires catégorielles (Desclés et Segond 1990 ; Lecomte 1990). Un peu plus tard, en 1993, Cori et Marandin ont publié leurs travaux sur « *Grammaire d'arbres polychromes (GAP)* »,(Cori et Marandin 1993).

3. Dans la partie linguistique de ma thèse, que j'avais conçue de manière indépendante de tout formalisme grammatical particulier, je proposais de dériver les structures interrogatives à partir de structures déclaratives, et ce au moyen de *transformations*. Je définissais celles-ci comme étant des outils descriptifs qui permettaient de relier les couples entrées-sorties de symbolisations de constituants. Cependant, cela divergeait des définitions habituelles de transformation des écoles de Pennsylvanie et de Cambridge : « Contrairement à Harris et à Chomsky nous n'utilisons pas la transformation en tant qu'élément d'une théorie explicative. [...] Notre but essentiel [...] est l'élaboration d'une description se servant de la transformation en tant qu'outil visant à faire des

et à d'autres modèles syntaxiques [4].

3.3.1.2 Analyseur lexico-syntaxique du français, ALSF

Dans le cadre de ma recherche québécoise, il s'agissait de proposer une description linguistique cohérente pouvant être implémentée ultérieurement dans l'analyseur ALSF (*Analyseur lexico-syntaxique du français)*, développé en collaboration entre le Centre d'ATO, université du Québec à Montréal et des chercheurs du CNRS à Paris, notamment Jean-Marie Marandin, que j'avais déjà côtoyé et beaucoup apprécié pendant mon stage doctoral québécois.

Mais avant de poursuivre sur le déroulement de mon année post-doctorale — afin d'en comprendre son assise théorique — remontons un petit peu dans le temps. L'héritage de Marandin provenait, entre autres, du programme de recherche de l'analyse de discours (AD) de Michel Pêcheux [5], et son *Analyse automatique du discours* de 1969 (AAD69) :

> AAD69 n'a pas été conçu comme modèle linguistique, ni plus généralement comme modèle des sciences du langage [...]. Il a été conçu comme modèle du sens, machine de guerre politique, et aussi comme dispositif pratique à visée documentaire ; ce qui explique qu'il s'est trouvé tiraillé entre la rigueur et la cohérence exigées d'un modèle théorique (théorie du discours) et les nécessaires ajustements et compromissions qui ont toujours été, et sont encore actuellement, le lot des dispositifs de traitement automatique des langues. (Léon 2010).

Maldider 1990 se souvient de la démarche tout à fait pionnière de Pêcheux pour l'époque : « C'est peut-être son rapport à l'informatique qui est sa plus grande originalité. Il ne voulait pas s'en servir, il voulait la faire servir. Contrairement aux premières démarches de l'intelligence artificielle, l'informatique devait selon

spécifications. » (Panckhurst 1990, p. 15).

4. Je m'intéressais également à d'autres approches et théories linguistiques, mais d'un strict point de vue d'une certaine culture générale. Entre autres : la théorie *Sens-Texte* de Igor Mel'čuk, les « espaces mentaux » de Fauconnier, les « univers de croyance » de Martin, la grammaire cognitive de Langacker, etc.

5. « Je dédie ce travail à la mémoire de Michel Pêcheux. Les motivations et les orientations du programme de recherche que je présente ici viennent par mille détours des inlassables discussions que nous avons eues à propos de l'analyse de discours. » (Marandin 1997).

lui permettre de reformuler les hypothèses, d'aller plus loin dans une lecture "où le sujet est à la fois dépossédé et responsable du sens qu'il lit" »
(Marandin et Pêcheux 1984) in (Maldider 1990, p. 90).

L'implémentation informatique de l'AAD69, dès 1971, allait au-delà d'une réflexion uniquement épistémologique, car « elle s'accompagnait de la construction d'un instrument et de la production de résultats expérimentaux » (Panckhurst (1997), Compte rendu (Hak et Helsloot 1995), *Michel Pêcheux. Automatic Discourse Analysis*). Cela étant, les critiques (émanant parfois de l'auteur lui-même) et les remaniements n'ont pas tardé à faire surface :

> Michel Pêcheux lui-même et son groupe de travail ont émis un certain nombre de critiques méthodologiques comme le caractère hybride des énoncés élémentaires, l'arbitraire des pondérations, ou la difficulté de comparer des structures complexes à l'aide des procédures algorithmiques. Pêcheux, 1975, in (Léon 2010).

Au début des années 1980, l'abandon d'AAD69 a été officialisé (Pêcheux et al. 1982) et la syntaxe a été clairement replacée au centre de l'analyse de discours (Marandin 1997 ; Marandin et Pêcheux 1984), (Léon 2010) :

> La notion de perception syntaxique fut, à l'origine, une manière de concevoir et d'expliquer le privilège que l'AD donne à la syntaxe, privilège qui la singularisait parmi les théories du discours qui ont émergé au tournant des années 1970. [Cela] conduisait nécessairement à un important travail sur le prédicat syntaxique et donc à un travail en syntaxe : il ne pouvait être rigoureux sans le recours à une formalisation digne de ce nom. En retour, cette formalisation ne pouvait être enracinée que par un surcroît d'attention à une empirie clairement définie.
> (Marandin 1997, p. 3).
> L'hypothèse centrale est qu'aucune manipulation d'expressions linguistiques n'est possible sans prendre en compte leur structuration syntaxique, considérant toutefois que le questionnement sur l'autonomie de la syntaxe dans les phénomènes discursifs implique que d'autres dimensions soient prises en compte comme l'énonciation, le lexique ou la séquence. (Léon 2010).

3.3.1.3 Déredec et FX

Le renouveau — vers une autre AAD — est ensuite arrivé du Québec, via le langage de programmation, Déredec, écrit en LISP, par Pierre Plante, du Centre d'ATO à l'UQÀM :

> [Déredec] pouvait être utilisé par des linguistes non-informaticiens tout en permettant un travail collectif et partagé. Écrit en LISP, un des deux langages créés au MIT au début des années 1960 pour l'intelligence artificielle et les analyseurs syntaxiques, il offrait enfin une interface adaptée au traitement des langues naturelles, aux représentations arborescentes, à l'analyse syntaxique. On va enfin pouvoir construire la grammaire de reconnaissance souhaitée par Michel Pêcheux dès 1967. Un nouveau projet d'AAD, AAD80, devient alors possible.
>
> (Léon 2010).

Déredec, grâce à son approche ascendante à partir d'items lexicaux, privilégiait une approche constructiviste. Il a été remplacé quelques années plus tard par le langage FX. [6]

Après avoir été initiée à Déredec par Jacqueline Léon à Paris (1986-1987), j'ai eu la chance de bénéficier de l'apport intellectuel des cercles de chercheurs clermontois, parisiens et québécois, eux-mêmes au centre des mutations des approches novatrices en linguistique-informatique. Dans les années 1980-1990, l'élaboration d'outils informatisés pouvant être rattachés à un analyseur/parseur morpho-syntaxique avait le vent en poupe (tout au moins dans certaines écoles). Comme ils permettaient une analyse automatique (ou semi-automatique) de données en langage naturel, ils étaient d'une aide précieuse aux terminologues, par exemple, et très recherchés par les acteurs des industries de la langue, pour la construction d'interfaces homme-machine permettant de construire et d'interroger des bases de connaissances, entre autres applications. Il faut se replacer dans le contexte,

6. FX (Plante, 1988-1996) incluait des fonctionnalités pour les stratégies d'analyse (descendantes, montantes, mixtes), l'itérativité, la récursivité, la programmation positionnelle, réflexive, etc. Il permettait à la fois une certaine sensibilité contextuelle et une autonomie déclarative et conceptuelle de ses primitifs, appelés des « faisceaux ». Différentes types d'interfaces étaient également directement programmables. Par ailleurs, FX, écrit en *Common Lisp/Le_Lisp* faisait partie d'un ensemble d'outils, Atelier-FX, qui se voulait être un système d'analyse linguistique, d'extraction d'informations dans les textes et de mise au point de progiciels à base de connaissances (Plante 1996).

bien sûr ; même si les méthodes de fouille de données étaient déjà utilisées dans des contextes essentiellement littéraire et sociologique, la constitution de corpus numérisés ou nativement numériques n'était pas encore devenue monnaie courante. Puis, il fallait attendre encore quelques années pour l'apport de la tradition de linguistiques sur corpus (Bilger 2000 ; Péry-Woodley 1995), ou linguistiques de corpus (Habert et al. 1997), à partir de *corpus linguistics*, en anglais, *cf.* aussi (Cori et al. 2008a). Je reviendrai sur le débat linguistique-informatique/linguistique de corpus (*cf.* § 3.3.1.6).

Dans le contexte d'élaboration d'un *analyseur lexico-syntaxique du français* (ALSF), Marandin s'intéressait aux travaux de (Milner 1989), qui avait critiqué — tout comme Bès — d'une part, l'évolution de la grammaire générative, et, d'autre part, proposait une approche *géométrique* (ou *positionnelle*, appellation préférée par Marandin). En effet, Milner prônait une distinction entre *termes* (le lexique) et *positions* (la syntaxe) :

> La théorie linguistique doit définir ce qu'est un terme, et elle doit définir ce qu'est une position. Elle contiendra donc nécessairement deux parties au moins : la théorie des termes — qu'on appelle le lexique — et la théorie des positions, qu'on appelle la syntaxe. (Milner 1989, p. 307).

3.3.1.4 Outil informatisé, sémantique lexicale, classification verbale

En route pour Montréal. Comme explicité précédemment, dans le cadre de ma thèse, j'avais effectué mes recherches **hors formalisme grammatical spécifique**. En ayant été sélectionnée pour une bourse post-doctorale, j'acceptais, de ce fait, d'adopter l'approche théorique choisie par Jean-Marie Marandin et les autres chercheurs du projet franco-québécois (Claude Ricciardi Rigault, Pierre Plante, Sophie David pour ne citer que mes collaborateurs les plus proches). Cela me paraissait être une ouverture stimulante. J'ai foncé. En septembre 1991, je suis repartie à Montréal pour 12 mois, en tant que boursière post-doctorale (Bourse d'Excellence AUPELF-UREF, domaine : lexicologie, terminologie, traduction (TAO)) sous la direction de Claude Ricciardi Rigault, du laboratoire d'accueil *Recherche et développement en linguistique computationnelle* (RDLC), du Centre d'ATO, à l'UQÀM.

L'objectif du projet post-doctoral, intitulé « Repérage et extraction automatique des unités polylexicales verbales (TERMINOV) », à la suite d'une analyse lexico-syntaxique, était de repérer et d'extraire de façon automatique les unités verbales polylexicales [7] d'un texte :

> Une unité verbale polylexicale est une unité linguistique composée d'un verbe et d'un ou plusieurs autres éléments le suivant, le tout format une *unité de sens*. *Enfoncer le clou, mettre sur pied, passer l'arme à gauche, perdre les pédales, rendre justice, trouver porte close,* etc. sont des exemples d'unités verbales polylexicales.
> (Panckhurst 1998b, p. 161).

Mais en amont, il fallait terminer la validation du modèle retenu pour la représentation du groupe verbal (GV) et la description de la sous-catégorisation verbale, puis effectuer son implémentation, avant d'envisager la recherche fondamentale concernant les unités verbales polylexicales (UVPL).

Ma période post-doctorale de 12 mois s'est donc déroulée en trois temps :

1. Approfondissement de la recherche fondamentale sur le groupe verbal et la sous-catégorisation verbale et classification manuelle de 500 verbes ;
2. Implémentation informatique (dispositif d'assignation lexicale (DAL) [8] et Scatlex : *Dispositif d'assignation lexicale pour la sous-catégorisation verbale*).
3. Recensement théorique des problèmes posés par les UVPL, en préparation d'une deuxième année post-doctorale consacrée à l'implémentation de TERMINOV.

7. L'appellation était la mienne. La terminologie courante de l'époque était diverse : *locutions/lexies/périphrases verbales, expressions/locutions figées/idiomatiques,* etc. J'écartais l'appellation *expression/locution figée*, car le *figement* peut laisser croire qu'aucune altération n'est possible. Or, il existe maintes situations où les insertions/intercalations sont envisageables. Néanmoins, bien que Gross utilise l'appellation *expression figée*, il était bien sûr conscient que même dans les cas les plus contraints, ces expressions ne sont qu'exceptionnellement entièrement figées.

8. *DAL fonctionnait comme une boîte à outils incorporant (i) un aspect de reconnaissance et attribution de propriétés ; (ii) une possibilité de fouille, de regroupement et d'altération de données. Après la phase de description*, lorsque la base de données lexicale est réalisée, à partir d'une classification des verbes du français, à l'aide de Scatlex (via une classification effectuée au clavier, ou à partir d'un fichier de verbes en entrée, ou encore à partir de classifications verbales déjà existantes), le dispositif DAL permet d'effectuer également des *requêtes/fouilles*. DAL permet de regrouper les verbes déjà classés en une famille de verbes, selon un parcours de chemin(s) spécifique(s) indiqué(s) par l'utilisateur-classificateur, et de les visualiser/sauvegarder, etc. Dans le cadre de cette habilitation, je n'évoquerai que Scatlex.

Dans (Panckhurst 1993a), (Panckhurst 1994a), j'explique l'emprunt du cadre théorique *positionnel* (Milner 1989) et du formalisme *Grammaire d'arbres polychromes* (CAP, (Cori et Marandin 1993), pour le traitement du groupe verbal, puis de la sous-catégorisation verbale, tout en le situant, et en le distinguant, par rapport aux autres approches de l'époque, (LFG, LG, GPSG, HPSG, FUG, TAG, UCG, etc.) :

> Alors que dans une approche positionnelle le verbe ne construit pas de positions, dans d'autres théories, les différentes positions sont directement insérées dans l'entrée lexicale verbale. [...] [Notre] conception devrait permettre d'effectuer une distinction entre l'information intervenant à différents niveaux : sémantique, syntaxique, lexical. (Panckhurst 1993b, p. 63).

Cela étant, l'existence du GV en soi, était loin de faire consensus. Comme le faisait remarquer (Milner 1989) :

> Certains auteurs ont douté de l'existence du Groupe verbal comme entité syntaxique (par exemple M. Gross). Leurs arguments ne sont pas dépourvus de sens. De façon générale, il est vrai qu'aucun événement syntaxique majeur ne met en cause le seul Groupe verbal, à l'exclusion du reste de la phrase. Cela étant admis, *la notion de Groupe verbal paraît utile pour déterminer la fonction des compléments*. On pourrait supposer que l'existence du Groupe verbal ne se manifeste pour la syntaxe que par là. (Milner 1989, p. 375).

Par ailleurs, pour Gross, il n'existe pas d'entrées lexicales correspondant à des verbes précis, mais plutôt des entrées correspondant à des structures de phrases. Pour lui, un verbe entre dans une *construction*. Ses tables ne sont pas des *entrées lexicales verbales*, mais plutôt des « tables de constructions » correspondant à différentes structures (Gross 1975, p. 55). Il s'agit de trouver toutes les combinaisons distributionnelles ; les descriptions sont faites « au moyen de séquences » (Gross 1975, p. 41). En empruntant cette démarche, la notion de sous-catégorisation verbale était écartée pour favoriser celle de « description de verbes ».

Mais pourquoi n'avoir pas tout simplement tenté d'implémenter les tables de (Gross 1975), dès 1991, au lieu de proposer un autre système de classification verbale ? Pour plusieurs raisons. La première raison, simple (voire simpliste),

était que les tables du LADL n'étaient pas librement accessibles (en dehors de l'ouvrage *Méthodes en Syntaxe*, désormais *MS*) comme elles le sont aujourd'hui. [9] Mais le problème était plus complexe que cela. Pour certains chercheurs, (Bès, notamment), les tables de *MS* étaient « formellement ininterprétables » (Bès 1994, p. 8), in (Panckhurst 1994b), ou bien elles étaient difficiles d'accès en raison d'un « formalisme de représentation opaque » (Alcouffe et Falcoz 1994, p. 9), in (Panckhurst 1994b). Ces derniers ont présenté leur système qui cherchait à unifier la représentation des tables par l'usage d'un cadre formel structurant, le modèle GENELEX. Cela étant, dans mon article (Panckhurst 1994e), j'évoquais le nécessaire rapprochement entre mon système d'analyse verbale, Scatlex, et « l'informatisation des tables du LADL » (Panckhurst 1994e, p. 131). Il est indéniable que les tables sont extrêmement riches et fines dans leur description, et, d'autre part, beaucoup de chercheurs travaillent sur l'implémentation des tables depuis plusieurs décennies maintenant (*cf.* quelques repères *infra*, p. 69).

Dans un article écrit 15 ans plus tard, intitulé, « Le lexique-grammaire est-il exploitable pour le traitement des langues ? », (Laporte 2010) revient sur la question de l'implémentation des tables :

> L'idée que le lexique-grammaire est difficilement exploitable pour le traitement des langues découle en partie de la présence d'erreurs et de lacunes, qui peuvent être corrigées, mais aussi d'un sentiment d'étrangeté que ressentent les spécialistes d'analyse syntaxique devant des choix qui sont peu courants dans la plupart des autres projets dont ils ont connaissance. Une prise en compte des différentes données du problème amène à nuancer cette vue, et à justifier la plupart de ces choix par les caractères originaux du lexique-grammaire : un vaste recensement du lexique et des constructions, la priorité donnée aux données factuelles sur les contraintes liées à des théories spécifiques, une exigence de reproductibilité des observations. Or ce sont justement ces caractères qui ouvrent des perspectives d'exploitation du lexique-grammaire dans des systèmes de traitement des langues.
>
> (Laporte 2010).

Laporte préconise une neutralité vis-à-vis des théories syntaxiques, d'où une volonté de formalisation moindre :

9. http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html

> La neutralité par rapport aux théories syntaxiques explique par ailleurs le choix d'un degré de formalisation limité. Un formalisme plus complexe, nécessairement plus dépendant d'une théorie, n'aurait -il pas gêné l'observation éventuelle de faits auxquels cette théorie n'aurait pas été adaptée ? (Laporte 2010).

Sa question est pertinente pour moi et je défendrais plus volontiers ce positionnement aujourd'hui. Dans mon propre cheminement de recherche, après avoir adopté telle ou telle approche formelle — surtout dans mes premières années de recherche post-doctorale — j'ai souvent choisi, en définitive, de ne pas m'« enfermer » dans celle-ci au moment de l'implémentation [10]).

Mais revenons pour l'instant à mon parcours historique.

En se re-situant en 1991 (trois ans avant les articles que j'ai évoqués ci-dessus), il ne fallait pas rêver : l'entreprise d'informatisation était irréalisable par une seule personne, et pour ma part, j'étais séduite à l'époque — *cela ne serait pas nécessairement le cas aujourd'hui* — par l'approche positionnelle/géométrique, dans laquelle lexique et syntaxe étaient distingués.

Au début des années 1990, je souhaitais proposer un système informatisé basé sur des *contrastes*, par opposition à une série de *combinatoires*. Mon idée était de fournir une taxinomie, sous forme arborescente, implémentable, contenant virtuellement toutes les entrées lexicales verbales du français. En proposant une série de tests basés sur des contrastes, un utilisateur-expert aurait été à même de procéder à une classification verbale interactive, aussi exhaustive que possible.

Groupe verbal

Dans le cadre de l'analyseur ALSF, le groupe verbal était structuré comme dans la figure 3.2, à l'aide de deux positions [11] :

À cette époque, en 1991-1992, je définissais la sous-catégorisation verbale (SCV) comme *les compléments nécessaires à l'établissement de la signification lexicale d'un*

10. *Cf.* également la remarque très juste de Jacqueline Léon à propos des nécessaires ajustement pour la phase de l'implémentation (page 72).

11. Bien entendu, une position intercalaire (pour l'occupation d'adverbes, par exemple, avec un V ou une UVPL : *aime follement Marie* ; *perd complètement les pédales*) était prévue, ainsi qu'une extension de la tête verbale pour permettre l'insertion de clitiques (*le lui donne*). La structuration interne du noyau permettait également de prendre en compte les auxiliaires (*avoir* et *être*) et de distinguer entre formes simples et composées (*mange, mange-t-il ?, a mangé, a été mangé, a été subitement mangé*, etc.). Je n'ai pas besoin de rentrer davantage dans ce détail ici.

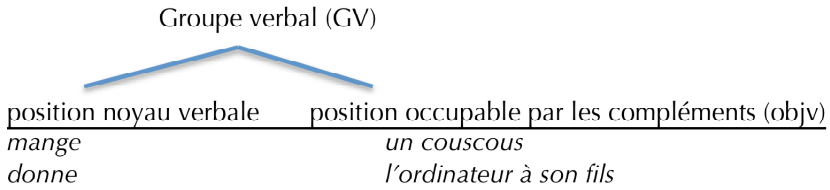


FIGURE 3.2 – Structuration du groupe verbal

verbe. Dans (PANCKHURST 1994e, p. 110-111)., je rappelle que dans le cadre de mon choix théorique positionnel/géométrique, d'une part, le verbe ne sous-catégorise pas le sujet, et, d'autre part, les entrées lexicales correspondent à des verbes précis (et non à des structures de phrases).

Afin de pouvoir fournir une base solide en entrée à l'analyse des UVPL, j'ai procédé à une classification manuelle de plusieurs centaines de verbes, provenant de sources variées (BOONS et al. 1976 ; GROSS 1975), Greidanus (1990), *Petit Robert, Trésor de la Langue Française* (TLF)). Ceci a permis d'effectuer une validation théorique supplémentaire, par rapport à celle qui existait à mon arrivée au Centre d'ATO.

Parallèlement à cela, j'ai opté pour une analyse de la sous-catégorisation verbale (SCV) par niveaux *stratifiés* (Putnam, 1970, in (PANCKHURST 1993b, p. 67)). Ma SCV distinguait cinq strates (STR), définies de manière indépendante (PANCKHURST 1994e, p. 113-114), *cf.* fig. 3.2) :

Sous-catégorisation verbale

La définition des strates (1 à 5) est liée au cadre théorique positionnel choisi :

- la STR-1 contient la part *interprétative* des types de verbe. Elle constitue la strate pivot, celle comportant le plus d'indices sémantiques, liant la syntaxe à la sémantique.
- la STR-2 contient également de l'information sémantique ; si la syntaxe permet de fournir les positions (par exemple, GV, V et Objet), elle ne permet pas de déterminer le *nombre d'arguments* d'un verbe.
- les STR-3, STR-4 et STR-5 déterminent le type de position *Objet* et son contenu ; il s'agit précisément d'un apport d'information syntaxico-lexicale.

Tableau 3.2 – Classification de la SCV par strates

Strates	Définition	Contenu spécifique
STR-1	Types de verbes (strate pivot : lieu de connexion syntaxe-sémantique)	Prédicatif ; argumental ; opérateur
STR-2	Nombre de compléments/arguments (indices sémantiques)	Valences : valence 0 ; valence 1 ; valence 2 ; valence 3
STR-3	Contraintes sur l'occupation de la position droite du GV	Sélection du complément/argument : 1. par le verbe directement (seul) ; 2. via une préposition-opérateur (via prép-op)
STR-4	Sélection catégorielle	GN ; GAdv ; GAdj ; GP ; GS
STR-5	Contraintes spécifiques sur la molécule lexicale	comp-de ; compØ ; sujpl ; ind ; subj

Plus précisément, dans (Panckhurst 1993b, 1994e), est explicité le contenu du tableau de classification des strates ci-dessus :

> ***STR-1 : Types de verbes***
> Le verbe prédicatif traite ses compléments comme des attributs. Quand il a des compléments (ce n'est pas le cas pour un verbe comme *exister* par exemple), la réalisation de ceux-ci est obligatoire (*avoir, falloir, importer, mesurer, peser, ressembler, rester*, etc.). Un verbe argumental traite ses compléments comme des arguments, la réalisation de ceux-ci étant obligatoire ou optionnelle. La classe des verbes argumentaux est très vaste : *manger, penser, oublier, donner, compter (sur)*... Un verbe opérateur ne traite son complément ni comme un attribut ni comme un argument. Il est « transparent » aux relations avec le sujet et se construit uniquement avec un verbe à l'infinitif (*arrêter, commencer, finir*, etc.).
> ***STR-2 : Valences***
> On admet quatre valences (à l'exclusion du sujet). Un verbe de valence 0 n'exerce aucune contrainte (formelle ou interprétative) sur la position sœur du GV (*dormir, exister*, etc.) ; un verbe de valence 1 exerce une contrainte sur la position sœur du GV sur un complément (*devenir, oublier, commencer*, etc.), un verbe de valence 2 sur deux compléments (*donner*, etc.) et un verbe de valence 3 sur trois compléments (*débarquer, envoyer*, etc.)
> ***STR-3 : Contraintes de sélection pour l'occupation***
> Le verbe sélectionne seul son complément/argument (*manger un couscous, donner un cadeau à son amant, oublier de venir*, etc.) ou via une préposition opérateur (*songer aux vacances, s'apercevoir de sa maladresse*, etc.)

> ***STR-4 : Sélection catégorielle***
> GN = Groupe nominal ; GAdv = Groupe adverbial ; GAdj = Groupe adjectival ; GP = Groupe prépositionnel ; GS = Groupe phrastique. La sélection de la préposition est proposée au sein du GP : datif (« à » – *Il donne un cadeau à son amant*) ; loc (*Il habite à Montpellier*) ; autres (*Il opte pour la retraite*). Le GS se sous-divise en phrases fléchies (S-fl) et non-fléchies (S-nfl) ; le S-fl se divise en complétive (S-que – *Pierre demande que Marie vienne*), interrogative (autre que « si ») /relative (S-ie – *Pierre demande/ignore/sait qui vient*), interrogative en « si » (S-si – *Pierre demande si Marie vient*) ; le S-nfl se divise en phrase infinitive déclarative (S-inf – *Pierre veut partir*) et en phrase infinitive interrogative (S-ie – *J'ignore qui regarder*).
> ***STR-5 : Contraintes molécule lexicale***
> Comp-de : complémenteur plein (*Il oublie de fermer la porte à clef*) ; compØ : complémenteur vide (*Elle a préféré s'abstenir*) ; sujpl : sujet « plein » (*Il regarde Marie tomber*) ; ind : indicatif (*Il s'aperçoit qu'il pleut*) ; subj : subjonctif (*Il veut que tu viennes*).
>
> (Panckhurst 1994e, p. 114-115).

Dans (Panckhurst 1995c, p. 201-202), je fournis des exemples détaillés de classification de verbes correspondant aux strates, repris ci-dessous :

> Je distingue trois types de verbes : prédicatif, argumental, opérateur. Sommairement, les verbes prédicatifs traitent leurs compléments comme des attributs. Ils correspondent surtout aux verbes copules et métrologiques.

Prédicatif	
valenceo	Dieu existe
valence1-seul-GP	Il ressemble à Marie Il compte parmi les meilleurs
valence1-seul-GAdj	Il pèse lourd
valence1-seul-GAdv	Il pèse beaucoup
valence2-seul&seul	Ceci semble intéressant à Paul

Les verbes argumentaux, dont la classe est très vaste, traitent leurs compléments comme des arguments :

	Argumental
valenceo	Il marche, Il dort
valence1-seul-GN	Il admire Marie Il mange un gâteau Il nomme Marie présidente
valence1-seul-GP-Sélection prép-datif(à)	Il téléphone à son ami
valence1-seul-GP-Sélection prép-loc	Il habite à Paris
valence1-seul-GP-Sélection prép-autres	Il opte pour des vacances en Nouvelle- Zélande Il compte sur ton aide
valence1-seul-GS-Sfl-Sque-ind/subj	Il oublie que tu viens Il veut que tu viennes Il comprend que tu es/sois venu
valence1-seul-GS-Sfl-Sie	Il ignore qui vient
valence1-seul-GS-Sfl-Ssi	Il ignore si Marie vient
valence1-seul-GS-Snfl-Sinf-comp-de	Il oublie de fermer la porte à clef
valence1-seul-GS-Snfl-Sinf-compo	Il préfère s'abstenir
valence1-seul-GS-Snfl-Sinf-sujpl	Il regarde Marie tomber
valence1-seul-GS-Snfl-S-ie	Il ignore qui regarder
valence1-via prép op-GN	Il songe à Paul Il renonce à sa venue Il s'aperçoit de son erreur Il pense au travail
valence1-via prép op-GS-Sfl-Sque-ind/subj	Il s'aperçoit qu'il se trompe Il pense qu'il travaillera
valence1-via prép op-GS-Snfl-Sinf	Il pense à travailler
valence2-seul&seul	Il donne un cadeau à son amant
valence2-seul&prép op	Il prévient Paul de son départ Il parle à Marie de son projet
valence3-seul&seul&seul	Jean envoie des marchandises à Marie à Montpellier Jean transporte des ordinateurs de Cordoue à Montpellier Max balade le spot de la porte à la fenêtre (*cf.* Guillet, Leclère)

> Les verbes opérateurs ne traitent leurs compléments ni comme des attributs ni comme des arguments. Ils sont « transparents » aux relations avec le sujet et se construisent uniquement avec un verbe à l'infinitif.

Opérateur	
valence1-seul-GS-Snfl-Sinf	Il commence à travailler Il arrête de travailler
valence1-seul-GP-Sélection prép-loc	Il court chercher le pain

(Panckhurst 1995c, p. 201-202), .

L'analyse stratifiée de la SCV fournit le contenu pour l'implémentation. Ensuite, j'ai élaboré un outil informatisé, intitulé Scatlex, (*Dispositif automatisé de classification lexicale pour la sous-catégorisation verbale*), à l'aide du langage de programmation FX, résumé ci-dessous. *Scatlex*

> Scatlex propose un protocole de reconnaissance et d'assignation des propriétés constructionnelles virtuelles d'un verbe. Dans la perspective d'un analyseur, Scatlex fonctionne comme un dispositif d'aide en fournissant un input d'information lexicale sous forme de schémas constructionnels. Par ailleurs, Scatlex permet de constituer une grande base lexicale [12] correspondant à l'explicitation du système verbal français. (Panckhurst 1993b, p. 61).

L'outil informatisé est représenté sous forme d'un arbre, basé sur une taxinomie (Panckhurst 1994e, p. 132), et reproduit en page suivante.

L'utilisateur, via une interface homme-machine, parcourt l'arbre et assigne des propriétés spécifiques aux verbes qu'il choisit de traiter, via l'aide proposée par l'outil :

> L'avancement du parcours au sein de l'arbre dépend crucialement de critères et de tests précis. L'utilisateur est constamment sollicité dans le choix de la direction à suivre. Ceci est fondamental, car sans les critères/tests qui sont rattachés à chaque nœud de l'arbre, le déplacement voire la constitution finale des entrées n'est pas envisageable. Techniquement ici, le trajet est quelque peu « déterministe ». Bien que l'on ne réfère pas au déterminisme au sens fort ici, le dispositif doit tout de

12. À cette époque, l'accès aux données massives (big data) n'était pas encore à l'ordre du jour. Je reviendrai sur ce point (p. 82).

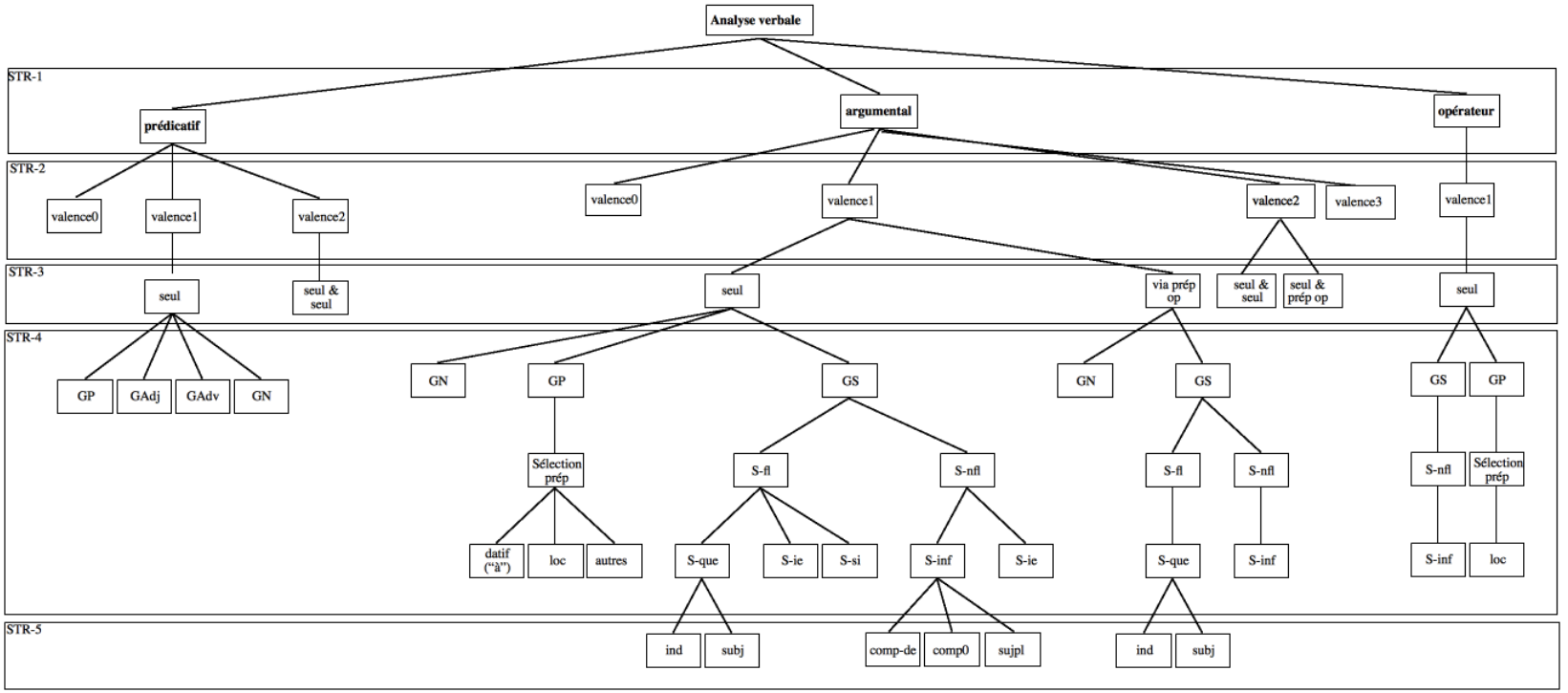
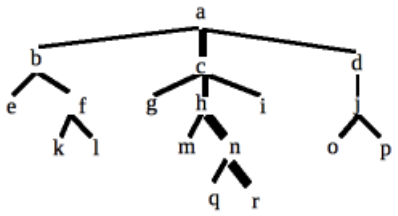


FIGURE 3.3 – Dispositif automatisé de classification lexicale pour la sous-catégorisation verbale en français (PANCKHURST 1994e, p. 132)

> même aider l'utilisateur à trouver le chemin correct dès le premier essai. Aucune stratégie de contrôle automatique n'a été programmée, car Scatlex ne descend pas dans un branchement pour ensuite effectuer un retour-arrière en cas d'échec, car la notion d'échec n'existe pas ; toutes les décisions sont à prendre par l'utilisateur qui est lui-même guidé par l'outil (bien sûr, si l'utilisateur se trompe, il a la possibilité d'effectuer un backtrack manuel, afin de ne pas être obligé de reprendre la totalité de sa classification !). (Panckhurst 1994e, p. 116-117).

Je fournis les explications de l'implémentation informatique (Panckhurst 1993b, p. 65-67) :

> Techniquement, le parcours dans l'arbre ne suit pas une stratégie de contrôle automatique tel qu'il pourrait être envisagé dans un arbre décisionnel classique. Dans l'arbre ci-dessous, si le chemin que je désire trouver en sortie finale correspond au tracé noirci, Scatlex ne descend pas dans un autre branchement, par exemple a : b : f : k pour effectuer ensuite un retour-arrière en cas d'échec. Le dispositif doit aider l'utilisateur à trouver le chemin a : c : h : n : r dès le premier essai, ainsi remplissant l'idéal déterministe : (Panckhurst 1993b, note 19, p. 67).



J'illustre, par photos d'écran, le suivi du déroulement de la classification verbale, effectuée en interface homme-machine (Panckhurst 1994e, p. 118-124). Enfin, je présente quelques résultats (à partir d'une classification d'environ 500 verbes) (Panckhurst 1995b) et rappelle la définition suivante du déterminisme pour un analyseur à l'époque — qui convenait effectivement pour l'utilisation de Scatlex :

> Parsers that (almost) never get fooled by local ambiguities, never have to backtrack, never change their mind, but which march inexorably to the single, correct, structural description (Gazdar et Mellish 1989) in (Panckhurst 1995b).

Par ailleurs, il existe une identité entre le dispositif et la structure de la donnée au sein de Scatlex :

> L'arbre analytique est censé contenir l'information nécessaire et suffisante pour assigner à un verbe ses propriétés d'analyse verbale (AV). Ainsi, en fournissant en entrée une forme lexicale de verbe, après parcours dans l'arbre, et cumul du savoir apporté au sein des différentes strates, une entrée lexicale verbale est créée. Par exemple, pour la description de la forme lexicale du verbe *donner*, on parcourt les strates, en cumulant l'information suivante :
>
> STR-1 verbe argumental (argu)
> STR-2 deux arguments (val 2)
> STR-3 le verbe sélectionne directement son complément/argument (seul), et ce, pour les deux arguments
> STR-4 SN ; SPrép
> STR-5 non pertinent
>
> avant que ne soit construite en sortie résultante l'entrée lexicale verbale correspondante :
>
> *donner* =>argu, val 2, seul + seul, SN ; SPrép
>
> Dans sa forme réelle, l'entrée lexicale est représentée de la manière suivante :
>
> ```
> (*donner* nil (argu nil (val2 nil (seul nil gn)
> (seul nil (gp nil datif))))(personne nil *rp*)
> (commentaire nil *1000F à Sidaction*))
> ```
>
> L'information cumulée à partir des strates constitue la structure de la donnée, c'est-à-dire l'entrée lexicale verbale. C'est en ce sens qu'il y a identité entre le dispositif et la donnée à construire. De surcroît, toutes les entrées lexicales verbales sont virtuellement contenues dans l'arbre de construction. Si le dispositif contient une trentaine de chemins effectivement distincts, pour chaque verbe un seul ou plusieurs chemins peuvent être empruntés. La combinatoire possible au niveau du passage de la structuration arborescente elle-même à la construction réelle de données lexicales est donc statistiquement élevée.
>
> (Panckhurst 1995b, p. 176-177).

Puis, comme la plupart des verbes en français présentent des parcours « multiples », dans l'arbre, et ce à tous les niveaux du système stratifié, je fournis quelques indicateurs sur les classifications mixtes que j'avais retenues, qui ne sont pas obligatoirement synonymes de polysémie : *polytype* (par exemple, *mesurer* qui peut être prédicatif : *Il mesure 1m90*, ou argumental *: Il mesure l'étagère) ; polyvalence* (par exemple, *tourner* qui peut être valence 0 : *La roue tourne* ou valence 1 : *Pierre tourne la roue*) ; *polycontraintes sur l'occupation*, *polycatégorie polycontraintes spécifiques* [13] (Panckhurst 1995c, p. 202-203)).

Si j'ai (quelque peu) détaillé ce travail post-doctoral sur la SCV, incluant *strates* et *arborescence*, en lien avec une application informatisée écrite en FX-LISP, c'est pour rappeler que les implémentations informatisées des phénomènes de ce type pour le français étaient plutôt rares, à l'époque, en dehors de l'énorme entreprise distributionnelle de Gross dans un effort d'énumération lexicale exhaustive des phénomènes. C'est également pour cette raison que j'ai assuré la direction d'un numéro des *Cahiers de Praxématique* en 1994, intitulé *Autour du verbe : théorie et implémentation*, (Panckhurst 1994b), pour que puissent y être recensés quelques approches théoriques, outils implémentés et modèles proposés, pour l'époque, par (Segond 1994), (Blache 1994), (Bès 1994), (Alcouffe et Falcoz 1994), (Panckhurst 1994e), (Ouellet et al. 1994), (Franckel 1994) [14].

Bien évidemment, de nombreux travaux, ressources et dictionnaires ont été publiés depuis les années 1990 sur le verbe et la sous-catégorisation verbale pour le français. Pour n'en citer que quelques-uns, par ordre chronologique, et concernant surtout les dictionnaires et les applications logicielles, voici quelques repères non exhaustifs.

> 1975 LADL – *Lexique-grammaire* : dictionnaire électronique (de prime abord, des verbes français) sous forme de tables, (initié par M. Gross (Gross 1975) et en constante évolution depuis cette date, au sein du Laboratoire d'informatique

13. Ces derniers aspects concernant les polycontraintes/polycatégories, impliquant ou non une sélection « seule » ou via une « préposition-opérateur » me paraissent peut-être un peu moins défendables aujourd'hui.

14. Concernant l'implémentation et la formalisation des tables de Gross, dans ce numéro des *Cahiers*, on se reportera aux articles (Alcouffe et Falcoz 1994 ; Bès 1994). Pour des travaux plus récents dans ce cadre, *cf*. la note 15.

Gaspard-Monge). [15] Malheureusement, les tables n'étaient pas encore disponibles pour consultation libre dans les années 1990.

Des travaux de recherche théorique et appliquée portent sur les tables, entre autres : (Constant et Tolone 2010 ; Danlos et Sagot 2007 ; Gardent et al. 2006 ; Laporte 2010 ; Laporte et al. 2013 ; Tolone et Sagot 2011). On se reportera également à : http://infolingu.univ-mlv.fr/ et, plus précisément, pour Unitex [16], à http://igm.univ-mlv.fr/~unitex.

(Panckhurst 1994b), *Autour du verbe : théorie et implémentation,Cahiers de Praxématique* ;

Pour les outils implémentés, on pourra également consulter, entre autres [17] : (Dubois et Dubois-Charlier 2013 (1997)) *Les Verbes Français* (1997, version informatisée 2013) (avec classification sémantique) http://rali.iro.umontreal.ca/Dubois ; 2003 (2010), DICOVALENCE Karel van den Eynde, Piet Mertens, http://bach.arts.kuleuven.be/dicovalence/ ; 2007, Rauzy et Blache, « Un lexique syntaxique des verbes du français : VfrLPL » ; 2008, Cédric Messiant, Thierry Poibeau, « LexSchem : A Large Subcategorization Lexicon for French Verbs » ; 2008, Anna Kupsc, Anne Abeillé : « Treelex : a subcategorization lexicon automatically extracted from a French Treebank » ; (Sagot 2010) ; (Ligia-Stela et Fuchs 2013). (Laporte 2015) indique, entre autres, les dictionnaires (français/anglais) suivants : FrameNet framenet.icsi.berkeley.edu, VerbNet, http://verbs.colorado.edu/verb-index/ (*cf.* également les travaux de Danlos), ComLex, Meaning-Text, etc.

15. *Cf.* la page officielle des lexiques-grammaires : http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html.

16. « Unitex est un ensemble de logiciels permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire. Elles sont issues de travaux initiés sur le français par Maurice Gross au Laboratoire d'Automatique Documentaire et Linguistique (LADL). Ces travaux ont été étendus à d'autres langues au travers du réseau de laboratoires RELEX. [...] Les grammaires sont des représentations de phénomènes linguistiques par réseaux de transitions récursifs (RTN), un formalisme proche de celui des automates à états finis. » (Paumier et Martineau 2016, p. 13), (Paumier 2003).

17. Seule une sélection des références suivantes figure en bibliographie.

3.3.1.5 Unités verbales polylexicales (UVPL)

La recherche initiale — sur le GV et la sous-catégorisation verbale — aurait pu donner l'illusion de constituer un détour par rapport au projet de la bourse post-doctorale concernant les unités verbales polylexicales (UVPL). Cependant, cette étape préalable était primordiale pour la bonne compréhension et le traitement efficace des UVPL. Pourquoi ? Dans l'approche *positionnelle*, au niveau syntaxique, j'estimais que l'UVPL était dotée des propriétés structurales d'un groupe verbal (GV) « normal » (c'est-à-dire un GV dont la molécule lexicale ne peut se rapporter à une UVPL) :

Dans l'UVPL « perd les pédales » (figure 3.4), « les pédales » occupe la même position que « son fils » dans « aime son fils ». Le travail sous-jacent au repérage et à l'extraction des UVPL concernait directement la structuration du groupe verbal (GV) et la sous-catégorisation verbale.

Par la suite, j'ai effectué un recensement de différentes approches concernant les problèmes posés par les UVPL. Si certains repérages syntaxiques sont envisageables pour les UVPL ne contenant pas d'article (*faire marche arrière, faire table rase, prêter main-forte*), le constat global était qu'un traitement sémantique s'imposerait, afin de résoudre le problème très complexe posé par la reconnaissance de ces unités [18]. En effet, comment reconnaître, dans une approche positionnelle, que l'on a affaire à une UVPL (*perdre les pédales)* plutôt qu'à un syntagme verbal ordinaire (*perdre son porte-monnaire*) [19] ?

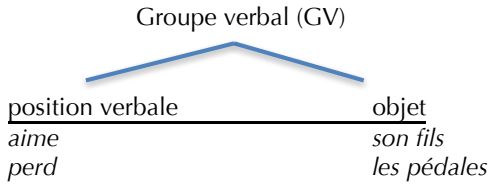


FIGURE 3.4 –

18. Contrairement au traitement des unités nominales polylexicales (*pomme de terre*, *système de gestion de base de données, garde forestier*, etc.), pour lesquelles un repérage syntaxique peut être partiellement envisagé (David 1992).

19. Rappelons qu'il s'agissait de travailler avec un analyseur morpho-syntaxique qui ne fonc-

Suite au travail de description théorique et de développement d'algorithmes effectués lors de l'année post-doctorale, il était prévu que je reste une deuxième année, afin de produire un module de repérage et d'extraction automatique des UVPL (TERMINOV). Cependant, ce travail a été interrompu, car j'ai obtenu mon poste de maître de conférences en septembre 1992 à Montpellier.

Trois points me paraissent essentiels à retenir du premier travail post-doctoral concernant la SCV :

1. Malgré le cadre théorique pré-défini, j'ai décidé, au final, pour l'implémentation, de défendre l'hypothèse de l'étude de données (ici, la classification lexicale) *indépendamment d'un formalisme grammatical*.
2. La constitution lexicale verbale était élaborée à partir de *contrastes* (figurant à chaque nœud de l'arbre), au lieu que ce soit à partir d'une série de *combinatoire*s.
3. Malgré la différenciation d'approches théoriques entre écoles, en ne me situant pas au sein d'un cadre grammatical prédéfini, le point 1) permettait d'envisager une ouverture éventuelle non-négligeable : « Il serait intéressant [...] de confronter cette recherche avec les résultats [...] sur l'informatisation des tables LADL [20][...] » (Panckhurst 1994e, p. 131).

Comme le fait remarquer à juste titre (Léon 2010), l'implémentation informatique exige de toute manière des ajustements :

> Les nécessaires ajustements et compromissions [...] ont toujours été, et sont encore actuellement, le lot des dispositifs de traitement automatique des langues.
> (Léon 2010, supra).

Le parfait respect par rapport à un cadre théorique sous-jacent est relativement utopique (même si j'ai défendu très longtemps la position qui consiste à exiger, dans des discussions face à des informaticiens « purs et durs » — en vérifiant parfois jusque dans le code informatique même — que la théorie linguistique soit parfaitement respectée). Mais le point 1) que j'évoque ci-dessus va au-delà de cette discussion. Depuis le début de ma carrière de chercheur — après avoir bien

tionnait pas avec des dictionnaires électroniques. Par conséquent, une liste d'expressions figées n'était pas à sa disposition.

20. *Cf.* également (Alcouffe et Falcoz 1994), in (Panckhurst 1994b, (coor)).

comparé différentes approches théoriques pour chaque travail de recherche — je suis constamment revenue à une position fondamentale, mienne, **hors cadre**.

Ceci m'amène au point 2). Le choix d'une arborescence pour le traitement de la SCV, basée sur des *contrastes*, en m'opposant à un traitement incluant une accumulation de *combinatoires* est **une position que je ne défendrais pas nécessairement aujourd'hui**. Si tout le système de SCV pour le français était effectivement potentiellement inclus dans l'arbre, via une classification active à l'aide de parcours simples ou complexes, *mono* ou *poly* (*polytypes, polyvalents, polycontraignants* (sur l'occupation argumentale ou sur la molécule lexicale), *polycatégoriels* (Panckhurst 1995c, p. 202-203), la connaissance bâtie, sous forme d'entrées lexicales verbales — exigeant d'ailleurs des connaissances linguistiques importantes de la part de l'opérateur humain — était-ce si différent, au fond, quant aux résultats, d'une approche de type lexique-grammaire ?

Je reviendrai sur cette question un peu plus tard (§ 3.3.1.6), lorsque j'évoquerai les mutations en linguistique-informatique et en TAL, concernant l'un des changements à mon avis majeurs, menant au choix de systèmes informatisés qui n'effectuent plus des analyses morpho-syntaxiques (de type ALSF, qui prenaient en considération des *positions*, et une distinction entre lexique et syntaxe), pour privilégier alors des systèmes plus simples techniquement (par exemple, inspirés des automates à états finis), basés sur un étiquetage POS (parties du discours). *Cf.* (Léon 2015), (Cori et al. 2008a).

D'ailleurs, bien que l'outil informatisé Scatlex ait été utilisé par des chercheurs au Centre d'ATO pour catégoriser plus de 500 verbes du français dans une première phase de test (Panckhurst 1995b), l'entreprise générale, concernant l'élaboration d'un véritable analyseur pour le français (ALSF) a été réorientée, voire abandonnée par la suite (Cori et Marandin 1993 ; Marandin 1993, 1997) (Léon 2010) :

> Le projet d'analyseur syntaxique s'autonomise du projet d'AD et devient un enjeu pour la théorie syntaxique elle-même jusqu'à ce que (Marandin 1993), montrant qu'un parseur syntaxique ne peut pas être une instance d'expérimentation pour la syntaxe, remette en cause la nécessité d'une interaction entre informatisation et théorie syntaxique. Le projet d'analyseur est à son tour abandonné.
>
> (Léon 2010).

Marandin (1993) annonce :

> Cet article est le résultat d'une longue période de sueurs et de colères froides face aux confusions et aux dialogues de sourds qui entourent la conception d'analyseurs syntaxiques. (Marandin 1993, p. 6).

Le chercheur se re-questionne (après de longues années de recherche appliquée) sur le statut de l'analyseur syntaxique comme dispositif expérimental pour la théorie syntaxique, et il en conclut : « un parseur ne peut pas être une instance d'expérimentation pour la syntaxe » (Marandin 1993, p. 5). Plus précisément :

> L'analyseur prend le statut d'une machine industrielle dans le TAL ; il est soumis aux impératifs de la technologie, ce ne sont pas toujours ceux d'un programme de recherche. Il est probable que l'analyseur n'a aucun statut en théorie de la syntaxe ; il n'en a assurément pas dans le cadre d'une conception où la syntaxe est un principe organisationnel des énoncés qui ne se réduit pas aux règles de combinaison des mots. (Marandin 1993, p. 31).

Dans le point 3) supra concernant mes premiers travaux post-doctoraux, on aperçoit également mon hésitation en 1994. Tout en m'étant inspirée d'une approche théorique précise, j'évitais de m'enfermer dans un formalisme donné, et je prônais une ouverture vers d'autres systèmes, d'autres modèles. Je crois qu'on peut entr'apercevoir ce qui commençait à me caractériser en tant que chercheur — et qui est peut-être vrai encore aujourd'hui : mon approche *variée* tendant vers une attitude sans doute un peu *empiriste* de la recherche.

J'allais d'hésitations en modifications. Dans (Panckhurst 1994e), mon appellation a évolué en *analyse verbale* et ma définition a légèrement changé :

> Dans le cadre du présent travail le terme *sous-catégorisation verbale* (SCV) a été retenu dans un premier temps. J'aimerais proposer désormais un autre terme : *analyse verbale* (AV). Pourquoi ce changement ? Précisément car dans le cadre présent la recherche proposée va au-delà d'une catégorisation stricto sensu, et incorpore en plus des aspects lexicaux et syntaxiques habituels, un pivot syntaxique-sémantique, permettant une classification des types de verbes [...]. Outre la définition des types de verbes [...] il nous importe de préciser :

> Les liens permettant de cerner les compléments nécessaires à l'établissement de la signification lexicale d'un verbe. (PANCKHURST 1994e), p. 110.

Entre autres facteurs, j'étais installée à Montpellier depuis un an et demi, et les recherches *praxématiques* [21] de mes collègues commençaient à m'interpeller :

> La signification ou identité lexicale du verbe est primordiale, et ce pris en dehors d'un contexte phrastique ; il ne s'agit pas de confondre les propriétés lexicales d'un verbe qui sont *virtuelles*, et *l'interprétation phrastique* que l'on peut en faire par la suite en contexte. (PANCKHURST 1994e), p. 110.

Dans (PANCKHURST 1995c), j'utilise une qualification typique de la praxématique :

> Un verbe peut être doté d'un certain savoir qui lui est inhérent, intrinsèque, et ensuite se voir attribuer un comportement différencié en discours *actualisé*. (PANCKHURST 1995c), p. 204.

Pour les recherches sur les UVPL débutées pendant la période post-doctorale puis menées pendant plusieurs années par la suite, je ferai essentiellement référence à trois publications (PANCKHURST 1996c, 1998b, 2001b).

Actualisation

En 1998, Barbéris, Brès et Siblot coordonnent l'ouvrage *De l'actualisation*, publié aux éditions CNRS, qui correspondait à l'aboutissement de quelques années de recherches effectuées dans le cadre du séminaire « Discours, textualité et production de sens » au sein du laboratoire Praxiling. Dans (PANCKHURST 1998b), j'indique que la définition d'actualisation empruntée à (ARRIVÉ et al. 1986, p. 32), englobant tout à la fois *actualisation, énonciation* et *référenciation*, était suffisante pour mon traitement des UVPL :

21. « La linguistique praxématique, née dans les années 70 à Montpellier, sous l'égide de Robert Lafont, et développée dans le cadre de l'UMR CNRS 5475, se donne pour tâche d'analyser la production de sens, à partir du repérage de ses marques dans le discours produit. » (Détrie, Siblot, Verine, 2001, p. 8). Cette théorie linguistique s'inscrit dans « un cadre anthropologique et réaliste » (idem, p. 261). « Elle met l'accent sur le fait que les mots effectuent des découpages non du monde réel, mais du monde vécu par les sujets parlants, les mots reflétant, véhiculant avec eux des expériences multiples du monde sensible, expériences anthropologiques, manipulatives ou culturelles. Les sujets parlants sont alors pris dans une double dialectique — celle du social et de l'individuel, celle du langage et du réel. » (Catherine Détrie, communication personnelle, le 8 juillet 2016).

> L'actualisation est l'opération par laquelle un morphème de la langue passe dans le discours. Cette opération, liée au phénomène de l'énonciation, s'accompagne de la référenciation : l'élément linguistique qui, en langue, n'a pas de référent, s'en voit affecter un dans le discours. (Arrivé et al. 1986, p. 32).

Ce positionnement diffère de celui de mes collègues, pour lesquels *actualisation* et *énonciation* se distinguent nécessairement (*cf.* (Détrie et al. 2017)). Pour ma part, ne voulant pas favoriser l'une ou l'autre terminologie, j'ai opté pour celle de (Milner 1978, p. 332), à savoir *référence actuelle* vs. *référence virtuelle* (Panckhurst 1998b, p. 163). Le lecteur pourra se reporter à (Panckhurst 1998b, p. 161-169), pour ces considérations théoriques, incluant un état de l'art succinct sur les UVPL (correspondant à d'autres appellations courantes : *locutions/lexies/périphrases verbales, expressions/locutions figés/idiomatiques*, etc.). Je m'interrogeais sur le degré de flexibilité dans la production-création des UVPL, et notamment les possibilités d'insertion-exclusion de l'article. Puis, la dimension sémantique de l'UVPL m'interpellait, plus précisément, comment reconnaître que l'on a affaire à une UVPL (*perdre les pédales*) correspondant à « une unité de sens », plutôt qu'à un syntagme verbal ordinaire (*perdre son porte-monnaie*) dont le sens est littéral ? Enfin, dans un cadre de TALNE, la préoccupation du travail descriptif linguistique préalable était toujours prégnante :

> Pour mener à bien une implémentation d'outil permettant d'extraire des UVPL d'un texte de manière automatique, il sera nécessaire de sérier les niveaux de formalisation des données linguistiques ; à cette fin, qui dépasse largement le seul cadre de cet article, il sera primordial de laisser l'ordinateur de côté dans un premier temps pour mettre l'accent sur le travail descriptif, crucial. (Panckhurst 1998b, p. 164).

Dans (Panckhurst 1996c), l'approche descriptive était également préconisée :

> Notre pari est le suivant : tant que nous ne nous préoccuperons pas de la description linguistique (syntaxique, mais surtout sémantique voire pragmatique) de ces unités de manière approfondie et fine, nous élaborerons des outils implémentés superficiels, non « robustes ». (Panckhurst 1996c, p. 468).

La conclusion logique de (Panckhurst 1998b) était que la *référence actuelle*, l'*actualisation* du nom n'est jamais possible au sein des UVPL :

> Si l'analyse de l'UVPL se révèle complexe, c'est que *l'unité* ne naît pas par simple contiguïté formelle, mais résulte de « processus intellectuels conceptuels et symboliques » (Gréciano 1982, p. 399) et constitue une « opération mentale » (ibid., 1982, p. 322). Ce point est fondamental ; l'UVPL ne correspond pas à une suite d'éléments juxtaposés représentant une compositionalité quelconque. On fera nécessairement la différence entre sens et référence (*cf.* Frege) d'une part, et sens individuel / sens global d'autre part :
>
> Tout en comprenant le sens individuel des mots, on peut très bien, hors de la circonstance, ne pas comprendre le sens qui résulte de l'assemblage des mots (Benveniste 1974, p. 226).
>
> La non-compositionalité impliquant l'*unité de sens* des UVPL nous permet d'avancer les hypothèses (1), (2) et (3) suivantes :
>
> (1) S'il y a unité de sens, la *référence actuelle* ou l'*actualisation* du nom en discours est écartée.
>
> L'hypothèse (1) est étayée par le fait qu'on ne puisse pas interroger ou traiter l'anaphorisation du nom apparaissant en position « objet » *: Il a perdu les pédales* → **Qu'a-t-il perdu ?* ; **Je les ai retrouvées* [22].
> À l'inverse :
>
> (2) La *référence actuelle* ou l'*actualisation* du nom en discours écarte l'interprétation d'UVPL pour une unité donnée.
>
> Par exemple, si on dote d'un article une UVPL ne permettant pas d'insertion d'article a priori, le nom se voit obligatoirement actualisé en discours : *Il a pris la racine* ne peut pas être interprété comme signifiant *Il a pris racine*.
> Enfin :
>
> (3) La présence d'un article n'indique pas obligatoirement une *référence actuelle* ou une *actualisation* du nom en discours.

22. Dans ces exemples, l'astérisque indique l'impossibilité d'une lecture UVPL, et non pas que la séquence est agrammaticale. Ces types de tests sont classiquement utilisés dans le cadre des LG également (précision apportée par Panayota Kyriacopoulou).

> Autrement dit, une UVPL permettant l'insertion d'un article introduit une ambiguïté ; l'*actualisation*, la *référence actuelle* est possible ou non, selon le contexte (*demander confirmation, demander la confirmation ; prendre place, prendre la place*…). Par contre, s'il y a actualisation, le sens de l'UVPL est suspendu. Dans ce cas, la présence de l'article « entraîne », « fait glisser » l'UVPL vers une lecture littérale où le nom est effectivement actualisé en discours — la lecture « oscille » alors entre unité de sens (sans *référence actuelle* du nom) et sens littéral (avec *référence actuelle* du nom). Il ne faut donc pas confondre une UVPL et un GV « en provenance » d'une UVPL, ayant regagné un sens littéral, ou ayant adopté un sens sémantiquement varié, dont jouent fréquemment les médias [23] :
>
> — Exemple du Canard Enchaîné du 25.10.95 :
> *Juppé n'est plus dans ses bottes ! Il est dans ses pompes…à phynances.*
>
> — Dans le film *Blazing Saddles*, il y a une séquence où un acteur donne un coup de pied dans un seau, avant que l'on ne voie mourir quelqu'un. Le jeu visuel est évident pour le spectateur anglophone — *to kick the bucket* est l'équivalent français de *passer l'arme à gauche*. L'UVPL est donc effective en tant qu'unité de sens (*X meurt*) mais devient également un GV détenant un sens littéral (*le coup de pied dans le seau*).
>
> (Panckhurst 1998b, p. 175-177).

Les recherches effectuées par les praxématiciens et mes propres pré-occupations TAListes — bien que très éloignées des leurs — ont pu être confrontées au sein du séminaire de Praxiling et dans l'ouvrage résultant. Même si je ne me suis jamais sentie praxématicienne, d'autres projets de publication ont également permis de maintenir un espace de discussion fructueux (*cf*. volet 1, Détrie et al. 2017, volet 2 Bres et al. 1999), voire un rapprochement partiel des chercheurs du laboratoire (*cf*. publications dans le cadre du volet 3, SMS, § 3.3.3).

Revenons-en aux problèmes spécifiques demeurant pour le repérage en TALN des UVPL (Panckhurst 1998b, p. 178, note 27) :

23. On pourra se reporter également, entre autres, aux travaux de (Kyriacopoulou 1989, 2010) pour le grec. Par ailleurs, une initiative multilingue récente en TAL, à propos des unités polylexicales (*multiword expressions*), verbales et autres types d'expressions, marie les approches formalisantes (HPSG, LFG, TAG, lexiques-grammaires) en analyse (*parsing*, incluant des techniques d'analyse symboliques, probabilistes et hybrides) et en ressources linguistiques (corpus et lexiques), *PARSEME–PARSing and Multiword Expressions* (Savary et al. 2015). Cela semble très prometteur.

La suite UVPL peut être formée d'un verbe suivi d'un nom. Encore faut-il désambiguiser une UVPL d'une unité nominale : *Il garde confiance vs. Un garde forestier*
Si le texte à analyser est écrit en style télégraphique, il est important de pouvoir distinguer l'utilisation d'un verbe et un nom inarticulé d'une part, d'une UVPL d'autre part : *Jeune fille donne cours d'anglais tous niveaux* *La patronne donne suite à votre demande toute affaire cessante*
Dans des exemples coordonnés, le verbe n'est pas obligatoirement répété : *Il avait faim et surtout très peur*
La désambiguïsation d'unités nominales / verbales est cruciale[24] : *Mettre en jeu des relations spatiales* → mettre en jeu (UVPL) + des relations spatiales → *mettre en + jeu des relations spatiales (unité nominale faussement repérée) *Tenir compte des résultats* → tenir compte (UVPL) + des résultats → *tenir + compte des résultats (unité nominale faussement repéré)

Dans (Panckhurst 1998b, 2001b) je m'intéressais aux UVPL dans l'optique de la construction d'un outil permettant le repérage et l'analyse automatique. Les approches par dictionnaires/lexiques électroniques, fournissaient déjà une aide précieuse pour ce faire, dans la mesure où les locutions/expressions figées insérées dans le dictionnaire/lexique pouvaient être ainsi repérées lors de l'étape de l'analyse automatique. Par exemple, on ne s'attendait pas à rencontrer **attraper un bouc par ses oreilles* plutôt que *prendre un taureau par les cornes* (Labelle, in Panckhurst 1996c, p. 470).

Si l'approche de type dictionnairique pouvait convenir pour le repérage d'unités extrêmement figées, j'ai souhaité proposer des *règles interprétatives*, permettant d'analyser des UVPL présentes sous forme de jeux de mots constamment utilisés par les médias :

> Juppé a du **plan** dans l'aile
> Ces Irlandais ne manquent pas d'**Eire**
> Canard Enchaîné, 29/11/1995, in (Panckhurst 1996c, p. 470).

Grâce à des tests de proximité de type phonétique/phonémique, des vérifications sur les homophones (*compte/conte* [kɔ̃t], synonymes (*pompes/bottes*) et des indications d'insertion incluses, notamment sur les noms, ces règles devaient potentiellement permettre des extractions automatisées plus efficaces.

> Unité potentielle, figurant dans le texte :

24. Exemples fournis par Sophie David, communication personnelle.

> → *avoir du **plan** dans l'aile*
> Première étape : recherche dictionnairique
> → *avoir du _____ dans l'aile*
> Deuxième étape : plan = [plɑ̃]
> Troisième étape : recherche de similitude
> → plan [plɑ̃] vs. plomb [plɔ̃] (Panckhurst 1996c, p. 472).

Les UVPL ne sont pas toujours figées et le repérage/extraction peut devenir fort complexe. J'avais constaté quatre cas de figure (*cf*. également les tables du Lexique-Grammaire, pour les expressions figées, http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html) :

1. **UVPL figée** (UF) : *mettre la charrue avant les bœufs*
2. **UVPL insertion impossible** (UII) : *prendre racine*
3. **UVPL insertion possible** (UIP) (avec une variation sémantique plus ou moins grande) : *rendre justice, rendre la justice, demander grâce, demander une grâce*.
4. **UVPL absence** (ou variation du type de déterminant) **impossible** (UAI) : *hocher la tête, *hocher tête, *hocher sa tête*).

Les catégorisations indiquées peuvent être agrémentées de marqueurs (à gauche et à droite du nom, par exemple) afin de permettre l'analyse automatisée des insertions (Panckhurst 2001b, p. 62).

3.3.1.6 Conclusion

Les recherches de cette première décennie ont été consacrées, dans un premier temps, aux propriétés d'analyse (ou sous-catégorisation) verbale aboutissant à un outil informatisé (Scatlex, programmé en FX/LISP), puis, dans un second temps, aux problèmes posés pour l'analyse automatique des UVPL. Après la description linguistique fournie pour l'étude des UVPL, et les règles interprétatives élaborées, leur implémentation informatique effective n'a jamais abouti. Sans doute parce que je commençais à consacrer davantage de temps à la direction du service METICE à l'université Paul-Valéry, au début des années 2000, et que mes

orientations de recherche ont pris d'autres directions (*cf.* volets 2 et 3, § 3.3.2, § 3.3.3).

Pendant les premières années de recherche post-doctorale, plusieurs publications indiquées au sein du volet 1 émanent du travail fourni pendant le doctorat. Je ne reprends pas le détail de leur contenu ici ; le lecteur pourra directement se reporter aux publications indiquées (PANCKHURST 1992a,b, 1994a).

Une dernière publication, que j'ai classée au sein du volet 1 (PANCKHURST 2003b), concerne le traitement automatisé de la glose [25]. Ma collègue Augusta Mela (MELA 2004, 2005 ; MELA et al. 2011), et moi-même avons développé des patrons d'extraction automatique sur des données étiquetées de manière morpho-syntaxique [26] à l'aide du langage de programmation Perl. Cependant, dans la conclusion de cet article, je mets en garde contre une approche de type patrons *vs.* une approche, par exemple, de type positionnel [27] : *Glose*

> Certains logiciels (DAVID 1993a,b) permettent de repérer des positions syntaxiques et non pas simplement des suites linéaires de catégories. C'est le cas du logiciel Termino [28], qui repère les candidats termes (de type nominal) au sein de SN.

25. Cet article aurait pu également être classé au sein du volet 3 (*cf.* § 3.3.3), qui concerne la CMO, le DEM, le DNM, car les données utilisées pour le traitement de la glose émanant de courriels, de forums et chats, sont issues du projet CMO. Cependant, j'ai préféré l'insérer dans le volet 1, étant donné sa dimension d'outil informatisé.

26. Le logiciel Cordial a servi pour l'étiquetage morpho-syntaxique. Puis, nous avons implémenté des patrons pour « approcher » la glose : 1) *N* ou *Adj* ou *V* suivi de « n'importe quelle chaîne (ne contenant pas *sens*) » suivi de *au sens* puis N ; 2) *N/V/Adj (ponct) prep=dans DET sens X*) ; 3) des combinatoires positionnels, SN/SV/UVPL/ + *c'est-à-dire* + qui/que/SV/SN/Sprep/PP en position adjectivale, etc. (PANCKHURST 2003b, p. 279, p. 290-292).

27. Je remercie Sophie David pour une discussion fructueuse à propos des exemples émanant de sa thèse sur les unités nominales polylexicales. Les exemples qui figurent dans cette citation sont les siens.

28. Le logiciel *Termino*, élaboré au Centre d'ATO, UQÀM, Montréal, par Sophie David, Lucie Dumas, Jean-Marie Marandin, Andre Plante, Pierre Plante, sous la responsabilité de Claude Ricciardi Rigault, est « un dispositif dont l'objectif est [...] la reconnaissance d'unités nominales polylexicales construites syntaxiquement », susceptibles de se lexicaliser (DAVID 1993a, p. 220). On pourra également se reporter à (SOUCHARD et al. 1997) pour une utilisation du logiciel *Termino*, et, entre autres, à (BOURIGAULT 1993 ; HABERT et JACQUEMIN 1993) pour des traitements automatiques de noms composés à cette époque. Pour un traitement spécifique en lexiques-grammaires, *cf.* le DELAC (COURTOIS et al. 1997) et les références complètes correspondantes : http://www-igm.univ-mlv.fr/~unitex/index.php?page=13. Dans un travail plus récent, KYRIACOPOULOU et MARTINEAU 2015 proposent un traitement d'extraction de « segments

> En partant d'un exemple comme : *un système de gestion de bases de données*, le logiciel repère la frontière droite (positionnelle) et permet d'extraire les candidats suivants : *bases de données, gestion de bases de données, système de gestion de bases de données*, mais il écarte automatiquement **système de gestion, *gestion de bases, *système de gestion de bases*. Une approche par patrons repère indifféremment (et par conséquent incorrectement) tous les exemples précédents.
> Par ailleurs, le logiciel positionnel ne peut pas systématiquement repérer les candidats correctement, mais pour des raisons linguistiquement valables. Par exemple, dans les deux exemples suivants, on constate qu'une information en provenance de la sous-catégorisation verbale (*parler de X à Y/ à Y de X*) doit être incorporée pour reconnaître la possibilité de synapse dans le cas « cause de divorce » et non pas dans le premier exemple « fille de programmation ».
>
> *Il a parlé à une fille de programmation*
> *Il a parlé d'une cause de divorce à sa sœur*
>
> Certaines ambiguïtés réelles ne peuvent être relevées par un logiciel de type Termino, à cause de l'identité syntaxique entre le candidat terme et une phrase canonique, et pour des raisons liées à la sous-catégorisation verbale. Dans l'exemple *J'ai rapporté un vase de Chine*, on ne sait pas, hors contexte, si le vase provient de Chine ou s'il s'agit d'un vase d'un type particulier (*cf.* par exemple, *j'ai ramené une belle assiette en porcelaine de Limoges de Paris*).
> Enfin, certains candidats termes n'en sont pas réellement car l'ambiguïté est réelle entre le terme (par exemple : *le mur du son)* et le SN classique (par exemple : *Le chat du voisin*).
>
> (Panckhurst 2003b, p. 287).

Si j'ai repris quelques notions évoquées par David et les chercheurs montréalais/parisiens avec lesquels je travaillais voici plus de 20 ans — dont les approches sont, à mon avis, tout à fait linguistiquement défendables à l'heure actuelle — c'est précisément car la linguistique-informatique a évolué, muté, depuis les années 1990. Depuis cette date, le traitement de grandes masses de données est devenu possible. Cela a provoqué, dans le domaine du TAL, l'arrivée accrue des linguistiques de corpus et des traitements statistiques/probabilistes [29]. Ceux-ci

complexes » pour enrichir les dictionnaires.

29. Roche 2004 propose un panorama des differentes approches linguistiques, statistiques

ont (parfois) tenté de supplanter, d'étouffer les approches syntaxiques et/ou lexicales. De mon point de vue, tout en apportant une dimension d'« outillage » intéressante, il n'en demeure pas moins qu'ils entretiennent le risque de refaire la roue.

(Léon 2015) évoque l'histoire du TAL comme étant non seulement « une histoire du récent mais également une histoire en cours » :

> On assiste actuellement à une accélération du changement de méthodes, entièrement parallèle aux développements technologiques, voire dominée par eux. La pression de la demande sociale et les ambiguïtés intrinsèques du TAL ont pour conséquence l'apparition de cycles d'abandon-reprise/oubli-redécouverte de méthodes qui vont en s'accélérant. (Léon 2015, p. 182).

Parfois, on ne sait même plus où on se situe, car il arrive que « l'alternance des modèles soit tellement rapide qu'on assiste à de véritables courts-circuits », poursuit Léon :

> Le 23 juin 2009, lors de la cérémonie du cinquantenaire de l'ATALA, un représentant allemand du TAL annonce que le courant actuel dominant en TA est fondé sur les méthodes statistiques, alors qu'au même moment, le représentant des États-Unis annonce que c'est la méthode *rule-based* (fondée sur des règles) qui a le vent en poupe, sans d'ailleurs que l'un ou l'autre fasse référence à la déclaration de l'autre. (Léon 2015, note 8).

Sans entrer dans le détail de l'*Histoire de l'automatisation des sciences du langage* — l'ouvrage éponyme par Jacqueline Léon (Léon 2015) fournit un excellent panorama du domaine — je reviens sur deux points qui m'interpellent particulièrement dans mon travail de chercheur, concernant cette première décennie, voire au-delà : les descriptions vs. les théories, les approches syntaxiques vs. les approches lexicalistes.

et mixtes. Il fait remarquer (Roche 2011, p. 48-49) que si les premiers systèmes d'extraction terminologique étaient syntaxiques (entre autres, *Termino* (David et Plante 1990), *Lexter* (Bourigault 1993), *Syntex* (Bourigault et C. Fabre 2000)), d'autres étaient clairement statistiques (*Xtract*, (Smadja 1993)). Enfin, certains systèmes sont mixtes (*Acabit*, (Daille 1994), *Exit*, (Roche 2004).

Depuis mes premières recherches, j'ai prôné la *description* — aussi exhaustive que possible — de propriétés (interrogatives, verbales, etc.) en essayant de faire en sorte qu'elles soient suffisamment abstraites pour que n'importe quel formalisme puisse les incorporer ultérieurement. Je demeurais hors cadre spécifique. Lors de mon année post-doctorale montréalaise, j'ai effectivement intégré une équipe utilisant une approche syntaxique *positionnelle/géométrique* précise, mais on remarquera que lors de mes implémentations informatiques (Scatlex en est un exemple), je suis restée indépendante vis-à-vis de cette théorie, pour la classification lexicale. Sans doute pour deux raisons : 1) l'implémentation informatique exige de toute façon des ajustements (Léon 2010), et il est extrêmement difficile de rester absolument fidèle à la théorie linguistique ; 2) je voulais maintenir la possibilité de m'ouvrir à d'autres horizons.

Ma position actuelle tendrait vers une *description empirique* nécessaire sans m'obliger à l'effectuer dans un cadre *théorique spécifique*. En cela, je rejoins la position, déjà ancienne, de Lakoff in (Parret 1974) :

> I should say that I do not think that theory construction and verification is the only or even the most important mode of doing linguistics. Theorizing is more glamourous these days than doing careful descriptive work. I think that is unfortunate. Linguistic description is still an art, and is not likely to become a science for a long time. Unfortunately it is an art that has begun to die just at the time when it should be flourishing most. [...] It has become clear in the past decade that no linguistic theory is anywhere near adequate to deal with most facts. [...] Any description of a language that adheres strictly to some formal theory will not describe most of what is in the language. (Lakoff in (Parret 1974, p. 152-153).

Dans un cadre lexicaliste, (Laporte 2015) évoque l'importance de l'observation empirique :

> Gross stressed formal procedures of empirical observation and systematic lexical studies. (Laporte 2015).

suivie d'une étape de formalisation descriptive lexicale :

> In a lexicon-grammar, every distinction with any reproducible property is formalized. (Laporte 2015).

Qu'en est-il des approches syntaxiques *vs*. lexicalistes ? S'opposent-elles obligatoirement ? Laporte évoque le « mépris » des linguistes syntacticiens envers ceux s'inscrivant dans la lignée de Gross et des lexiques-grammaires :

> The widespread impression among linguists that syntactical studies are somehow more scientific than lexical studies has deterred them from studying the lexicon. Syntax is where the theories are, and where would we be without the theories ? « Picking up shells on the beach, » some linguists scoff. That is where we would be. (Laporte 2015).

Or, précise-t-il, la rigueur scientifique exigée pour établir des lexiques-grammaires n'est pas donnée à tous [30].

Par ailleurs, des approches syntaxiques, permettant le repérage et l'extraction de données qui correspondent à des catégories syntaxiques (par exemple, *Termino* et les unités nominales polylexicales), et non uniquement à des positionnements linéaires (par exemple une approche par patrons de fouille), sont également valables.

Can we meet half way ? L'important, me semble-t-il, est de se rencontrer à mi-chemin. Déjà, dans (Panckhurst 1994e), je me posais la question du rapprochement entre mon outil informatisé Scatlex (basé sur une approche positionnelle, même si l'outil implémenté correspondait à une structure de classification indépendante vis-à-vis d'un formalisme grammatical) et l'informatisation des tables du LADL, des lexiques-grammaires.

Au final, il me semble important de retenir la notion de *souplesse* : toutes les approches sont possibles et respectables (à condition de les connaître *a minima* et d'éviter de refaire la roue), et l'on se doit d'éviter un jugement quelconque. Le tout est de faire en sorte que les chercheurs investis dans différents domaines du champ vaste que sont les sciences du langage, se rencontrent, se confrontent à un moment donné, mais se sentent libres d'investir les activités de recherche qui leur plaisent. En cela, je rejoins les préconisations de Lakoff qui me semblent toujours d'actualité, 40 ans plus tard :

30. Précision apportée lors d'une discussion avec Éric Laporte, le 29 juin 2016.

> It is important to recognize that there is no one particular « way » to do linguistics. What there are, are variant strategies, some more productive at present than others. (Lakoff in (Parret 1974, p. 153).

Pour ma part, la *description empirique* et le maintien *hors cadre* continuent à me caractériser.

Le volet 1 était surtout investi d'une recherche fondamentale puis applicative via l'élaboration de mes propres outils implémentés. Le volet 2 sera consacré à l'utilisation d'outils par autrui et ce que cela apporte dans un enlacement entre recherche appliquée et sa dimension réflexive.

3.3.1.7 Encadrement spécifique : volet 1

Tableau 3.3 – Encadrement de la recherche : volet 1*

2 Directions de mémoires	
Maîtrise en Sciences du Langage, mention Industries de la langue : 1) « La glose et une application informatique du phénomène. Création d'une rubrique consacrée à la glose à l'aide d'un système de publication par Internet » (C. Porta). 2) « Enquête, logiciel, tutoriel : conception, évaluation, modification » (S. Lafaye).	2001-2002
3 Co-directions de mémoires	
DEA : « L'accès à l'information sur Internet » (F. Pascual, co-directeur : P. Siblot). (Volets 1 et 2)	1998-1999
DEA : « Introduction à une étude des difficultés du français et de leur traitement par ordinateur », (M. da Conceição, co-directeur : A. Coïaniz). (Volets 1 et 2)	1996-1997
DEA : « La préposition "à" au sein des constructions locatives en français » (S. Lescure, co-directeur : J. Bres).	1994-1995
7 Participations à jurys de soutenance	
M1 Recherche : « Outils pour l'analyse automatique du discours » (S. Riou, Département de Sciences du Langage, directeur : P. Siblot). (Volets 1 et 3)	2003-2004
Maîtrise en Sciences du langage, mention Industries de la langue : « La traduction automatique : quelques pistes de réflexion à partir d'une première expérience pratique sur le système de T.A. du LIRMM » (V. Vedel, direction : A. Mela).	2001-2002
Maîtrise en Sciences du langage, mention Industries de la langue : « Repérage automatique des sources d'énonciation » (F. Ruas, direction : A. Mela).	2000-2001
DEA : « La phrase clivée c'est...qui/que et ses équivalents en polonais » (A. Nowakowska, directeur : J. Bres). Maîtrise : « Phonétique Expérimentale : Traitement et analyse numérique de la parole. Essai d'une description acoustique des voyelles nasales du français assisté du logiciel Signalyze 3.12. » (C. Pujol, direction : M. Lanvin).	1997-1998
D.E.S.S : « Terminologie du traitement automatique du langage naturel » (Mourad Amine, Ingéniérie multilingue, CRIM, INALCO, direction : M. Slodzian). DEA : « Quand la langue devient un produit industriel : le projet Translearn » (K. Wagner, Département de Grec Moderne, université Paul-Valéry Montpellier 3, direction : M.-P. Masson).	1994-1995

*Dans ce tableau ne sont inscrits que les encadrements en rapport avec mes sujets de recherche. J'ai également encadré d'autres mémoires « généralistes », qui ne sont pas indiqués ici. *Cf.* le tableau 3.1 pour ce détail.

Récapitulatif des publications sélectionnées du volet 1

Pendant mes premières années de recherche j'ai surtout publié en auteur unique.
Quatre publications sont liées au **doctorat** : 2 actes de colloque, *CIL*, Québec, (Panckhurst 1992a) et *ILN*, Nantes, (Panckhurst 1993a), 1 résumé publié dans *TAL*, (Panckhurst 1992b), et 1 article, *CHUM*, (Panckhurst 1994a). Les points abordés concernent plus particulièrement la consultation de **bases de données**).
Six publications concernent **l'outil** en sémantique lexicale que j'ai implémenté (en FX/LISP) pour la classification verbale en français, **Scatlex** (*cf.* 1 coordination de revue (*Cahiers de Praxématique*, (Panckhurst 1994b)), 2 articles (*ICO* Québec, (Panckhurst 1993b) *Cahiers de Praxématique*, (Panckhurst 1994e)), 3 actes de colloque (*TALN*, Marseille, (Panckhurst 1995b) *LGC*, Montréal, (Panckhurst 1995c), *ACH/ALLC*, Santa Barbara, (Panckhurst 1995a)).
Une sélection de **trois publications** a été effectuée sur les recherches que j'ai menées à propos du fonctionnement des **unités verbales polylexicales** (UVPL). Ces recherches s'inscrivaient à la fois dans le cadre du séminaire de recherche du laboratoire Praxiling « Discours, textualité et production de sens » : 1 chapitre dans l'ouvrage collectif (*De l'actualisation*, paru aux éditions CNRS, (Panckhurst 1998b)), 1 acte de colloque (*ILN*, Nantes, (Panckhurst 1996c)), 1 chapitre dans l'ouvrage collectif à propos du repérage automatique des UVPL (*La locution et la périphrase du lexique à la grammaire*, (Panckhurst 2001b)), puis au sein du projet 1.1.4. « Le verbe et son environnement », axe 1 *: Sujet, praxis et production de sens*, Programme : *Lexique et discours* du laboratoire Praxiling. Le projet portait sur l'**informatisation** ultérieure des **données linguistiques formalisées**, à la fois sur le plan de la recherche fondamentale et de la recherche appliquée pour l'analyse syntaxique et en sémantique du discours.
Une autre publication (1 chapitre dans l'ouvrage collectif *Le mot et sa glose*, (Panckhurst 2003b)) porte sur un travail d'**extraction automatisée** de la **glose**, à l'aide de Perl.
Enfin, **une publication** et **un addendem** concernent la mise en place des enseignements de TAL dans un contexte montpelliérain, (Panckhurst 1996a,b).

3.3.1.8 Sélection des publications : volet 1

Remarque. — Dans le Volume II, je scinde les publications en quatre sections, pour le volet 1 : 1) Publications liées au doctorat ; 2) Outil informatisé, sémantique lexicale, classification verbale ; 3) Unités verbales polylexicales (UVPL), actualisation, glose ; 4) Publications pédagogiques. Le lecteur pourra s'y reporter pour ce détail.

PANCKHURST, Rachel (1992a). « Comment allier les besoins du linguiste et l'utilisation intelligente de bases de données ? » In : *Actes du XVe Congrès international des Linguistes (CIL)*. Québec : Les Presses de l'Université Laval, Sainte-Foy, p. 301–304.

– (1992b). « Description linguistique et implémentation en FX des structures interrogatives (directes) du français. Résumé de thèse. » In : *Traitement automatique des langues (T.A.L.), Spécial trentenaire*. 33.1-2, p. 250–252.
– (1993a). « Analyseurs et bases de données pour des besoins spécifiques ». In : *Actes du colloque Informatique et langue naturelle, ILN '93*. Nantes, p. 207–222.
– (1993b). « Scatlex : une aide informatisée pour la construction d'entrées lexicales verbales ». In : *Revue de la liaison de la recherche en informatique cognitive des organisations (ICO)* 5.3, p. 61–67.
– (1994a). « A Database for linguists : intelligent querying and increase of data. » In : *Computers and the Humanities* 28.1, p. 39–52.
– éd. (1994b). *Cahiers de Praxématique, PULM* 22. URL : http://praxematique.revues.org/1887.
– (1994e). « Une structure de classification verbale basée sur des contrastes ». In : *Cahiers de Praxématique* 22, p. 105–134. URL : http://praxematique.revues.org/2275.
– (1995a). « Behind the scenes : Building a tool for Verb Classification in French. » In : *Actes du colloque Association for Computers and the Humanities/Association for Literary and Linguistic Computing, ACH/ALLC*. Santa Barbara, p. 89–92.
– (1995b). « Décrire le système verbal indépendamment d'un cadre grammatical. » In : *Actes du colloque Le traitement automatique du langage naturel, TALN*. Marseille, p. 172–180.

Panckhurst, Rachel (1995c). « Poly…quelque chose et classification verbale. » In : *Actes du colloque Lexiques-Grammaires comparés et traitements automatiques, Université du Québec à Montréal (UQAM)*. Montréal, p. 199–206.
– (1996a). « Formation en linguistique-informatique : une expérience montpelliéraine ». In : *Traitement automatique des langues (T.A.L.), Enseignement du TAL*. 37.1, p. 51–64.
– (1996b). « Linguistique-informatique : la crise (Addendum) ». In : *Traitement automatique des langues (T.A.L.), Grammaire et théorie de la preuve* 37.2, p. 176–177.
– (1996c). « Quelques problèmes posés pour l'analyse automatique des unités verbales en français. » In : *Informatique et langue naturelle (ILN)*. Nantes, p. 465–476.
– (1998b). « Des unités verbales polylexicales ». In : *De l'actualisation*. Sous la dir. de Jeanne-Marie Barbéris, Jacques Bres et Paul Siblot. Paris : CNRS-Éditions, p. 161–178.
– (2001b). « Les unités verbales polylexicales : problèmes de repérage en traitement automatique ». In : *La locution et la périphrase du lexique à la grammaire*. Sous la dir. de Francis Tollis. Paris : L'Harmattan, p. 55–63.
– (2003b). « La glose, le document électronique et l'extraction automatisée ». In : *Le mot et sa glose*. Sous la dir. d'Agnès Steuckardt et Aïno Niklas-Salminen. T. Langues et langage. 9. Publications de l'université de Provence., p. 271–292.

3.3.2 Volet 2 : Formation, évaluation, réseaux pédagogiques (1996-2012)

Le deuxième volet de mon triptyque s'ancre dans la formation et l'évaluation, toujours du double point de vue théorique et appliqué, avec un intérêt, dans la deuxième partie de cette période, à partir de 2006, pour les réseaux pédagogiques, que je définis plus bas.

Si mes recherches du volet 1 avaient porté essentiellement jusqu'alors sur ma propre élaboration d'outils/prototypes en TAL (RISIF, CBD, DAL, Scatlex, UVPL), désormais, je souhaitais déplacer le curseur afin de mettre mes compétences à la disposition des autres. Mes motivations tenaient — et tiennent toujours — dans ma volonté d'être au service du public, de répondre à leurs besoins et à leurs attentes, précisément au sein d'un établissement de « service public ». Je tenais à aider autrui — non seulement le public étudiant, bien entendu, mais également les enseignants-chercheurs et les personnels administratifs et techniques — à s'approprier et apprivoiser les outils informatiques nécessaires dans leur propre travail (Panckhurst et Pérez 2000), et leur donner les clefs pour un apprentissage ultérieur autonome. Ce deuxième volet est également caractérisé par le questionnement à propos d'une pédagogie renouvelée pour l'autoformation et l'autoévaluation (Vincent-Durroux et Panckhurst 2002), par une réflexion sur l'évaluation de logiciels dans un cadre européen (Panckhurst et al. 2004a), (David et al. 2005), (Amar et al. 2008) et sur la formation ouverte et à distance (FOAD, eLearning), et, notamment, sur les réseaux d'échanges pédagogiques dans ce cadre (Panckhurst et Marsh 2006, 2007, 2008a,b, 2009, 2011a,b), (Panckhurst 2011, 2012).

Les volets du triptyque se chevauchent en terme de périodicité. Si le premier volet (prototypes et outils) ne se termine réellement qu'en 2003, le deuxième volet démarre dès 1996 [31]. À l'évidence, certains travaux sont menés en parallèle, même si, pour être honnête, la plupart des publications de la deuxième

31. Cela est également le cas pour le troisième volet. Je reviendrai sur ce point ultérieurement (*cf.* § 3.3.3). En réalité le deuxième volet débute en 1994, avec le premier concours européen d'évaluation de logiciels (EASA, *European Academic Software Award*). Cependant, comme 1996 marque un tournant dans les orientations de ma recherche à plusieurs niveaux, j'ai préféré effectuer le découpage ainsi.

période du volet 1, à partir de 1996 (à l'exception du travail sur la glose, en 2003), correspondent à des recherches déjà amorcées quelques années auparavant.

3.3.2.1 Formation pour tous les personnels de l'université et pour les doctorants (1996-2003)

En 1996, j'étais à l'université Paul-Valéry Montpellier 3 depuis 4 ans. À l'issue de la première année de ma prise de fonction, en juin 1993, j'avais proposé aux enseignants-chercheurs du département des Sciences du Langage, une modeste formation de formateurs à propos des outils bureautique de base. C'était un avant-goût de la mise en place nationale des formations pour l'obtention du *Certificat Informatique et Internet* (C2i, devenu « compétences numériques » : https://c2i.enseignementsup-recherche.gouv.fr/), bien que celles-ci s'adressent exclusivement aux étudiants. Mais les enseignants-chercheurs ressentaient déjà une certaine appréhension, voire une véritable « crainte », à l'idée d'être « dépassés » par les étudiants en matière d'acquis technologiques. La formation que j'ai assurée pour mes nouveaux collègues en 1993 m'a indéniablement ouvert la voie pour mes prises de fonctions quelques années plus tard :

- responsable de la formation pour tous les personnels/acteurs de l'université (service informatique, CRIT-Formation, 1996-2003)
- directrice du service commun METICE (*Multimédia, Enseignement, Technologies de l'Information et de la Communication Éducatives*, 1999-2001), nommée par la présidente de l'université.

Comme indiqué précédemment (*cf.* § 2.3), j'ai assuré la coordination et la co-formation pour un public tripartite au sein de l'université Paul-Valéry Montpellier 3 (enseignants-chercheurs, personnels techniques et administratifs, doctorants) entre 1996 et 2003. À cette époque — de l'enseignant-chercheur « tout puissant » présumé « détenteur du Savoir » — les formations étaient uniquement prévues pour les personnels administratifs et techniques dans les budgets ministériels. Tout était à inventer pour les enseignants-chercheurs. Rapidement, j'ai commencé à réfléchir aux mutations — y compris en recherche — que cela allait engendrer.

3.3.2.2 Formation et recherche

Le livre que j'ai co-écrit (Panckhurst et Pérez 2000) décrit les activités de formation menées pendant l'année universitaire 1998-1999 dans le cadre de CRIT-Formation (utilisation de base d'un Macintosh ou d'un P.C. ; courrier électronique ; accès à Internet ; formats de fichiers ; outils de navigation, etc.). Mais il était important de réfléchir à la façon dont je pouvais rattacher la formation (continue), puis plus tard l'évaluation, à mes activités de recherche. Le lien s'est tissé tout naturellement. Au moment de co-rédiger cet ouvrage, je m'intéressais depuis quelques années, en recherche, à toutes les formes de communication via des ordinateurs interposés. Je menais notamment une réflexion qui avait pour but d'étudier le statut du type de discours *médié* véhiculé par le courrier électronique échangé dans un contexte universitaire (*cf*. Volet 3, § 3.3.3). Outre des analyses de corpus de messages électroniques, je m'intéressais également aux comportements des sujets devant l'ordinateur (qui ont été filmés pendant le déroulement d'une expérience menée sur l'utilisation d'un correcteur grammatical [32] par des étudiants étrangers (Charnet et Panckhurst 1998). J'exposais le lien entre mes objets de réflexion en recherche et les « travaux pratiques » en formation de la manière suivante :

> Quel lien existe-t-il entre les sujets de recherche mentionnés et le livre ici-même, qui se veut une introduction aux nouvelles technologies de l'information et de la communication dans un contexte universitaire ? Il nous semble que pour bien comprendre le contexte global (le développement d'une forme nouvelle de discours, le comportement de l'individu devant la machine, etc.) — donc, pour avancer vis-à-vis de notre sujet de recherche actuel en linguistique et informatique —, il est tout à fait intéressant, voire nécessaire, d'assister à l'appropriation de ces outils par le néophyte. À l'origine de ce livre se situent donc les activités de formation menées dans un cadre universitaire pendant l'année 1998-1999, qui nous ont permis, précisément, de mieux comprendre les difficultés, les besoins et les attentes des utilisateurs des technologies de notre époque. Et, à l'origine de ces formations, s'ancre la réflexion que nous avons menée concernant le manque de formation prévue, notamment, pour le public d'enseignants-chercheurs, le tout

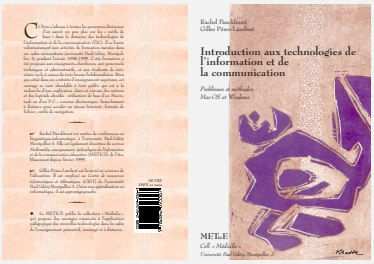
32. Le correcteur grammatical *Le Correcteur 101* a été commercialisé par *Machina Sapiens*, à Montréal, en 1992.

> ayant été motivé par nos préoccupations en recherche.
>
> (Panckhurst et Pérez 2000, p. 2).

À l'évidence, je me posais néanmoins la question du positionnement intellectuel pour ce type d'ouvrage :

> S'agit-il alors d'un livre de recherche ou de pédagogie ? La question n'a pas une importance ou une pertinence véritable. Dans les années à venir, il est certain que la recherche réellement appliquée et pratique prendra le pas sur une recherche qui demeure uniquement fondamentale. Les différents domaines d'enseignement et de formation (présentiel, aménagé, à distance) actuels exigent une réflexion et une recherche pluridisciplinaires urgentes (pour ne nommer que quelques disciplines : sciences du langage, informatique, information et communication, sciences de l'éducation, psychologie, etc.). La linguistique-informatique, la recherche documentaire, le traitement automatique des langues, l'intelligence artificielle, les sciences cognitives, seront (ou sont déjà) au cœur des avancées liées aux pratiques d'utilisation sur le réseau Internet. Par exemple, dans quelques années (quelques systèmes sont d'ores et déjà à l'essai sur Internet), tel étudiant travaillant à distance, pourra consulter le fonds documentaire des différents cours qu'il suit et qu'il a téléchargés à partir d'un serveur distant, par une recherche de concepts, réellement sémantique, et non pas par une recherche à l'aide de mots-clefs et d'opérateurs booléens. L'indexation et la recherche d'information qui s'en suit tiendront le haut du pavé de la recherche pluridisciplinaire dans les années du début du nouvel millénaire ; d'aucuns le savent depuis longtemps. Enfin, pour répondre à la question posée en début de paragraphe, c'est grâce aux réflexions menées en recherche, que des livres du type que nous proposons ici peuvent surgir. Recherche, didactique, pédagogie ne s'opposent donc pas ; l'un ne peut être amoindri comparé aux autres, le tout s'imbrique nécessairement.
>
> (Panckhurst et Pérez 2000, p. 2-3).

(Panckhurst et Pérez 2000) *Introduction aux technologies de l'information et de la communication. Problèmes et méthodes : MacOS et Windows.*, coll. MédiaTic n° 1, université Paul-Valéry Montpellier 3.



Un compte rendu de notre ouvrage figure en ligne dans la revue *Apprentissage des Langues et Systèmes d'Information et de Communication* https://alsic.revues.org/1869 (Pothier, vol. 3, numéro 2, décembre 2000).

Précision : distinction volets 2 et 3

Mais, une question demeure : pourquoi avoir prévu deux volets distincts (volet 2 : formation/évaluation ; volet 3 : communication/discours), alors que je mentionne les influences de mes recherches concernant le volet 3 sur l'époque concernant la formation, et ce dès 1996 ? Je n'aurais pas pu le comprendre à l'époque, mais maintenant, avec le recul d'une décennie, je constate qu'il y a bien une cohérence pour moi. Les recherches que j'ai menées à partir de 1996 sur la *communication médiée par ordinateur* (Herring 1996, Panckhurst 1997a, volet 3) ont été véritablement impulsées par des comportements, voire des mutations au niveau des pratiques scripturales, à la fois du côté estudiantin et chez mes collègues. J'ai pu les observer et les comprendre grâce aux formations à la fois initiale et continue, et par le biais des échanges communicationnels par ordinateur. Je reviendrai sur cela quand je présenterai le volet 3 (§ 3.3.3). Mais, insisteriez-vous, si ces recherches ont été initiées suite aux formations, pourquoi ne les avoir pas tout simplement intégrées directement dans le volet 2 ? Ma réponse est la suivante : la formation initiale, élargie au cadre de la formation continue, a également propulsé une réflexion sur l'évaluation — quand on forme, on évalue aussi, bien évidemment — puis, vis-à-vis des pratiques émergentes en matière de pédagogie dans le contexte de ce qu'on appelait à l'époque, les

nouvelles technologies de l'information et de la communication éducatives (NTIC(E), d'abord, TIC(E) ensuite) dans l'ère du tout-numérique à venir. J'ai donc préféré scinder les volets en deux. Cela apparaîtra plus clairement, je l'espère, à l'issue de la lecture des 2e et 3e volets, menés temporellement de manière quasi synchrone, jusqu'en 2012, date de la fin officielle de mes recherches du volet 2.

3.3.2.3 Publications pédagogiques (1998-2001)

Outre l'article dans *ALSIC* (Charnet et Panckhurst 1998), qui comme son titre l'indique s'intéresse à une utilisation pédagogique d'un correcteur grammatical (en concluant à une utilisation maîtrisée d'un logiciel de cette nature pour des tâches morpho-lexico-syntaxique, mais en excluant une stimulation de la production sémantico-pragmatique), à d'autres moments de ma carrière, j'ai souhaité fournir des écrits qui cherchaient à répondre aux questions suivantes : qu'est-ce que le TAL ? À quoi sert le TAL ? Peut-on implémenter les phénomènes sémantiques, voire pragmatiques, etc. ? Dans *Hommages à Xavier Mignot*, le chapitre que j'ai rédigé est intitulé « Sens et Informatique ». J'avais décidé de partir d'un paragraphe de l'ouvrage de Baylon et Mignot (Baylon et Mignot 1995) dans lequel les auteurs justifiaient une non-inclusion de chapitre sur ce thème :

> On s'étonnera peut-être de ne pas trouver dans l'ouvrage un chapitre sur « sens et informatique ». [...] Il nous est apparu qu'on ne disposait pas encore de données réellement instructives justifiant la présence d'un chapitre dans un ouvrage d'initiation.
>
> (Baylon et Mignot 1995, p. 5).

Même si j'aboutis à la conclusion (semblable) suivante (*cf.* ci-dessous) :

> Bien que les recherches menées en direction d'une implémentation éventuelle de phénomènes sémantiques (essentiellement en utilisant des formalismes logiques) et pragmatiques existent depuis fort longtemps, il est encore trop tôt pour parler d'un véritable « sens » machinal. Pour y parvenir au moins partiellement, le « traitement sémantique des textes doit s'ouvrir aux situations. [Il nous semble effectivement que seule] la prise en compte de facteurs sociaux, ethnographiques et anthropologiques, conditionne les progrès ergonomiques des traitements automatiques » (Rastier *et al.* 1994, p. 201). [...] La mise en machine de phénomènes

> linguistiques et extra-linguistiques passe nécessairement par une réflexion menée loin de l'ordinateur, et peut-être ne parviendrons nous jamais à approcher un tant soit peu un traitement « artificiel » de « l'intelligence » :
>
> Quelque ressemblance superficielle que l'on ait observée entre le raisonnement de l'homme et celui de la machine, l'intervalle qui les sépare demeure encore immense, et nul ne sait s'il pourra jamais être réduit. (Ganascia 1996, p. 44).
>
> (Panckhurst 1997b, p. 128-129).

la réflexion induite m'a permis de proposer un écrit qui m'a servi en situation pédagogique à maintes reprises, dans la mesure où j'expose des points classiques du traitement automatique [33] (chaîne de traitement : prétraitement, analyses morpho-lexicale et syntaxique, découpage, étiquetage, traitement terminologique, etc.), et que j'évoque des considérations sémantico-pragmatiques de l'époque (mes recherches sur les UVPL (*cf.* volet 1, § 3.3.1), entre autres).

Dans le même sens, j'ai contribué à l'ouvrage coordonné par Détrie, Siblot, Vérine (2001), sur l'analyse du discours d'un point de vue praxématique. N'étant pas moi-même praxématicienne, j'ai néanmoins souhaité porter ma (minuscule) pierre à l'édifice de cet ouvrage dictionnairique collectif, afin d'y apposer mon point de vue sur le TAL et sur le traitement informatisé de corpus, dans des articles brefs et utiles en circonstances pédagogiques. En 2017, l'ouvrage *Termes et Concepts pour l'analyse du discours* a été réédité et augmenté (Détrie et al. 2017). Cela a fourni l'occasion d'investir de nouveaux terrains, en matière de courts articles que j'ai proposés et de réactualiser les articles précédents (Panckhurst 2001, 2017) [34].

3.3.2.4 Mutations. Vers une pédagogie renouvelée (1999-2002)

De mon propre positionnement d'enseignant-chercheur en linguistique-informatique, je me posais constamment des questions sur la mutation des supports d'enseignement/recherche. Cela apparaît clairement dans la préface de (Panckhurst

33. Comme indiqué précédemment, pour une histoire très fournie sur la discipline, j'invite à lire l'excellent ouvrage de Jacqueline Léon (Léon 2015), *Histoire de l'automatisation des Sciences du langage*.

34. Si j'ai gardé le même intitulé pour l'article *TAL*, j'ai modifié *traitement informatisé de corpus* en *linguistique(s) de corpus*, pour l'édition de 2017.

et Pérez 2000) rédigée par Daniel Savey, enseignant-chercheur en linguistique anglaise, à l'époque directeur du CRIT (service informatique) de notre université :

> Les applications de traitement de l'information et du texte [...] sont le pain blanc d'une université comme la nôtre. Et aussi se prépare la grande mutation des supports d'enseignement, que nous devons apprendre à maîtriser en universitaires respectueux des valeurs intellectuelles et culturelles, un peu perdus dans cet océan d'outils d'incitation à la Consommation qu'est devenu l'Internet — initialement conçu par et pour les chercheurs. Gardons la foi : des équipes comme celle qui propose cet ouvrage sont de taille à résister aux errements et à tirer profit des techniques pour maintenir la culture et la science.
>
> (Préface, D. Savey, in (Panckhurst et Pérez 2000)).

Lorsque j'ai été nommée directrice du service commun METICE, nous étions au tournant numérique, à l'horizon 2000. Il s'agissait de faire le passage d'un service d'enseignement à distance qui fonctionnait depuis 25 ans avec des envois papier à une gestion en ligne via des plateformes de cours (*cf.* (Panckhurst 2001a) pour une évocation plus détaillée du processus). Mais au-delà des aspects techniques, organisationnels, logistiques énormes, c'était la question des *mutations* qui m'interpellait : mutations des supports, mutations des savoirs, mutations des pratiques scripturales [35]. Le METICE était divisé en trois branches [36], et l'une d'elles, *l'Espace Multimédia* — lieu d'auto-formation pour étudiants anglicistes, au sein de l'université — était conjointement dirigée par deux collègues, Laurence Vincent-Durroux et Cécile Poussard. Toutes deux sont enseignants-chercheurs et leurs recherches portaient, entre autres, sur les dispositifs autonomisants pour la formation en langues (*Cf.* (Vincent-Durroux et Poussard 2014) pour une publication récente à propos de *Macao*, le logiciel pédagogique qu'elles ont conçu).

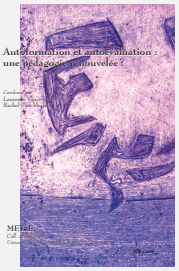
En 2000, Laurence et d'autres collègues de l'université Paul-Valéry Montpellier 3 ont organisé conjointement avec le *Pôle universitaire européen de Montpellier*

35. Les mutations dans le cadre des pratiques scripturales seront abordées de manière privilégiée dans le cadre du volet 3 (*cf.* § 3.3.3).

36. Les trois branches du METICE étaient : le SUAV (service universitaire audio-visuel), le SÉAM (service d'enseignement aménagé et à distance), l'EMM (Espace Multimédia). Ce service a été entièrement restructuré, puis inséré directement au sein de la DSI (Direction des services informatiques) récemment.

et du Languedoc-Roussillon (devenu ensuite le PRES-UMSF, *Pôle d'enseignement supérieur et de recherche, université Montpellier Sud de France*, puis, plus récemment la COMUE), et avec le soutien de l'ISIM (Institut des sciences de l'ingénieur de Montpellier) et du service METICE, que je dirigeais, une journée d'étude intitulée « Une pédagogie renouvelée par l'autoformation et l'autoévaluation ». Il s'agissait de faire le point sur « l'autonomie et l'autoévaluation dans les dispositifs d'apprentissage mis en place dans l'enseignement supérieur grâce au développement des environnements multimédias » (*cf.* la quatrième de couverture, (Vincent-Durroux et Panckhurst 2002)). Les nouveaux rôles des enseignants et des tuteurs dans ces contextes étaient également pris en compte. En 2001, Laurence et moi avons co-coordonné la publication d'une sélection des communications remaniées, afin de faire bénéficier la communauté scientifique des réflexions de spécialistes du domaine. Tout comme (Panckhurst et Pérez 2000), nous avons publié l'ouvrage dans la collection *MédiaTic* [37].

(Vincent-Durroux et Panckhurst 2002, coord.), *Autoformation et autoévaluation : une pédagogie renouvelée ?*, coll. MédiaTic n° 2, université Paul-Valéry Montpellier 3.



Maggy Pézeril, alors conservateur général et chargée de mission TIC au *Pôle*, également l'une des organisatrices de la journée, contextualisait les préoccupations des chercheurs dans l'avant-propos de notre ouvrage :

> À l'ère des « campus virtuels » ou « numériques » il était important de replacer l'étudiant au centre de la réflexion sur les dispositifs d'enseignement médiatisé.

37. Le METICE publiait la collection *MédiaTic* qui proposait des ouvrages consacrés à l'application pédagogique des nouvelles technologies dans le cadre de l'enseignement présentiel, aménagé et à distance.

> Le concept d'« e-learning » rend-il bien compte des véritables enjeux de la formation ? N'est-il pas souvent le faire-valoir de services de distribution de cours qui fleurissent sur le Web, cachant sous l'abondance du catalogue des cours et sous la technicité des outils et plates-formes utilisés, un manque de garantie quant au véritable service d'enseignement offert ?
>
> (Pézeril, in (Vincent-Durroux et Panckhurst 2002, p. 11)).

Avec les nouvelles possibilités d'offrir des cours via des plateformes d'enseignement à distance, les responsables ministériels songeaient à une économie budgétaire drastique. Nous, chercheurs dans ces disciplines, avons expliqué que si les innovations numériques permettaient effectivement de proposer de nouveaux outils de transmission et d'acquisition du savoir — que ce soit au niveau de supports supplémentaires pour les étudiants en présentiel ou d'outils de mise à disposition de cours pour des étudiants en formation ouverte et à distance (FOAD/eLearning) — il fallait effectuer un travail approfondi théorique sur les nouvelles pratiques pédagogiques. De plus est, j'étais consciente que la mise en place technique et pédagogique que nous avons impulsée en l'an 2000, devait être accompagnée :

> From the METICE's perspective, we knew that trying to change attitudes in order to adapt to modern trends in education was an enormous challenge and that staff would need to be « coached », guided and supported. (Panckhurst 2001a)

Par ailleurs, à mon sens, il n'était pas du tout sûr que ces formations soient moins onéreuses par rapport aux formations classiques en présentiel.

Cette journée d'étude et l'ouvrage résultant apportent cette réflexion scientifique, d'un double point de vue de repères théoriques (liens entre autoformation, individualisation et autonomie ; représentations des acteurs apprenants/enseignants en matière d'isolement et de solitude ; mutation des rôles : enseignants devenant tuteurs et médiateurs ; guidage de l'apprenant afin de privilégier son autonomie ; évaluation et réorientation voire amélioration du parcours de l'apprenant) et de mises en œuvre de dispositifs/outils intégrant de l'autoformation ou de l'autoévaluation (formation en langues : éclatement de l'intervention enseignante et prise de distance des apprenants face à leur formation ; baladeurs/podcasts pour la compréhension en langue étrangère orale, visant une écoute autonome grâce

à une approche cognitive de comparaison ; présentation des plateformes européennes de l'époque (Ariadne, etc.), dans un souci de guidage pédagogique et de fédération de ressources pédagogiques multimédias de partenaires européens) (Vincent-Durroux et Panckhurst 2002, p. 14-15).

J'avais une double casquette de directrice d'un service commun ayant impulsé et accompagné la transition des cours papier aux cours sous format numérique (plateforme universitaire initialement choisie : WebCT, remplacée par Moodle quelques années plus tard (Panckhurst 2001a), et d'enseignant-chercheur très impliqué, d'une part, dans le choix d'outils appropriés (généralistes et/ou plus spécialisés en TAL) pour les étudiants, et, d'autre part, ayant une longue expérience en enseignement présentiel/à distance et hybride (mixte entre présentiel et eLearning). Tout cela a confirmé mon intuition du départ voici plus de 15 ans : la formation ouverte et à distance n'est pas moins coûteuse que la formation en présentiel. Au contraire, si elle est organisée comme il se doit, elle exige une très grande attention de la part des tuteurs/médiateurs, dans le suivi plus personnalisé des étudiants, et demande de jongler entre plusieurs types d'outils de communication (courriels, forums, chats, et, plus récemment, SMS, utilisation d'applications comme *WhatsApp*, etc.).

J'aborderai les recherches que j'ai menées sur des façons innovatrices d'aborder la formation des apprenants dans un instant (*cf.* 3.3.2.6). Le virage du numérique était pris avec la mise sur plateforme des cours pour les étudiants en FOAD/eLearning.

En attendant, en 2001, j'avais l'impression d'avoir fait ce que j'avais à faire, en apportant une certaine expertise dans la transmission papier vers numérique pendant deux ans et demi à la tête du METICE. J'avais 40 ans (*Life begins at 40*, as they say), et il était temps que je m'investisse dans d'autres projets.

3.3.2.5 EASA (European Academic Software Award) ; évaluation (1994-2004)

Mais revenons un temps en arrière. Parallèlement à la recherche que j'avais menée en sémantique lexicale théorique et appliquée (volet 1, § 3.3.1), j'ai fourni un travail de recherche théorique et appliquée concernant l'évaluation de logiciels en nouvelles technologies. Le cadre retenu était le concours EASA (*European Academic Software Award* - Concours européen d'évaluation de logiciels pour

l'enseignement supérieur et la recherche). Ce concours européen permettait aux universitaires (enseignants-chercheurs, chercheurs, ingénieurs, étudiants) qui ont développé des logiciels, en collaboration ou non avec des entreprises, de les soumettre pour évaluation à un comité d'experts européens. J'ai été contactée pour participer en tant que membre du jury évaluateur et expert « coordonnateur de discipline » dans le cadre de ce concours pendant une décennie (1994-2004), (*cf.* § 3.1.2). La France était officiellement (et financièrement) impliquée entre 1999 et 2006, puisque notre université avait accepté d'assurer la coordination française, sous ma responsabilité ; le Ministère nous avait, par ailleurs, accordé une subvention (d'un montant de 70 000 F en 1998) pour mener à bien ce projet. À partir du moment où la France participait via une subvention ministérielle, on m'a également sollicitée pour intégrer le comité d'organisation de l'association European Knowledge Media Association (EKMA), qui organisait le concours. J'ai assumé la responsabilité de représentative nationale pour la France (puis membre associé [38]) entre 1999 et 2006.

Les objectifs de l'association à but « non lucratif » EKMA étaient :

(a) stimuler la compréhension, le développement et l'usage des médias à travers l'Europe, tout particulièrement au niveau universitaire ;

(b) propager en Europe l'information concernant la compréhension, le développement et l'usage des médias de la connaissance. (*cf.* (Panckhurst et al. 2004a, Statuts, Annexe B)).

L'association EKMA a organisé le concours EASA pendant une décennie (1994-2004) avant d'être dissoute en 2006. Au niveau européen, les nombreux acteurs, s'échelonnant d'universitaires à de hauts responsables dans la fonction publique et/ou dans des institutions/associations dans leurs propres pays, ont collaboré ensemble sans intervention d'une « agence externe » :

> A striking feature of EASA/EKMA has been the way the member organisations representing each European country have come together entirely of their own accord and have each found the resources to enable EASA to happen. No external

38. À partir de 2003, la France est devenue membre associé, puisque la subvention n'a pas été renouvelée et la France n'a pas pu honorer l'exigence financière de cotisation annuelle pour adhérer à l'association suisse, EKMA (15 000 CHFr 14 000 €).

> agency was required to create EASA and the six EASA competitions [39] have only been possible through a very substantial amount of voluntary effort.
>
> (Jonathan Darby, Preface, in (PANCKHURST et al. 2004a)).

Ma participation tripartite (organisation et coordination d'une discipline pour l'évaluation de logiciels soumis ; membre du jury à la finale ; membre du comité d'organisation du concours, EKMA) a nourri mes travaux de recherche en m'incitant à mener une réflexion approfondie à la fois à propos de l'évaluation (comment évaluer un logiciel ? qu'est-ce qu'on évalue ? peut-on comparer deux logiciels foncièrement différents, bien qu'inscrits dans une même discipline ? etc.) et au niveau de la formation ouverte et à distance (FOAD/eLearning), puisque, les années passant, de plus en plus de logiciels soumis se positionnaient dans cette mouvance.

L'expérience EASA/EKMA a été absolument passionnante, dans la mesure où (entre autres facteurs), c'était l'occasion rêvée de partager des idées, de se confronter à ses pairs sur des sujets de recherche d'actualité. En tant qu'enseignant-chercheur en linguistique-informatique, pouvoir rencontrer et échanger avec des collègues, des étudiants, des responsables ministériels, etc. à propos de l'*évaluation*, de *logiciels* et d'*eLearning* me confortait dans le chemin que j'avais emprunté : après avoir créé des outils/prototypes moi-même (volet 1, § 3.3.1), maintenant j'allais pouvoir m'intéresser à l'évaluation des logiciels des autres — tout en intégrant ce travail au niveau de mes recherches, et également impliquer mes étudiants directement dans la phase d'évaluation des logiciels du concours.

Les finales du concours EASA étaient des lieux d'échanges extrêmement enrichissants. Non seulement plusieurs centaines de personnes se réunissaient pour une finale dans le but d'une évaluation sous forme de compétition menant à des prix, mais tous les acteurs pouvaient échanger à divers niveaux : les auteurs des logiciels pouvaient comparer leurs outils entre eux, les universitaires-évaluateurs étaient dans une atmosphère de colloque [40], pluridisciplinaire, avec

39. À l'issue de la période de sélection et d'évaluation des logiciels, une finale s'est déroulée dans un des pays membres d'EKMA. Les finales se sont déroulées dans les villes suivantes : Heidelburg (Allemagne, 1994), Klagenfurt (Autriche, 1996), Oxford (Angleterre, 1998), Rotterdam (Pays-Bas, 2000), Ronneby (Suède, 2002), Neuchâtel (Suisse, 2004).

40. Les finales EASA les plus récentes étaient souvent jumelées avec des colloques, afin de créer un environnement ouvert et propice aux échanges.

un public élargi au-delà de leurs pairs habituels, incluant un public estudiantin. Les quelques jours de la finale étaient donc toujours intenses mais très stimulants.

À la finale d'EASA 2002, à Ronneby, j'ai fait la rencontre de Shirley Alexander, qui donnait une conférence plénière à *NetLearning* 2002. Actuellement Deputy Vice-Chancellor et Vice-President à University of Technology (UTS), Sydney, Australie, Shirley est (entre autres) une experte en eLearning/humanités numériques et a longtemps dirigé l'*Institute of Interactive Media & Learning* (IML) à UTS. C'est à Ronneby également que j'ai proposé au comité d'organisation EKMA l'idée de rédiger un livre qui ferait une synthèse de nos expériences et de notre expertise pendant une décennie, et qui proposerait des pistes et d'orientation de recherche pour la décennie suivante. Ma proposition a été acceptée avec enthousiasme. Restait la rédaction !

Très intéressée par l'approche et les travaux de Shirley Alexander, j'ai sollicité un congé de recherches en 2003, afin de l'effectuer au sein de son institut. Elle a accepté ma demande. J'ai obtenu mon CRCT illico, à prendre sur le deuxième semestre de l'année universitaire 2003-2004 et j'ai eu la chance d'intégrer l'Institut IML à l'UTS, qu'elle dirigeait. J'y ai passé 5 mois de février à juin 2004 inclus. Mon fils, né quelques années auparavant, en 1999, a pu, de cette manière, s'entraîner à imiter l'accent australien pendant qu'il apprenait à lire en anglais, dans une école publique (*Darlinghurst Public*) à Sydney…

Pendant ce premier semestre 2004, riche en rencontres/séminaires et échanges avec des collègues australiens, j'ai co-coordonné un ouvrage avec Sophie David et Lisa Whistlecroft sur l'évaluation en eLearning, appliquée au concours EASA. J'ai décidé de faire la mise en pages moi-même à l'aide de TEX-LATEX. Comme j'enseignais effectivement depuis l'année 2000 un cours d'initiation à la *publication assistée par ordinateur* (PAO), l'occasion d'utiliser le célèbre système de préparation/composition de documents était une façon de joindre mes activités de recherche appliquée à mes enseignements [41].

41. Bien que je continue actuellement à enseigner la PAO, je n'ai pas créé de cours exclusivement dédié à TEX-LATEX. J'ai fait varier les logiciels en fonction de leur évolution et du lien avec le monde professionnel de l'édition (Pagemaker, QuarkXpress, Adobe InDesign), et j'ai expliqué comment utiliser l'outil de conversion texte-HTML, Markdown, dans le cadre du processus de

(Panckhurst et al. 2004a, coord.), *Evaluation in e-learning : the European Academic Software Award*, coll. MédiaTic n° 3, université Paul-Valéry Montpellier 3, xxii + 134 p.

Un compte rendu de l'ouvrage a été publié dans les *Cahiers de Praxématique* : http://praxematique.revues.org/1699 (Vincent-Durroux, 44, 2005).

L'ouvrage a été lancé lors de la finale à EASA 2004, en Suisse (*cf.* figure 3.5 pour la photo souvenir des trois auteurs et quelques membres fondateurs d'EKMA).

L'objectif de notre ouvrage était le suivant, tel que précisé dans le sommaire :

> [Aborder] la question de l'évaluation dans le domaine de la formation ouverte et à distance (FOAD) et plus spécifiquement celle des problématiques et des méthodologies mises en place dans le cadre d'un concours européen, créé [en 1994], le concours European Academic Software Award (EASA). Cette compétition s'est donné comme objectif d'évaluer des logiciels développés et mis au point, dans des établissements d'enseignement supérieur et de recherche, par des enseignants, des chercheurs, des ingénieurs, des étudiants, etc., et ce, avec ou sans la collaboration d'entreprises privées. Le concours EASA est organisé par l'association European Knowledge Media Association (EKMA).
> Différentes personnes, organisateurs ou jurés, impliquées ces dernières années dans le concours EASA ou dans l'association EKMA, ont voulu faire partager leurs savoirs et leurs expériences dans ce domaine.
> Le but de ce livre est non seulement de donner un aperçu des différentes compétitions passées, mais aussi d'approfondir la réflexion sur différentes questions

publication en autoédition, à l'aide du site Leanpub (https://leanpub.com/). Un professionnel de l'édition, Gilles Pérez, qui m'a grandement aidée avec la réalisation de la mise en pages de l'ouvrage de 2004, inclut une partie de cela dans un cours qu'il co-enseigne avec moi.

Figure 3.5 – Jonathan Darby (Royaume-Uni), Lisa Whistlecroft (Royaume-Uni), Adolf Schreiner (Allemagne), Rachel Panckhurst (France), Sophie David (France), Martin Lehmann (Suisse), lors du lancement de notre livre *Evaluation in e-learning : the European Academic Software Award* à Neuchâtel, le 27 septembre 2004.

> qui importeront dans les années qui viennent, notamment : la portabilité européenne, la diversité des langues européennes, la création d'une maison d'édition, l'établissement de standards, la réalisation d'un répertoire d'experts évaluateurs, l'établissement de protocoles et de directives pour développer et appliquer les critères d'évaluation. [...] Plusieurs chercheurs (européens, australiens et américains) nous ont fait part de leurs réflexions prospectives à propos de l'évaluation et de la formation ouverte et à distance. (Panckhurst et al. 2004b, p. xv).

Au sein de cet ouvrage, la démarche s'annonce comme étant plus pratique que théorique. En effet, l'accent a été mis sur la transmission de l'expérience des années et les concours passés, avec l'apport d'une réflexion sur les moyens à mettre en œuvre pour les concours ultérieurs.

Le concours EASA se déroulait en trois étapes, décrites par Wim Liebrand :

> The first stage consists of a broad call for submissions, encouraging as much participation as possible by inviting entries from all over Europe. At the end of the first stage, only those entries that do not meet the minimal requirements are eliminated, usually about 5%. Normally an EASA competition receives an average of 200 stage 1 entries distributed over 15 disciplines [42].
> In the second stage, entries are categorised by discipline and sent to discipline coordinators who recruit qualified jurors in the relevant discipline. Jurors are teachers, students and practitioners in the various disciplines, and they evaluate entries on both academic and technical content. Each submission is reviewed by three jurors of different background and different countries. At the end of this stage, the best 30 submissions, distributed over disciplines, are selected for the final stage.
> During the finals, usually organised back-to-back with an existing educational conference, the finalists present their application to the audience and to a team of 4 or 5 jurors who evaluate and score the submission on the same criteria as used in the second stage:
>
> — Innovation
> — Design and ease of use
> — European portability
> — Educational materials and approach
> — Evaluation of use
>
> A sophisticated algorithm [43] is used to provide a first ranking of the 30 submissions. An extensive discussion between all the finals jurors (usually about 20) finally yields the best 10 submissions, which receive the prestigious EASA award.
>
> (Liebrand, in (Panckhurst et al. 2004a, p. 3))

Nick Hammond résume le détail des critères d'évaluation [44] des étapes 2 et 3 (Hammond, in (Panckhurst et al. 2004a, p. 64).), *cf.* tableau 3.4 :

42. Les 15 disciplines étaient comme suit, par ordre alphabétique: Arts & Humanities, Biology, Life Sciences and Environment, Chemistry, Computer Sciences, Disabled, Economics, Education, Education — Electronic learning environment, Engineering, Generic, Languages and linguistics, Mathematics, Medicine, Physics, Social and Behavioural Sciences.

43. *Cf.* Hammond in (Panckhurst et al. 2004a, p. 68-70).

44. Dans l'annexe de (David et al. 2005), figure le formulaire d'évaluation utilisé par les membres des jurys EASA.

Tableau 3.4 – Summary of criteria used in EASA Round 2 evaluations. (Hammond, in (Panckhurst et al. 2004a, p. 64).)

Criterion	Short definition	Key aspects
Innovation	Is the project novel in approach or in terms of the activities it supports ?	— Added value — Distinctiveness — Effectiveness — General use of technology
Design, and ease of use	Is the product or approach well-designed and easy to use or apply ?	— Installation or access — User interface — Support — Screen design — Transferability
European portability	Can the software or approach be used (or adapted for use) across Europe ?	— Language for use — Language for support — Language adaptability — Portability of materials — Portability of approach
Educational materials and approach	Are the materials and the approach educationally sound ?	— Users and objectives — User needs — Pedagogical approach
Evaluation of use	Has the software or approach been evaluated, and how good is the evaluation ?	— Thorough evaluation procedure — Results of evaluation

Au niveau des solutions originales pour l'évaluation dans le cadre d'EASA, on peut retenir les aspects suivants : 1) les jurys d'EASA étaient toujours constitués de manière pluridisciplinaire ; 2) les différents critères d'évaluation étaient pondérés par les jurés en fonction de l'importance que ceux-ci avaient pour eux ; 3) le travail d'évaluation durant la finale était systématiquement mené en équipe de 4 ou 5 jurés ; 4) le temps de discussion lors de la finale pour aboutir à une décision collective au sein des équipes de jurés était privilégié.

Outre les points évoqués dans le cadre de (Panckhurst et al. 2004a), discutant, entre autres, des questions de fiabilité, de validité et d'explicitabilité dans le processus d'évaluation (*cf.* Hammond, chapitre 6, p. 61-74), ou analysant l'évolu-

tion du processus pour les coordonnateurs de discipline depuis les évaluations individuelles aux décisions collectives (*cf.* Whistlecroft (chapitre 7, p. 75-82), ou encore explicitant le concours du point de vue des compétiteurs (*cf.* (David et Panckhurst 2004) chapitre 8, p. 83-88) et précisant les améliorations apportées (*cf.* (Panckhurst et Cordewener 2004), chapitre 3, p. 23-41), il fallait approfondir les questionnements à propos du processus d'évaluation.

Si j'ai rappelé ci-dessus le déroulement du concours et les critères d'évaluation, c'est parce que nous avons formé un groupe de réflexion sur les problèmes posés par l'évaluation, qui semblaient se répéter à chaque finale. En effet, à l'issue de concours EASA 2004, nous avons évoqué la nécessité de créer un *Evaluation Working Group (EWG)* [45], qui aurait pour tâche d'effectuer une révision de la grille d'évaluation EASA. David, Panckhurst, Whistlecroft (2005) examinent les problèmes spécifiques pour EASA, mais qui s'ancrent dans une réflexion menée sur l'évaluation dans le cadre du TAL (Cori et al. 2002b) et en recherche et traitement de l'information (Chaudiron 2004) [46]. L'article fait suite au travail mené par les membres du *EWG* de novembre 2004 à avril 2005 [47]. Tout d'abord, les problèmes liés à la grille d'évaluation ne pouvaient être résolus que si EASA prenait une position ferme concernant les points suivants :

1. How does EASA position itself compared to other national, European and international competitions?
2. What values are promoted by EASA?
3. As a consequence: What types of "objects" are allowed to enter?

(David, Panckhurst, Whistlecroft, 2005)

.

45. Les 8 membres étaient comme suit : Peter Baumgartner (Allemagne), Jonathan Darby, Lisa Whistlecroft (Royaume-Uni), Per-Gotthard Lundquist (Suède), Sophie David, Debra Marsh, Rachel Panckhurst (France), Sabine Payr (Autriche).

46. *Cf.* (Segond 2002) pour un ouvrage écrit deux années auparavant à propos du traitement de l'information en contexte multilingue.

47. Les conclusions détaillées sont présentées dans David, Panckhurst, Whistlecroft, (2005), « Many Forms of the Future. A report on future options for the organisation of EASA », rapport interne soumis au comité EKMA et présenté à Oxford le 11 avril, 2005, 38p + vii. *Ce rapport n'est pas inclus dans ma sélection, étant donné sa nature confidentielle*.

Les réponses sommaires à ces questions sont les suivantes (*cf.* David, Panckhurst, Whistlecroft (2005) pour une discussion plus fournie) :

1. EASA must be a unique European competition;
2. EASA promotes the following values: differing European cultures and languages, innovation, best practice, standardisation, accessibility;
3. EASA evaluates products emanating from academia.

Une fois ces réponses déterminées, nous pouvions discuter et proposer une grille d'évaluation révisée. Les 5 critères d'évaluation ont d'abord été discutés :

> **Innovation**
> The main problem with this criterion was that there was a confusion between programming innovation vs. content innovation. As EASA is a "black-box" [48] competition, there is no way in which technical aspects can be verified precisely. This criterion is now explicitly conceptual and not at all technical.
> **Design and ease of use**
> The "Design and ease of use" criterion maintained the ambiguity between the architecture of the system itself and the user interface. The new criterion only looks into installation/access and the interface.
> **European portability**
> Several different interpretations were also possible for portability: computer portability or portability from one natural language to another. We have removed this ambiguity by requiring entries to specify, in some way, their precise European nature, from a linguistic and cultural perspective.
> **Educational materials and approach**
> This criterion posed a problem for research software and generic tools. By reformulating it to read: "Users, approach and content", target users and needs, as well as approach and content, via particular activities, are now clearly defined.
> **Evaluation of use**
> At the moment, it is extremely difficult to check that a piece of software has been evaluated by *real users in real-life situations*, because the current grid does not

48. "The *blackbox* method focuses solely on input and output. The evaluator does not have access to any details of the computer process, which remains a black box. This method is generally used when there is intellectual or commercial copyright, and is often used in competitions. The *glassbox* method implies that the evaluator has access to the whole computing process (structure, algorithms, programming). It includes detailed evaluation, and is often accompanied with measures of intrinsic performance of the software." (Amar et al. 2008).

> address this problem clearly. In addition, more importantly, as we mentioned above, EASA is not a real-user oriented procedure.
>
> — evaluation of use should not be reduced to a mere estimate of the number of users of a software entry, or to the (positive) appreciations indicated by the authors themselves;
>
> — evaluation of use should not rely solely on jurors' intuitions, even if, as experts in their discipline and/or in e-learning, they may have valid intuitions about the real use of the entry. If real evaluation of use is to be conducted, we need to answer the following questions: who is the entry used by? at what moment? in what context? what kind of other resources does it compete with? etc. To summarise, *evaluation of use* versus *evaluation of systems* require radically different methodologies (Le Marec 2004). We have proposed to suppress this criterion entirely.
>
> (David, Panckhurst, Whistlecroft, 2005)

Puis, en fonction de nos recherches théoriques, les échanges au sein du EWG et nos expériences en tant que jurés, coordonnateurs de discipline EASA et au sein du comité d'organisation EKMA, nous étions à même de proposer une réorganisation des 5 critères en 4. Nous avons retenu : *l'innovation ; l'accès/l'installation et l'interface ; Europe : langue et culture ; usagers, approche et contenu*. J'indique ci-dessous la grille d'évaluation révisée, prévue pour le concours EASA 2006 :

1. **Innovation**

Recommendation: content innovation should be the only innovation criterion.

New grid: criterion name maintained

a) *Novelty* (The product includes activities/approaches not covered by other products)
or

b) Improvement (The product includes activities/approaches which are already covered by other products, but it does it in a better/more effective way)

2. **Installation/access and interface**

Recommendation: the installation/access procedure and the interface are checked from the user's point of view.

New grid: change of criterion name from *Design and ease of use*

a) Installation/access (the product is easy to install or access)

b) User interface

   (a) Design and ease of use: the interface is easy to use and the screen design is attractive, effective...)

   (b) Up-to-date standards

   (c) Is the interface appropriate for different end-users (i.e., undergrad/post-grad/students, or research end-users, provisions for handicapped)?

c) Documentation

   (documentation, online help, etc., is provided, and is appropriate and of high quality)

3. **Europe: language & culture**

Recommendation: we need to include multilingual, cultural and curricular aspects which are European specific.

New grid: Change of criterion name from *European portability*

a) Multilingual

   (a) Software exists in 2 or more languages (n/a only for discipline reasons)

   (b) Documentation/online help exists in 2 or more languages (glossary)

   (c) Interface exists in 2 or more languages

b) Culture: differing methods — does the software help bring Europeans together or help understand cultural differences between Europeans?

4. **Users, approach and content**

Recommendation: the new criterion is more general and should be applicable to the different types of entries (research, teaching/learning, generic tools)

New grid: Change of criterion name from *Educational materials and approach*

a) Users & objectives

   (a) the intended use of product is clear and adequately defined;

   (b) the target users of product are clear and adequately defined.

b) User needs (the project addresses real user — teacher, researcher, learner — needs)

c) Approach (the educational/theoretical approach is appropriate)

d) Activities (appropriate activities/content are/is proposed; feedback is provided)

Evaluation of use

Recommendation: We suggest this criterion be eliminated for the following reasons:

— difficult to verify

— other criteria allow for discrimination (and check quality of product) anyway

— users can be checked at stage1 in document submitted by authors.

(David, Panckhurst, Whistlecroft (2005), Annexe 2).

En 2004, Göran Petersson, Bas Cordewener et Lisa Whistlecroft avaient proposé leur vision de la décennie suivante d'EASA en suggérant des *recommandations* :

> In its first decade EKMA has succeeded in maintaining an organisational body to run a biennial competition with broad panels of expert jurors, and to build the basis of an expert community. EKMA is now moving into a new period, a second decade, which will see a growth in professionalism and maturity. EKMA's second decade will be successful only if more European countries become members and

> are represented on the board. This would then enable the establishment of a solid organisation with sufficient financial capability to continue to promote and organise the biennial EASA competition, to lay the groundwork for a European clearing house for quality educational software and provide a quality « marque » for such independently refereed software, to develop and publish guidelines for the production of high quality materials which would be of value across Europe, and to build a community of expert practice, linked by shared Web-based communication, which could work together to enhance learning and teaching across an expanding Europe.
>
> (Petersson, Cordewener, Whistlecroft in (Panckhurst et al. 2004a, p. 96).)

Mais malheureusement, le concours EASA 2006 n'a jamais eu lieu et l'association EKMA — malgré toutes les bonnes volontés et l'énergie sans cesse renouvelée de ses membres — a été dissoute la même année. Malgré l'intérêt manifesté par les instances européennes, le concours devenait de plus en plus difficile à organiser, et le coût, prohibitif [49].

Je retiens de cette formidable aventure, qui a duré 12 ans, (pour ce qui concerne mon implication, et, par conséquent, des collègues/étudiants que j'ai pu impliquer à mon tour dans le processus), des échanges extrêmement fructueux tripartites : intellectuels, langagiers [50], culturels. C'était un vrai travail à envergure européenne, dépassant les frontières et créant un réseau solide d'intellectuels pluridisciplinaires, guidés et motivés par des passions et des objectifs communs, résumés par Wim Liebrand :

49. L'un des problèmes posés était le fait que le coût annuel de la cotisation devait être versé à une association domiciliée en Suisse, et les instances ministérielles nationales n'étaient pas toujours convaincues que les retombées pour leur pays seraient manifestes. Après une première subvention française, l'implication nationale à EKMA/EASA n'a pas été renouvelée.

50. Je me souviens d'une anecdote à propos des langues utilisées pour le développement de logiciels à EASA, lors d'une réunion EKMA. Il était possible de soumettre un logiciel à EASA dans n'importe quelle langue européenne, à condition que les coordonnateurs puissent trouver des jurés pour l'évaluer. Mais pour certains, cette ouverture était trop large. Pour ma part, je défendais bec et ongles le plurilinguisme — en bonne anglophone installée en France depuis des lustres (et ayant vécu dans la belle province de la *loi 101*, le Québec), et défendant la langue française. Un collègue suédois m'a alors rétorqué : « Mais enfin, Rachel, personne ne va jamais soumettre un logiciel en suédois ! ». L'année suivante, à EASA 1998, Kjell Jerselius a non seulement présenté un logiciel en suédois, intitulé *CUT !*, dans le domaine du cinéma, mais, de plus est, il a gagné l'un des dix prix lors de la finale ! Vive la défense de la pluralité linguistique…

> Stimulating the development and use of outstanding academic software [which] would both improve the quality of education and training, and allow Europe to achieve a stronger and more independent position in the global information society. (Liebrand, in (PANCKHURST et al. 2004a, p. 5)).

En 2004, Shirley Alexander donnait quelques pistes pour l'avenir de l'évaluation en FOAD/eLearning, à mon avis toujours d'actualité, plus d'une décennie plus tard :

> The only way forward in my view, is for e-learning practitioners to undertake this important role of evaluation, and engage in evaluation practices which foreground the learners' experience of e-learning (with the technologies and context playing a supporting role), and undertake more holistic, longitudinal evaluation studies, rather than the short term "Polaroid" evaluations of the present. We need to understand the consequences of e-learning over time and within a variety of contexts. As noted by Castells (2001, p. 28):
>
> We engage in a process of learning by producing, in a virtuous feedback between the diffusion of technology and its enhancement... It is a proven lesson from the history of technology that users are key producers of the technology, by adapting it to their uses and values, and ultimately transforming the technology itself.
>
> Gaining a deeper understanding of the ways in which learners experience e-learning and then, as designers of e-learning, adapting e-learning to their uses and values, will help us to develop a deeper, more evidence-based understanding of important questions such as:
>
> — which learners benefit from e-learning?
> — how do learners approach e-learning in a variety of contexts and what do learners think that e-learning is good for?
> — what is best achieved face-to-face, and what is best achieved online?
> — what do students believe they have gained and/or lost as e-learners?
> — how has learners' use of e-learning changed?
>
> We can only start to answer these questions through detailed studies which take a holistic and longitudinal approach to the evaluation from the users' perspective. (Alexander, in (Panckhurst et al. 2004a), Chapter 10, p. 98-99).

L'expérience d'EASA/EKMA, en tant qu'enseignant-chercheur, m'a permis de conjuger recherche fondamentale — concernant les aspects d'évaluation — et recherche appliquée — par le biais des pratiques directes menées dans le cadre du concours, à tous les niveaux. Mes étudiants ont été également impliqués dans le processus d'évaluation dès le départ de mon adhésion au projet, que ce soit en tant que réels évaluateurs pour le concours, ou en phase test. Ils appréciaient notamment la dimension réelle/pratique de leur participation. L'une de mes étudiantes en Master, Virginie Vedel, a préparé le travail de questionnaire auprès des compétiteurs d'EASA (*cf.* chapitre 8, David et Panckhurst 2004).

Même si EASA/EKMA s'est terminé en 2006, la recherche fondamentale en évaluation continuait à nous interpeler. En 2008, à *LREC* (Marrakech), Muriel Amar, Sophie David, Lisa Whistlecroft et moi-même, dans une approche pluri-disciplinaire (conservatrice des bibliothèques, linguistes-informaticiennes et musicienne, compositrice, et *sound designer* !), avons présenté nos travaux concernant des procédures de classification pour l'évaluation de logiciels. Cela tombait à pic dans le cadre d'une conférence dont l'intitulé comporte le nom *évaluation* [51]. Je reprends notre résumé de l'article publié dans les actes :

> A methodological perspective for the classification [of software] is adopted rather than a conceptual one, since a number of difficulties arise with the latter. We focus on three main questions: what to evaluate? how to evaluate? and who evaluates? The classification is therefore hybrid: it allows one to account for the most common evaluation approaches and is also an observatory. Two main approaches are differentiated: system and usage. We conclude that any evaluation always constructs its own object, and the objects to be evaluated only partially determine the evaluation which can be applied to them. Generally speaking, this allows one to begin apprehending what type of knowledge is objectified when one or another approach is chosen. (Amar et al. 2008).

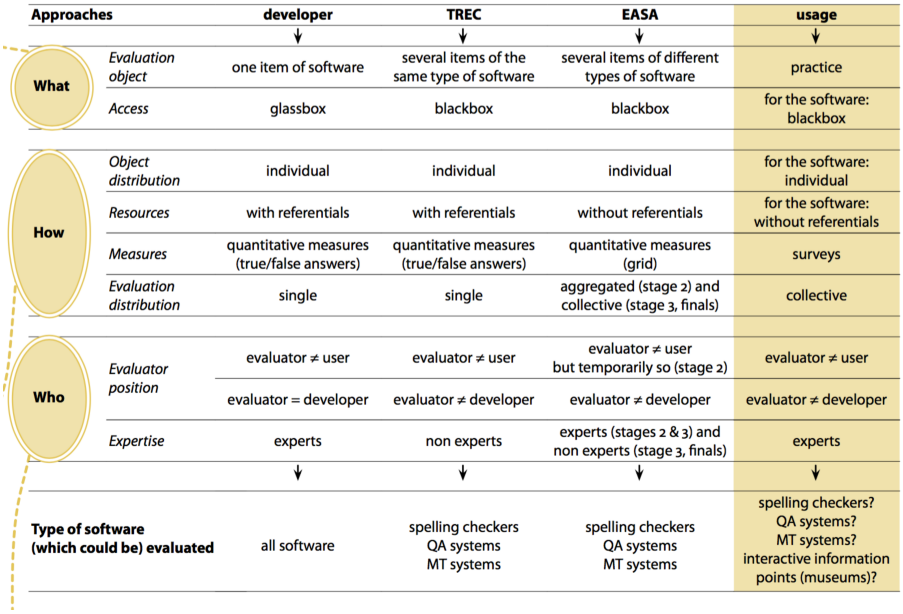
Nous avions réfléchi aux trois questions centrales dans le cadre de l'évaluation de logiciels : Qu'est-ce qu'on évalue ? Comment on évalue ? Qui évalue ? À la première question (*what ?*), on peut répondre qu'on évalue les objets (les logiciels), isolés ou au contraire intégrés dans leur contexte d'utilisation. On évalue également en utilisant des approches (*blackbox* vs *glassbox*, *cf.* note 48). À la deuxième question

51. *LREC=Language Resources and Evaluation Conference*.

(*how ?*), on distribue les objets et on les évalue un à un, individuellement, ou de manière comparative. On évalue (ou non) à l'aide de ressources, ou de *referentials*, qui correspondent à des connaissances normées, consensuelles, stables. On peut mesurer également de manière quantitative (tests/questions, etc.) ou qualitative (utilisation de méthodologies spécifiques, questionnaires, etc.). Enfin, l'évaluation est menée de manière *simple* (un seul évaluateur), *agrégée* (plusieurs évaluateurs, et les résultats de plusieurs évaluations sont ensuite combinés), *collective* (plusieurs évaluateurs produisent une évaluation unique, à propos de laquelle ils ont négocié/se sont mis d'accord). La troisième question (*who ?*) explore, d'une part, la *position* de l'évaluateur : a) *évaluateur et développeur* : ce cas de figure est rare, sauf dans des méthodes *glassbox*. Dans des concours, l'éthique exige qu'il s'agisse de deux personnes distinctes ; b) *évaluateur et utilisateur* : si l'évaluateur observe l'utilisateur, les deux personnes sont nécessairement distinctes ; l'évaluateur peut temporairement adopter la posture de l'utilisateur. D'autre part, l'*expertise* de l'évaluateur est nécessairement évaluée : a) on a souvent recours à des évaluateurs non-experts dans des méthodes avec des référentiels, car on leur fournit une série de points (check-list) à vérifier ; b) les évaluateurs-experts interviennent dans des méthodes avec ou sans référentiels et ils jugent la qualité/pertinence de la réponse dans un contexte donné. (Extraits synthétisés et traduits de Amar et al. 2008.)

La figure 3.6 reprend les points évoqués ci-dessus et compare plusieurs approches/concours (*cf*. Annexe, Amar et al. 2008).

(Amar et al. 2008) est la dernière publication à laquelle j'ai participé portant sur l'évaluation. Depuis 2006, je m'étais rapprochée de Debra Marsh, experte en FOAD/eLearning pour l'apprentissage des langues étrangères. Je l'avais rencontrée à EASA 1998 (elle était finaliste, et la plateforme *Merlin* qu'elle présentait a gagné l'un des 10 prix décernés). Suite à son déménagement dans le sud de la France, Debra et moi avons décidé de creuser nos recherches ensemble, ayant de nombreux centres d'intérêt en commun. Je l'avais sollicitée dans le cadre de l'ouvrage de 2004, afin qu'elle donne sa vision d'une approche « intégrée » de l'évaluation en FOAD/eLearning. Avec le logiciel *Merlin*, elle avait montré que si certaines plateformes/VLE (*Virtual Learning Environments*) pouvaient être effectivement sophistiquées d'un point de vue technologique, d'autres se consacraient davantage aux aspects concernant les innovations pédagogiques.



Approaches		developer	TREC	EASA	usage
		↓	↓	↓	↓
What	*Evaluation object*	one item of software	several items of the same type of software	several items of different types of software	practice
	Access	glassbox	blackbox	blackbox	for the software: blackbox
How	*Object distribution*	individual	individual	individual	for the software: individual
	Resources	with referentials	with referentials	without referentials	for the software: without referentials
	Measures	quantitative measures (true/false answers)	quantitative measures (true/false answers)	quantitative measures (grid)	surveys
	Evaluation distribution	single	single	aggregated (stage 2) and collective (stage 3, finals)	collective
Who	*Evaluator position*	evaluator ≠ user	evaluator ≠ user	evaluator ≠ user but temporarily so (stage 2)	evaluator ≠ user
		evaluator = developer	evaluator ≠ developer	evaluator ≠ developer	evaluator ≠ developer
	Expertise	experts	non experts	experts (stages 2 & 3) and non experts (stage 3, finals)	experts
		↓	↓	↓	↓
Type of software (which could be) evaluated		all software	spelling checkers QA systems MT systems	spelling checkers QA systems MT systems	spelling checkers? QA systems? MT systems? interactive information points (museums)?

Figure 3.6 – Évaluation. (Amar et al. 2008)

En même temps, dans sa contribution sollicitée (Marsh in (Panckhurst et al. 2004a)) elle n'était pas dupe ; elle était consciente qu'il ne fallait pas opposer les deux considérations technologique et pédagogique, mais faire en sorte qu'elles se soutiennent mutuellement :

> The future of evaluation in e-learning will necessarily see a more integrated approach to establishing its objectives, desired outcomes and overall effectiveness. Evaluation will no longer consist of either a consideration of the technology or of the pedagogical innovation, but will consist of an approach which considers how the one supports the other, how together they represent an integral part of the learning process, and why in some cases the combination of a particular technology and a specified pedagogy are inappropriate.
>
> (Marsh, in (Panckhurst et al. 2004a), p. 99).

Par le biais d'une citation de Thoreau en exergue, que je trouve très juste, elle met en garde contre la domination par la seule technologie :

> Men have become tools of their tools (Henry David Thoreau (1817-1862))

Si les tuteurs/apprenants cherchent constamment à s'adapter à la technologie, il faut rappeler, au contraire, que la technologie doit s'adapter, être le soutien, à l'apprentissage :

> No matter how technically sophisticated the technology may be, no matter how much video or audio can be delivered directly to the individual learner, and no matter what complexity or array of communication tools are available to the learners…the future of e-learning lies in the individual learner's ability to learn. And, in order to learn, the individual requires the tools which will support him/her appropriately within a given context and for a given need. The tools, the context and the need cannot be separated out, but require an integrated approach to evaluation in order for us to move forward. Only in this way will we avoid the perpetuation of current practice which sees many first time users of e-learning tools « seek to adapt the way they work to the way the software needs things to be done »
>
> (Britaine & Liber, 2004). (Marsh, in (PANCKHURST et al. 2004a), p. 100-101).

La théorie du pionnier ayant co-dirigé le laboratoire d'intelligence artificielle au M.I.T., Seymour Papert (décédé récemment en 2016 [52]), qui s'inspirait de Piaget, préconisait le *constructionisme* par opposition à l'*instructionisme*. Pour celui qui a inventé le langage informatique *Logo*, dans l'objectif d'initier les enfants à l'informatique — d'ailleurs le premier langage dont j'ai appris à me servir avec enthousiasme dans mon cursus universitaire — il préférait que les étudiants bâtissent leurs connaissances en se servant de situations et d'objets concrets, plutôt qu'à partir de propositions abstraites. Par ailleurs, souligne-t-il :

> The deep difference between education past and future: in the past, education adapted the mind to a very restricted set of available media; in the future, it will adapt media to serve the needs and tastes of each individual mind.
>
> (Seymour Papert, *Wired Magazine*, 1993)

52. http://www.nytimes.com/2016/08/02/technology/seymour-papert-88-dies-saw-educations-future-in-computers.html?_r=0, (consulté le 9 janvier 2017).

De mon point de vue, si la technologie ne peut s'adapter aux besoins des usagers, soit on la modifie soit on change d'outil. Souvent, en FOAD/eLearning, on n'a pas toujours besoin d'outils technologiquement sophistiqués — au contraire, je les ai parfois fuis (*cf.* § 3.3.2.6 pour un exemple de réseau d'échanges pédagogiques avec le Royaume-Uni) — le tout est dans l'accompagnement approprié et l'apprentissage de l'autonomie guidée. À suivre avec les questionnements sur les innovations pédagogiques dans le prochain paragraphe.

3.3.2.6 Réseaux d'échanges pédagogiques en FOAD/eLearning (2006-2012)

À la rentrée 2004, un Master professionnel en Sciences du langage — unique en France, car le premier à délivrer un diplôme dont le M2 était entièrement mené à distance « Gestion des connaissances, apprentissages et formation ouverte et a distance » [53] (sous la responsabilité de Chantal Charnet) — a été mis en place au sein de notre département. M'intéressant depuis quelques années aux dispositifs technologiques/pédagogiques novateurs en eLearning, mes enseignements, et, par conséquent, une partie de mes travaux de recherche, menés en collaboration avec Debra Marsh [54] ont porté sur l'apprentissage collaboratif, en plaçant réellement les apprenants au centre du processus. Nous voulions évaluer les avantages et les défis lies à l'incorporation de réseaux sociaux pour des échanges pédagogiques, à l'ère du Web sémantique, au sein de cursus universitaires pré-établis. Pour nous, mettre en place un apprentissage en ligne efficace impliquait : *autonomie, indépendance, partage, soutien, création, responsabilité*. Puisque la première promotion d'étudiants avait fini leur Master en 2006, il nous semblait important de vérifier le taux de satisfaction des étudiants

53. Au moment de la création du diplôme, le M1 était enseigné de manière hybride : partiellement en présence, partiellement à distance. Le M2, en revanche, était entièrement mené à distance. Le diplôme actuel « Humanités Numériques » est uniquement fourni à distance pour les deux années, avec une période de regroupement obligatoire, généralement au premier semestre.

54. Debra est chef d'entreprise et experte en eLearning. Basée en France depuis plus d'une décennie, elle avait été auparavant responsable de la formation ouverte et à distance a l'université de Hull pendant de nombreuses années, et elle y a dirigé le projet Merlin (*cf.* p. 117) — la plateforme d'enseignement à distance de l'université de Hull — et ce à partir de la phase d'élaboration jusqu'à son implémentation. Ses intérêts en recherche incluent les aspects pédagogiques et le design en eLearning et elle a co-publié un ouvrage sur l'éducation en ligne (Bennett et al. 2007). Après avoir travaillé avec l'université de Cambridge et Cambridge University Press, elle travaille désormais en tant que chef d'entreprise de *Learning Design, France*.

et vérifier que l'équipe pédagogique avait répondu à leurs besoins et attentes. Nous avons exposé les résultats du questionnaire distribué aux étudiants (aux tuteurs et aux enseignants) — afin de recueillir leur feedback qualitatif et ensuite procéder à une légère modification du cursus — puis les recommandations pour les années à venir dans (Panckhurst et Marsh 2006) :

> In order to respond to specific professional needs at Master's level, a review of the overall pedagogical design is required. A lockstep, building-block model (Salmon 2000) [55]) may be used at undergraduate and initial M1 level, but M2 students need a more flexible model which is responsive to their individual and collective needs and contextualises their learning within their personal professional context. Thus, networking becomes central to the learning activity. Emphasis is placed on collaborative learning during which information exchange and knowledge construction become cyclical and applicable to the differing professional contexts.
>
> (Panckhurst and Marsh 2006)

Notre conclusion majeure était qu'il fallait proposer une approche plus expérimentale voire *expérientielle*, centrée sur les apprenants eux-mêmes :

> We concluded there was a clear need to explore a more experiential approach to learning and teaching on the programme. An approach in which the focus of learning becomes the individual learners themselves — their backgrounds, experiences, interests, capacities and needs, and in which learners take more responsibility for their own learning "as members with rights and responsibilities, power and vulnerability, and [who] learn to act responsibly, considering the best interests of themselves, other individuals, and the group as a whole"
>
> (Carver, 1997, p. 146, in (Panckhurst and Marsh 2007)).

David White le résume joliment ainsi :

> Education courses need to be tailored to people and not people tailored to courses.
>
> [David White, conférence plénière, EADTU, Lisbonne, novembre 2007, in (Panckhurst et Marsh 2008b, p. 3).

55. Salmon's 5-step model includes: 1. Access and motivation, 2. Online socialisation, 3. Information exchange, 4. Knowledge construction, 5. Development (Salmon 2000, p. 26).

Les étudiants avaient besoin d'une mise en réseaux, centrale à leur activité collaborative. Afin de mettre l'accent sur l'utilisation de réseaux, en 2006-2007, nous avons mis en place une expérimentation qui a consisté à créer un réseau d'échanges pédagogiques en eLearning (RÉEL), eLearning exchange network (ELEN), à l'aide de Ning [56] (www.ning.com). Le choix de ce réseau social était réfléchi. Facebook et Ning existaient tous deux depuis 2004 ; nous avons choisi le second afin d'éviter de « mélanger les genres » en ne proposant pas le même outil pour les échanges sociaux et pédagogiques. Mais pourquoi n'avoir pas tout simplement choisi la plateforme de cours universitaire (à l'époque, WebCT, au sein de notre établissement) ? Outre les raisons pédagogiques (cours sur l'évaluation de logiciels et les innovations pédagogiques), notre expérimentation a consisté à créer un réseau bilingue et biculturel entre les étudiants de M2 de notre université et ceux de l'université de Hull, en Angleterre. De ce fait, il aurait été impossible pour les étudiants anglais de se connecter et partager les discussions sur notre plateforme institutionnelle, seulement accessible aux étudiants inscrits au sein de l'université Paul-Valéry Montpellier 3. Nous avons donc privilégié l'échange via le web, mais au sein d'une communauté privée.

On consultera (Panckhurst et Marsh 2007), pour une analyse du processus et des remarques des étudiants ayant participé à l'exercice, notamment concernant la facilité d'utilisation/d'accès, leur sentiment de liberté, d'appropriation du processus d'apprentissage, leur prise de responsabilité concernant l'apprentissage autonome, la partage des connaissances dans un contexte d'apprentissage collaboratif, leur appartenance à une communauté virtuelle qui permettait, à leurs yeux, de promouvoir l'innovation pédagogique, etc.

Nous avons conclu que l'apprentissage collaboratif peut être mené de manière efficace à l'aide d'un réseau social :

> This pilot study has shown that collaborative learning can take place in a social network and that collaborative learning can be effective online for the purpose of achieving an academic goal and that an active exchange of ideas within small groups not only increases interest among the participants but also promotes

56. Quelques années plus tard, nous avons remplacé Ning, (devenu payant), par grou.ps (http://www.grou.ps). Plus récemment, dans mes cours menés depuis 2012, j'ai choisi d'évoluer vers *Google+* (https://plus.google.com), puis plus précisément vers les communautés *Google*, que j'utilise à l'heure actuelle.

> critical thinking (Gokhale, 1995 ; Johnson & Johnson, 1986 ; Totten, Sills, Digby, & Russ, 1991). In addition, interest and motivation can be sustained online, despite the contrary being reported that in many online groups, participation can drop to « zero » and that despite the initial vibrancy of online communities, large numbers of them fail (Ling *et al.*, 2005) [57].
>
> (Panckhurst et Marsh 2007).

Cependant, pour que les RÉEL/ELEN soient des outils d'apprentissage réussis, la recherche que nous avons menée stipule plusieurs critères fondamentaux :

> This research has identified the following as fundamental for success of an international bilingual social network. Learners:
>
> — must be familiar with working online
>
> — need to be aware with basic principles of eLearning
>
> — need to be used to self-directed/placed in learning context where they take responsibility for own learning
>
> — need to grasp use of ELENs within general curriculum in order to understand purpose/focus.
>
> (Panckhurst and Marsh 2007)

Cette première expérimentation nous a fourni des pistes pour la recherche et le développement/approfondissement des réseaux ultérieurs :

> If eLearning is to offer improved learning opportunities, educators will have to rethink the models that underlie eLearning [...] Progress will depend on embracing learner-centered models that place the student at the focal point, not the teacher and not the classroom [...] While eLearning based on classroom-centered models is not necessarily poor instruction, it certainly fails to optimize what eLearning could be and fails to optimize the students' learning experiences.
>
> (Carver el al. 2007, in (Panckhurst et Marsh 2007)).

À travers deux expérimentations pédagogiques à l'aide de RÉEL/ELEN menées en 2007 et pendant l'année universitaire 2007-2008, (Panckhurst et Marsh

57. On se reportera à (Panckhurst et Marsh 2007) pour les références précises de ces auteurs.

2008a), nous avons affiné nos critères et principes précédents, notamment en insistant davantage sur le « guidage » des apprenants vers une démarche d'appropriation, puis d'évaluation et d'autonomie semi-guidée. Je résume les critères clefs sous forme de 4 points ci-dessous :

1. ***Opportunities presented by social networking tools for development of communities of practice across different professional groups***
   - ease of use,
   - access to multicultural groups within same/similar professional context,
   - access to individuals/groups which is not possible if limited to face to face content.
2. ***Benefits of Social Networking Tools (which use the Open Web) over institutional platforms***
   - ease of use/setting up,
   - autonomy for learners and tutors i.e. no need for specialist technical support,
   - choice of features/design/look and feel left to learners/tutors,
   - each network has individual/group feel,
   - dynamic network built from scratch; nothing imposed from "outside" the group.
3. ***Identifying key fundamental principles/criteria required to promote effective communities***
   - a sense of purpose,
   - group cohesion,
   - tutor guidance becomes learner self-group management,
   - learners provided with guidance to encourage/promote independence/autonomy,
   - learners need sense of "ownership",
   - teaching staff need to be prepared to "let go" and "take a back seat",
   - if assessment required think appropriately: number of postings not appropriate; reflection/diary/summary of activity relating to own professional context is appropriate.

4. ***Opportunities presented by social networking tools for multicultural and bilingual collaborative learning***
   — ease of use/connection if sitting outside institutional technical constraints. (Panckhurst and Marsh 2008a, p. 10-11).

Cette étude a confirmé nos convictions : le web peut être utilisé à bon escient pour l'apprentissage collaboratif, à condition que les tuteurs et les apprenants, acceptent quelques modifications vis-à-vis des rôles « traditionnels ». Afin de créer un RÉEL/ELEN efficace qui continue à motiver et susciter l'intérêt de ses membres, les tuteurs doivent déployer une énergie initiale très importante dans la phase de mise en place du réseau pédagogique, et ce de manière à *soutenir* la communauté :

> La qualité de l'apprentissage peut être améliorée à l'aide des réseaux sociaux à condition que l'apprenant et l'enseignant soient prêts à assumer des rôles respectifs différents voire novateurs : l'apprenant doit prendre la responsabilité par rapport à son propre apprentissage et l'enseignant ne doit plus diriger l'apprentissage à proprement parler mais il doit être en mesure de l'orienter, de l'encourager et de le faciliter. (Panckhurst et Marsh 2008b, p. 9-10).

Puis, les étapes de planification et de structuration du réseau sont cruciales afin que les apprenants puissent prendre la responsabilité de leur propre apprentissage. Des tâches spécifiques seront ensuite proposées à chaque apprenant, afin de privilégier l'autonomie :

> Learning will become « autonomous » as long as tutors/facilitators « guide » rather than « manage » and change their roles so that learners feel comfortable with taking the initiative. (Panckhurst et Marsh 2008a, p. 11).

Notre quatrième étude de cas (Panckhurst et Marsh 2009) conduite pendant l'année universitaire 2008-2009, autour de fils de discussion menés par les étudiants, conclut sur la nécessité de progresser vers des RÉEL/ELEN de deuxième génération. En effet, *a few years down the track*, les étudiants avaient déjà une habitude certaine des réseaux sociaux, et le sentiment de « nouveauté » de cette utilisation pédagogique était caduque. Un point d'ancrage ou de motivation supplémentaire était nécessaire. Après tout, que pouvions-nous offrir de plus ?

> Simply continuing with the same formula as previous case studies is no longer sufficient. Students' needs, expectations and skills in online exchange have moved on. As a consequence, it is time to evolve to the next generation of social networks in education. This does not require a change of technology, but a step forward in thinking and approach in how to organise and support social networking and learning. In other words, the very nature and purpose of engagement and motivation online needs to be reviewed and the results acted upon.
>
> (Panckhurst et Marsh 2009)

La recherche menée pendant cette période (Anderson, 2009, Conole *et al*., 2008, Downes, 2008, Siemens 2010, Weller 2008, etc.), nous a convaincues : on devait adopter une approche implémentant des *objets d'apprentissage sociaux,* qui « facilitent la conversation et ainsi l'interaction sociale » (Weller, 2008, in (Panckhurst et Marsh 2009)).

> La leçon qui est difficile à retenir dans ce contexte pour des universitaires est que la valeur éducative n'est pas directement présente dans le contenu lui-même, mais dans l'interaction sociale qui est engendrée par ce contenu.
>
> (Weller, 2008, notre traduction, in (Panckhurst et Marsh 2011a), p. 294).

Je reviendrai sur ce point, dans (Panckhurst 2012), concernant le rôle voire le *lâcher prise* des enseignants/tuteurs, à mon avis fondamental pour la réussite effective des RÉEL/ELEN.

Par ailleurs, il est important de retenir que l'outil peut avoir un « profil technologique bas », aisé d'utilisation et permettant de faire relativement peu de choses. Il s'agit surtout de réfléchir à la mise en place et l'accompagnement pédagogiques de l'utilisation et de l'exploitation de l'outil :

> Les réseaux sociaux exigent des objets sociaux. Ces objets sociaux facilient l'interaction sociale. Par ailleurs, les réseaux sociaux n'exigent pas une technologie novatrice, mais plutôt une approche pédagogique novatrice et approfondie. Anderson (2009) évoque l'intérêt d'avoir recours aux réseaux d'apprentissage pédagogiques dans un cadre d'enseignement supérieur :
>
> Le design des réseaux sociaux ouvrira des voies plus efficaces, plus efficientes et plus motivantes pour l'apprentissage que toutes les formes précédentes — y

> compris l'éducation traditionnelle en présentiel et la formation en ligne.
>
> (Anderson, 2009, notre traduction), in (Panckhurst et Marsh 2011a, p. 294).

À travers notre cinquième étude de cas (année universitaire 2010-2011), notre RÉEL/ELEN de deuxième génération adopte l'utilisation d'*objets d'apprentissage sociaux*. Précédemment, des fils de discussion étaient menés individuellement par les étudiants. Désormais, ils avaient à accomplir des travaux collaboratifs imposés (dont le contenu devait servir d'objet d'apprentissage social) en groupes de 4-5, incluant une période de préparation, une discussion menée entre et pour les pairs, et une rédaction synthétique finale, remise aux pairs et aux tutrices, le tout dans le cadre d'un planning très strict et très précis (Panckhurst et Marsh 2011b). C'était un pari, car on craignait de compromettre *la diversité, l'autonomie, l'ouverture et l'interaction* — concepts préconisés par (Downes 2008), associés à l'utilisation de réseaux. Mais les retours des étudiants ont été très encourageants, à la fois à propos des *sujets imposés* :

> D'abord, le fait que le sujet soit imposé. Ceci permet de traiter d'un sujet auquel on n'aurait pas nécessairement pensé, puis cela représente un gain de temps considérable. On rentre directement dans le vif du sujet sans perdre 2 à 3 jours dans le choix du sujet. De plus, on trouve enfin le temps de créer un dossier plus original et plus interactif. (Étudiante, 2009-2010).
>
> (Panckhurst et Marsh 2011a, p. 296).

puis concernant la *responsabilisation*, l'*autonomie*, la *confiance*, la *cohésion* du groupe, leur *appartenance* au RÉEL/ELEN, l'*indépendance* vis-à-vis des enseignantes/tutrices, l'*apprentissage collaboratif* entre pairs :

> L'autonomie presque totale que nous avons eu pour les activités sur Ning a été un plus. Cela nous a appris à gérer une consigne en groupe, à confronter nos idées, et nos doutes, sans qu'un enseignant vienne nous orienter. Je pense qu'une confiance a été mise envers les étudiants et cela a payé par nos réalisations. Encore une fois, on peut mettre cette situation en parallèle avec le monde du travail. (Étudiant, 2009-2010)
>
> [...] le fait d'avoir accès au travail de chaque groupe est une démarche rare, et très bénéfique qui nous donne un maximum de connaissances, de notions très

> intéressantes et utiles pour la suite. (Étudiant, 2009-2010)
>
> (Panckhurst et Marsh 2011a), p. 296

Les critères clefs pour un apprentissage efficace indiqués supra ((Panckhurst et Marsh 2008a)), peuvent être schématisés dans la figure 3.7 :

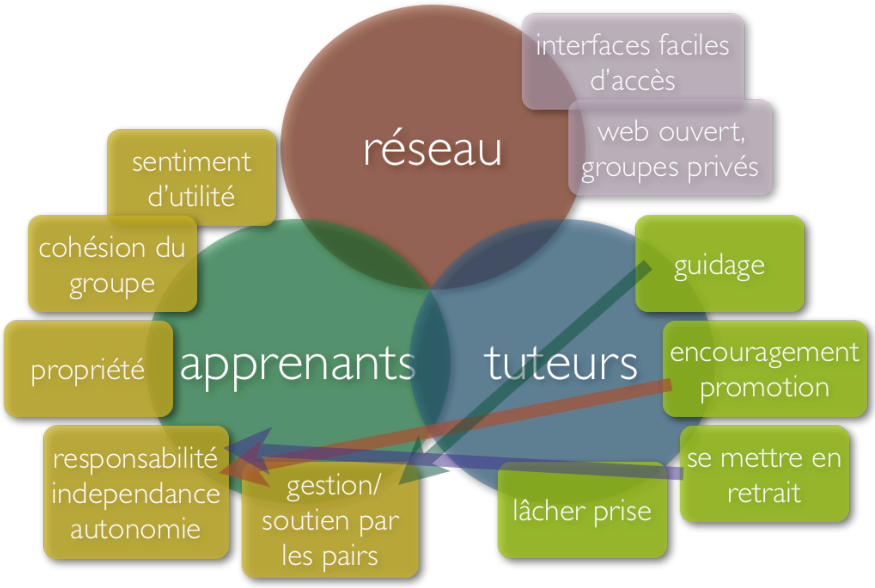


Figure 3.7 – Critères clefs pour un apprentissage efficace au sein d'un RÉEL/ELEN. (Panckhurst et Marsh 2011a).

Ces critères rejoignent les 7 points clefs pour un apprentissage efficace, énoncés par Siemens (2010) :

1. Les apprenants doivent avoir le *contrôle* de leur propre apprentissage. L'*autonomie* est la clef. Les éducateurs peuvent initier, peser le pour et le contre, et *guider*. Mais un apprentissage significatif exige des *activités menées par les apprenants*.
2. Les apprenants ont besoin de faire l'expérience d'une certaine *confusion* et de *chaos* pendant le processus d'apprentissage. Clarifier le chaos constitue le cœur de l'apprentissage.

3. Du *contenu* et une forme *d'interaction ouverts* augmentent la perspective de *connexions au hasard* qui entraînent l'*innovation*.
4. L'*apprentissage* exige à la fois du *temps*, de l'*approfondissement*, de la *pensée critique* et de la *réflexion*. Ingérer de l'information nouvelle exige du temps pour digérer. Trop de personnes font de *l'engloutissement digital* sans prendre le temps de la *digestion*.
5. *L'apprentissage est la formation de réseaux*. Les *connaissances* sont *distribuées*.
6. La *création* est *vitale*. Les apprenants doivent créer des artéfacts afin de les partager avec autrui et afin de recentrer l'exploration au-delà des artefacts fournis par l'éducateur.
7. *Comprendre* la complexité exige des systèmes *sociaux* et *technologiques*. Nous effectuons le premier mieux que le second.

   (Siemens, 2010, notre traduction, in (Panckhurst et Marsh 2011a), p. 297).

Pour que le RÉEL/ELEN réussisse, il faut que les objets d'apprentissage sociaux soient *attirants*. Les étudiants ont adopté très rapidement ces nouvelles procédures, impliquant *autonomie, indépendance, partage, soutien, création, responsabilité*.

En revanche, ce qui commençait à me préoccuper davantage maintenant était le tuteur, l'enseignant. La conclusion de (Panckhurst et Marsh 2011a) annonce cette orientation :

> Quels rôles, quelles places, quelles attitudes, l'enseignant doit-il assumer, adopter, dans ce nouveau paradigme ? Cette posture n'est pas simple ; loin s'en faut. L'enseignant hésite, ne sait pas réellement où est sa place, à quel moment il doit intervenir. Ce sera la prochaine étape de l'étude : le tuteur et le changement de paradigme. (Panckhurst et Marsh 2011a), p. 297.

Les deux dernières études de cas, que j'ai menées à partir d'enseignements assurés en tandem avec un jeune collègue de mon département [58] jusqu'en 2012, ont

58. Debra Marsh a été contrainte d'interrompre ses activités d'enseignement ponctuel au sein de notre établissement à cause d'une charge professionnelle trop contraignante. J'ai donc sollicité un collègue en interne afin de partager le cours/tutorat avec moi.

permis de dégager les six critères clefs suivants (*cf.* tableau 3.5) pour un tutorat efficace et réussi :

En revanche, et cette dernière expérience a permis de le démontrer : tous les enseignants n'ont pas la capacité de travailler avec les étudiants/apprenants dans ce type d'environnement. La conclusion de (Panckhurst 2012) l'atteste :

> Digital tutors/educators/facilitators are currently living in an era of *unease*. The paradigm shift has happened. A junior colleague who recently joined me on the second-year distance-education Master's course I run online, expressed the following: "I feel that my presence is totally unnecessary in this course; I don't know what position to adopt and it makes me uncomfortable." He subsequently suggested another colleague take his place next year, while I urged him to stay on. After hesitating for one or two months, he finally decided to stop co-tutoring the course. Notwithstanding, it is highly important to persevere. As digital tutors, we may well feel *uncomfortable*, as if we are doing very little when behind the scenes instead of being centre stage. But in *learning for the future,* less leadership and control from teachers and more ownership, responsibility and autonomy for students are crucial for learners. As digital tutors, we will gradually adjust and accept being able *to lose control and make mistakes*, learning both *with* and *from* our students — the true *future makers*. (Panckhurst 2012), p. 738.

Dans notre publication sous forme de monographie de 20 pages en une version bilingue anglais/espagnol, toute la recherche fondamentale et appliquée menée en collaboration avec Debra Marsh est synthétisée et présentée à l'aide

Tableau 3.5 – Six étapes importantes pour un tutorat réussi au sein de RÉEL/ELEN (Panckhurst 2012)

1.	Spend initial time carefully setting up and structuring the eLEN.
2.	Initiate ice-breaking activities (including surprise activities that create suspense) so that students can gain confidence, take ownership and feel trusted.
3.	Introduce structured imposed activities and stringent timelines.
4.	Make sure two tutors work together[59] on each online course; this is important, so that they can provide support for each other and decide whether they should or should not intervene at precise moments.
5.	Let go of control — sit back and trust that learners, after initial tutor support and guidance, will progressively learn autonomously providing peer-support for each other.
6.	Accept and expect to make mistakes.

d'exemples détaillés et de citations d'étudiants/apprenants (Panckhurst et Marsh 2011b). On pourra également consulter la vidéo en ligne de ma conférence invitée à la deuxième journée TICE à l'université d'Avignon et des Pays de Vaucluse (UAPV)(Panckhurst 2011).

3.3.2.7 Conclusion

Si le volet 1 (§ 3.3.1) était consacré à mes recherches fondamentale et appliquée en proposant mes outils implémentés, le volet 2 (§ 3.3.2) est caractérisé, d'une part, par l'utilisation d'outils de type TICE/FOAD/eLearning par autrui, et, d'autre part, par la réflexion et la recherche induites de ces usages. *(Auto)formation, (auto)évaluation, pédagogie renouvelée, réseaux d'échanges pédagogiques, mutations*, sont les maîtres mots de cette période plurielle. En matière de disciplines CNU, je me situe, dans le cadre de ces recherches, à la frontière des sections sciences du langage/linguistique-informatique (CNU 7), sciences de l'éducation (CNU 70), sciences de l'information et de la communication (CNU 71) [60]. Par ailleurs, dans des positionnements ultérieurs, il pourrait être intéressant d'envisager plus en détail le mariage disciplinaire entre enseignement du TAL et eLearning, notamment pour la création de contenus pédagogiques ouverts (*Open Educational Resources*) permettant ainsi d'approfondir la collaboration en inter-sections CNU (*Cf.* à ce sujet (Krstev et al. 2015)).

On remarquera, depuis 2004 (était-ce un tournant suite à mon CRCT à Sydney ?), une augmentation sensible de mes publications en binôme (ou à plusieurs auteurs, notamment pour le volet 3) [61]. Il faut dire que cela est devenu ma marque de fabrique de travail en recherche. Je ne conçois absolument plus la recherche comme un exercice isolé, à mener seul. Cela ne m'intéresse plus. En sciences

60. En revanche, le volet 3 (§ 3.3.3), partiellement mené en parallèle au volet 2 (§ 3.3.2), est entièrement inscrit en linguistique-informatique/TAL : les chercheurs appartiennent aux sections CNU 07 et 27.

61. Les rares exceptions sont lorsque j'écris des hommages (Panckhurst 1997b), (Panckhurst 2009) ou bien, pour un colloque donné, lorsque seuls les auteurs présents ont le droit de faire publier leur communication dans les actes. Par exemple, c'était le cas pour (Panckhurst 2006b, 2010, 2012, 2013, 2015). D'ailleurs, même quand j'écrivais des articles, ou je menais quelques recherches seule, j'essayais de partager le plus possible avec mes pairs, de présenter mes travaux en séminaire, à colloque, ou devant les étudiants afin d'avoir un retour précieux.

« dures » la recherche collaborative demeure plus courante qu'en sciences humaines et sociales, mais mis à part ce point, la recherche ne m'apporte que par sa dimension partagée, par ses moments d'échanges et de discussions approfondis. De même que mes étudiants nourrissent mes recherches par leurs questions et leurs remarques souvent astucieuses, mes collègues *keep me on my toes*, en réagissant, en échangeant, en me faisant avancer, tout simplement, dans ma réflexion, et vice versa. Je reviendrai sur ce point lors des recherches du volet 3 (§ 3.3.3), dont l'apogée de la recherche collaborative se situe entre 2011 et maintenant.

Comme stipulé *supra*, le volet 2 est une période d'exploration par excellence, pluridisciplinaire, pendant laquelle j'ai continué à élargir mon réseau de chercheurs au-delà des frontières des sciences du langage [62]. J'ai évolué d'une recherche au sein de laquelle la donnée de type *exemple* était privilégiée (volet 1) à des *situations authentiques* dans lesquelles étaient investis les acteurs enseignants/chercheurs/étudiants (volet 2). Le cheminement continue (en parallèle) en privilégiant cette mouvance vers le recueil et l'analyse de *données authentiques* (volet 3).

Les actions d'(auto)formation, d'(auto)évaluation, de mise en place de réseaux d'échanges pédagogiques (RÉEL/ELEN), et les recherches induites qui caractérisent ce deuxième volet m'ont beaucoup appris sur la recherche pluridisciplinaire dans le cadre d'un réseau de chercheurs élargi. Le positionnement voire le rapprochement par rapport aux questionnements scientifiques des chercheurs de mon laboratoire de recherche *Praxiling* (qui étaient pour certains, praxématiciens — tout au moins à l'origine) avait tendance à être ponctuel (Panckhurst 2001, 2017). Cela n'est pas étonnant, dans la mesure où le traitement automatique de la pragmatique pose énormément de problèmes encore non résolus de nos jours — que je n'ai jamais tenté d'explorer, il faut le reconnaître, me situant plutôt du côté de la syntaxe et de la sémantique lexicale. Comme la praxématique peut être comparée (dans certains cas) à la pragmatique [63], notre rencontre intellectuelle pouvait sembler difficile à imaginer. En tout état de cause, le travail

62. Comme indiqué, non seulement mes recherches impliquaient les sciences de l'éducation et les sciences de l'information et de la communication, mais par le biais du concours EASA j'ai élargi mon réseau de chercheurs à une quinzaine de disciplines différentes. C'était vraiment stimulant.

63. Pour les membres de l'école de pensée praxématicienne, praxématique et pragmatique se rapprochent par certains points et se distinguent par d'autres : « Pour la praxématique, le

sur les *données authentiques* — et plus généralement sur l'*analyse de discours* — que j'ai démarré réellement en 1996, découle partiellement de mes discussions avec certains membres de *Praxiling*. Leur volonté se caractérisait par une ouverture d'esprit manifeste, qui permettait d'accueillir d'autres approches, comme la mienne, entre autres, totalement différente des leurs. Le rapprochement le plus important a eu lieu entre 2011 et 2016, avec deux chercheurs du laboratoire, Catherine Détrie et Bertrand Vérine, avant leur prise de retraite. Je reviendrai sur cela au sein du volet 3 (§ 3.3.3).

Le volet 3 sera l'occasion d'explorer de manière plus approfondie le lien entre *données authentiques*, *analyse du discours électronique médié*, et *mutation(s) des pratiques*.

langage sert à 1) agir dans et sur le réel, ce qui rejoint la pragmatique, et 2) à représenter les points de vue des locuteurs sur le réel et sur l'action pragmatique, ce qui est moins souvent pris en compte, puis 3) à construire l'identité des locuteurs par la façon dont ils nomment les choses, dont ils réagissent aux discours, dont ils racontent les événements et dont ils reçoivent les récits que les autres font des mêmes événements, etc. » (Bertrand Verine, communication personnelle, le 2 août 2016).

3.3.2.8 Encadrement spécifique : volet 2

Les 12 (co)directions de mémoires de Master (M1 et M2) et de rapports de stage, les 2 participations à jurys de soutenance, l'encadrement de 2 stagiaires en formation continue, affectés au METICE pendant ma direction, attestent de cette implication auprès des étudiants pour la période du volet 2.

Tableau 3.6 – Encadrement de la recherche : volet 2*

9 Directions de mémoires	
M2 : « Jeux sérieux dans un contexte pedagogique. Memoire de recherche sur les potentialites educatives des jeux video. », (S. Reverte).	2011-2012
M1 : « Téléphones intelligents et tablettes : usages pédagogiques des logiciels d'apprentissage mobile (Mobile learning software). » (L. Gobert).	2010-2011
M1 : 1. « L'utilisation des TICE dans la rééducation orthophonique : les troubles du langage écrit » (E. Pennes) ; 2. « Tice et dyspraxie/dysgraphie : Les TICE permettent-elles une meilleure inclusion des élèves dyspraxiques/dysgraphiques en milieu scolaire ordinaire ? » (D. Bartholemé)	2009-2010
M1 : 1. « TICE et surdité en milieu scolaire : de la théorie à la réalité » (S. Pollet) ; M2 : « Réseaux sociaux et enjeux pédagogiques » (B. Naoul) ; M1 : « Jeux en ligne et stratégies pédagogiques » (V. Lucas).	2008-2009 2007-2008
Mémoire/rapport de stage de M2, Gestion, apprentissages et formation ouverte et à distance « Conception et réalisation de la mise en ligne d'un cours dans le cadre d'une formation au sein du Département de Français de l'université Pédagogique de Maputo », (N. Bernard).	2004-2005
Maîtrise : « Étude linguistique sur les correcteurs grammaticaux », (C. Tresallet).	1998-1999
3 Co-directions de mémoires	
M1 : « L'identité dans une communauté en ligne, l'exemple du forum rock6070.com » (D. Blind, co-directrice : C. Béal).	2009-2010
DEA : « L'accès à l'information sur Internet » (F. Pascual, co-directeur : P. Siblot). (Volets 1 et 2)	1998-1999
DEA : « Introduction à une étude des difficultés du français et de leur traitement par ordinateur », (M. da Conceição, co-directeur : A. Coïaniz). (Volets 1 et 2)	1996-1997
3 Participations à jurys de soutenance	
M2 Gestion, apprentissages et formation ouverte et à distance : 1. « Stage de production d'un site Intranet, Sufco, université Paul-Valéry Montpellier 3 », (A. Diallo) ; 2. « Mise en place du site Internet Coaching-lombalgie.com » (C. Sesboué).	2005-2006
Maîtrise : « La rupture publicitaire » (N. Llinares, direction : C. Charnet)	1995-1996
2 Stagiaires affectés au Metice, sous ma direction	
Encadrement de deux stagiaires SUFCO (Formation Continue) en DU (diplôme d'université) de concepteur médiatique : 1. Marilyne Martin : « Médiatisation de cours en ligne », septembre et octobre 2001 ; 2. Sabine Cotreaux : « Présences unies vers cités distantes », avril et mai 2001.	2001

*Dans ce tableau ne sont inscrits que les encadrements en rapport avec mes sujets de recherche. J'ai également encadré d'autres mémoires « généralistes », qui ne sont pas indiqués ici. *Cf.* le tableau 3.1 pour ce détail.

Récapitulatif des publications sélectionnées du volet 2

Ce volet inclut davantage de publications à deux ou plusieurs auteurs. Mes co-ateurs relèvent d'une grande variété de disciplines/professions : linguiste-informaticienne, chercheure au *CNRS* (Sophie David) ; conservatrice des bibliothèques au Centre Pompidou, docteure en sciences du langage (Muriel Amar) ; typographe (Gilles Pérez) ; experte en eLearning/FOAD, (Debra Marsh) ; musicienne, compositrice, et *sound designer*, (Lisa Whistlecroft) ; professeure en linguistique anglaise (Laurence Vincent-Durroux) ; professeure en sciences du langage (Chantal Charnet) ; *International Facilitator* à *JISC*, Royaume-Uni, précédemment à *SURF, Collaborative organisation for ICT in Dutch education and research* (Bas Cordewener).
Huit publications attestent de mes implications en recherche fondamentale et appliquée : 1 livre (dont je suis co-auteur) concernant la **formation** des personnels de l'université, *Introduction aux technologies de l'information et de la communication. Problèmes et méthodes : MacOS et Windows*, (Panckhurst et Pérez 2000), 1 acte de colloque (*ICDE*, Düsseldorf), à propos de l'évolution des **outils/pratiques** dans un contexte d'enseignement supérieur et de recherche à l'**ère numérique** (Panckhurst 2001a), 2 co-coordinations de livre : l'un à propos de l'**autoformation** et l'**autoévaluation**, intitulé *Autoformation et autoévaluation : une pédagogie renouvelée ?*, (Vincent-Durroux et Panckhurst 2002), l'autre concernant l'**évaluation de logiciels**, *Evaluation in e-learning : the European Academic Software Award, EASA*, (Panckhurst et al. 2004a) ; 2 chapitres dans ce dernier ouvrage : le premier dresse un bilan du prix *EASA* 2000, (Panckhurst et Cordewener 2004), le second analyse les résultats d'un questionnaire adressé aux candidats de la compétition, (David et Panckhurst 2004) ; 2 actes de colloque : le premier propose une révision de la procédure d'évaluation pour le prix EASA, (*European University Information Systems, EUNIS*, Manchester), (David et al. 2005), et le second évoque des critères/procédures de classification pour l'évaluation de logiciels (*Language Resources and Evaluation, LREC*, Marrakech), (Amar et al. 2008).

Puis, sont évoquées mes recherches à propos des **réseaux d'échanges pédagogiques** (RÉEL/ELEN) dans un contexte de **formation ouverte et à distance** (FOAD/eLearning).
Neuf publications sélectionnées témoignent de cette réflexion : 6 actes de colloque (*Australasian Society for Computers in Learning in Tertiary Education, ascilite*, 2006, Sydney, (Panckhurst et Marsh 2006), et 2012, Wellington, (Panckhurst 2012), *European Association of Distance Teaching Universities, EADTU*, Lisbonne, (Panckhurst et Marsh 2007), *iLearn Forum conference*, Paris (Panckhurst et Marsh 2008a), *Association internationale de pédagogie universitaire, AIPU*, Montpellier, (Panckhurst et Marsh 2008b), *Online Educa*, Berlin, (Panckhurst et Marsh 2009)) ; 1 monographie (20 pages) en ligne en version bilingue anglais-espagnol, revue *RUSC, International Journal of Educational Technology in Higher Education/Revista de Universidad y Sociedad del Conocimiento*, Universitat Oberta de Catalunya (Barcelona), (Panckhurst et Marsh 2011b) ; 1 conférence invitée à la 2e journée TICE à l'UAPV, Avignon, sous forme de vidéo (47 minutes) en ligne, (Panckhurst 2011), https://pod.univ-avignon.fr/video/0108-r-panckhurst-journee-tice-2011/ ; 1 chapitre dans 1 ouvrage à propos des TIC, (Panckhurst et Marsh 2011a).
Enfin, **trois** publications sélectionnées concernant des articles/chapitres **pédagogiques**, 1 hommage à un collègue sur *sens et informatique*, (Panckhurst 1997b), 1 article dans la revue *Apprentissage des langues et systèmes d'information et de communication, ALSIC* concernant les correcteurs grammaticaux et leur usage en situation pédagogique), (Charnet et Panckhurst 1998), puis, au sein de l'ouvrage coordonné par (Détrie et al. 2017), des fiches terminologiques* définissant les notions suivantes : *Discours électronique médié (DEM), Linguistique(s) de corpus, Néographie, Traitement automatique des langues, (TAL)*, (Panckhurst 2001, 2017).

**Remarque* : bien que les fiches DEM et néographie soient plus appropriées au sein du volet 3, pour ne pas démultiplier les citations, je les regroupe toutes à ce stade, sous forme d'une publication unique.

3.3.2.9 Sélection des publications : volet 2

Remarque. — Dans le Volume II, je scinde les publications en trois sections, pour le volet 2 : 1) Formation, évaluation ; 2) Réseaux d'échanges pédagogiques en eLearning, *RÉEL/ELEN* ; 3) Publications pédagogiques. Le lecteur pourra s'y reporter pour ce détail.

Amar, Muriel, Sophie David, Rachel Panckhurst et Lisa Whistlecroft (2008). « Classification procedures for software evaluation ». In : *Actes du colloque LREC*. Marrakech, p. 623–630. URL : www.lrec-conf.org/lrec2008/.

Charnet, Chantal et Rachel Panckhurst (1998). « Le correcteur grammatical : un auxiliaire efficace pour l'enseignant ? Quelques éléments de réflexion ». In : *ALSIC* 1.2, p. 103–114. URL : https://alsic.revues.org/1494.

David, Sophie et Rachel Panckhurst (2004). « Questionnaire results : from the competitors' point of view ». In : *Evaluation in e-learning : the European Academic Software Award*. Sous la dir. de Rachel Panckhurst, Sophie David et Lisa Whistlecroft. MédiaTic 3. Service des publications, Université Paul-Valéry Montpellier 3, p. 83–88.

David, Sophie, Rachel Panckhurst et Lisa Whistlecroft (2005). « Revising the evaluation procedure of the European Academic Software Award, Eunis ». In : *Actes du colloque Eunis*. Manchester. URL : http://web.archive.org/web/20061009022604/http://www.mc.manchester.ac.uk/eunis2005/medialibrary/papers/paper_111.pdf.

Panckhurst, Rachel (2001, 2017). « Discours électronique médié ; Linguistique(s) de corpus ; Néographie ; Traitement automatique des langues ». In : *Termes et concepts pour l'analyse du discours. Une approche praxématique*. Sous la dir. de Catherine Détrie, Paul Siblot, Bertrand Verine et Agnès Steuckardt. Nouvelle édition augmentée. Paris : Honoré Champion, p. 103–105, 205–207, 239–240, 406–408.

– (1997b). « Sens et informatique ». In : *Hommages à Xavier Mignot*. Sous la dir. de Paul Siblot. Université Paul-Valéry Montpellier 3, p. 115–130.

– (2001a). « Distance, open and virtual lifelong learning : shaping the transition within a French University ». In : *Proceedings, 20th World conference on open learning and distance education*. Sous la dir. de Norvège ICDE – Oslo et Allemagne FernUniversität Hagen. Düsseldorf. ISBN : ISBN-NR.3-934093-01-9.

Panckhurst, Rachel (2011). « Réseaux sociaux et pédagogie dans un contexte d'enseignement supérieur français ». In : *2e journée TICE à l'UAPV*. Avignon, Voir la vidéo de l'intervention (47 minutes). URL : https://pod.univ-avignon.fr/video/0108-r-panckhurst-journee-tice-2011/.

– (2012). « The digital tutor : accepting to lose control and make mistakes ». In : *Actes du colloque ascilite "Future challenges/Sustainable challenges"*. Wellington, p. 735–739. URL : http://www.ascilite.org/conferences/Wellington12/2012/images/custom/panckhurst,_rachel_-_the_digital.pdf.

Panckhurst, Rachel et Bas Cordewener (2004). « A review of the European Academic Software Award : Year 2000 ». In : *Evaluation in e-learning : the European Academic Software Award*. Sous la dir. de Rachel Panckhurst, Sophie David et Lisa Whistlecroft. MédiaTic 3. PULM, Université Paul-Valéry Montpellier 3, p. 23–41.

Panckhurst, Rachel, Sophie David et Lisa Whistlecroft, éds. (2004a). *Evaluation in e-learning : the European Academic Software Award*. MédiaTic 3. Service des publications, Université Paul-Valéry Montpellier 3.

– (2004b). « Overview, Sommaire ». In : *Evaluation in e-learning : the European Academic Software Award*. Sous la dir. de Rachel Panckhurst, Sophie David et Lisa Whistlecroft. MédiaTic 3. Service des publications, Université Paul-Valéry Montpellier 3, xi–xiv and xv–xviii.

Panckhurst, Rachel et Debra Marsh (2006). « A French Master's degree in eLearning : are the students' needs met ? » In : *Actes du colloque ascilite*. Sydney, p. 985–986. URL : http://www.ascilite.org/conferences/sydney06/proceeding/pdf_papers/p25.pdf.

– (2007). « eLEN — eLearning Exchange Networks : reaching out to effective bilingual and multicultural University collaboration ». In : *Actes du colloque EADTU*. Lisbon.

– (2008a). « Communities of Practice. Moving from Institutional Platforms to the open Web as a platform ». In : *Actes du colloque iLearn Forum*. Paris. URL : http://www.eife-l.org/publications/proceedings/ilf08/contributions/designing-estrategies-for-learningorganisations/panckhurst/_marsh.pdf/view.

– (2008b). « REEL : réseaux d'échanges pédagogiques en eLearning. Améliorer la qualité de l'apprentissage en favorisant l'autonomie des apprenants ». In :

25e Congrès de l'AIPU, Le défi de la qualité dans l'enseignement supérieur : vers un changement de paradigme. Sous la dir. de Chantal CHARNET, Claire GHERSI et Jean-Louis MONINO. Montpellier, Actes consultables en ligne. URL : www.aipu2008-montpellier.fr.

– (2009). « eLEN2 — 2nd generation eLearning Exchange Networks ». In : *Actes du colloque Online Educa*. Berlin, p. 245–248. URL : http://www.online-educa.com.
– (2011a). « Les frontières pédagogiques sont-elles remises en question par l'utilisation des réseaux sociaux ? L'implémentation d'objets d'apprentissage sociaux dans un espace de communication électronique médiée ». In : *La communication électronique : enjeux de langues*. Sous la dir. de Fabien LIÉNARD et Sami ZLITNI. Lambert-Lucas, Limoges, p. 293–301.
– (2011b). « Using Social Networks for Pedagogical Practice in French Higher Education : Educator and Learner Perspectives ». In : *RUSC, Revista de Universidad y Sociedad del Conocimiento, « Globalisation and Internationalisation of Higher Education »* 8.1, Monographie en ligne. ISSN : 1698-580X. URL : http://www.raco.cat/index.php/RUSC/article/view/225632/306988.

PANCKHURST, Rachel et Gilles PÉREZ (2000). *Introduction aux technologies de l'information et de la communication. Problèmes et méthodes : MacOS et Windows*. MédiaTic 1. Service des publications, Université Paul-Valéry Montpellier 3.

VINCENT-DURROUX, Laurence et Rachel PANCKHURST, éds. (2002). *Autoformation et autoévaluation : une pédagogie renouvelée ?* MédiaTic 2. Service des publications, Université Paul-Valéry Montpellier 3.

3.3.3 Volet 3 : Communication médiée par ordinateur (CMO), discours électronique médié (DEM), discours numérique médié (DNM) (1996-2017)

Les volets 2 et 3 débutent la même année, en 1996. Le volet 3 se poursuit au-delà de la fin du volet 2 (2012), et englobe une recherche s'étendant sur plus de deux décennies, fortement imprégnée du recueil et de l'exploitation de données authentiques. Dans un premier temps, j'ai collecté des données émanant de courriels, puis dans un second temps, de forums et de messageries instantanées (ou chats), le tout dans un contexte exclusif d'enseignement supérieur et de recherche. Puis, j'ai élargi le contexte de collecte auprès du grand public, dans le cadre du projet SMS *sud4science LR* puis du projet Délégation générale à la langue française et aux langues de France (DGLFLF), en embarquant dans l'aventure avec moi, 5 collègues et 8 stagiaires étudiants.

Je dirais que ce troisième volet constitue le plus important à la fois en matière de périodicité et d'orientations de recherche, peut-être parce qu'il m'a permis, avec le recul, de comprendre la nature de la recherche qui me passionne le plus : la recherche collaborative et interdisciplinaire, qui jongle entre recherche fondamentale, descriptive, analytique, applicative, autour d'un objet commun — par exemple, un corpus. De plus est, toujours dans un contexte de service public, la dimension de mise à disposition des résultats, de diffusion d'une recherche réellement accessible, et non pas hermétique (!), auprès du grand public — d'un retour sur investissement des usagers (par exemple, après nous avoir fait don de leurs SMS) — m'importe énormément. La recherche publique doit non seulement revenir (et rendre des comptes) au grand public, mais — et là, nous avons tous des efforts à faire pour convaincre — elle doit être davantage directement prise en considération au niveau ministériel. Je reviendrai sur ce dernier point dans la section § 3.4 Réseaux, diffusion et valorisation (*cf.* § 3.4.5.1).

Le volet 3 peut être lui-même scindé en deux périodes : la première, s'échelonnant de 1996 à 2006, caractérisée par les définitions et les premières analyses dans le cadre de la CMO, (Panckhurst 1997a)), à partir de courriels (Panckhurst 1998a,c, 1999a,b ; Panckhurst et Bouguerra 2003), forums et chats (Panckhurst 2003a, 2006b) ; la seconde, menée de 2006 à 2017, concernant une ré-orientation terminologique, *discours électronique médié, discours numérique*

médié (DEM/DNM, Panckhurst 2006a, Panckhurst 2007), et englobant l'élaboration de ma typologie concernant l'*écriture SMS* en français (eSMS, Panckhurst 2009), et en comparaison avec l'espagnol et l'italien (Panckhurst 2010), le recueil et l'analyse des données SMS (Lopez et al. 2015 ; Panckhurst 2013, 2015 ; Panckhurst et al. 2013 ; Panckhurst et Moïse 2014 ; Panckhurst et al. 2015b, 2016b ; Roche et al. 2016), et la mise à disposition du corpus *88milSMS* et des réponses à un questionnaire sociolinguistique associé (Panckhurst et al. 2014a), http://88milsms.huma-num.fr/, puis des recherches sur les applications possibles en TAL (Accorsi et al. 2014 ; Lopez et al. 2014, 2015).

Si le premier volet était plutôt consacré à ma propre élaboration d'outils/prototypes en TAL, et le second volet, à l'utilisation d'outils par autrui, (puis la réflexion et la recherche autour de ces usages), dans ce troisième volet, je reviens à l'implémentation informatique, mais elle est désormais envisagée uniquement en collaboration avec des informaticiens. Mon rôle de linguiste-informaticienne se situe maintenant dans une dimension de dialogue fructueux avec eux. Si j'ai programmé toute seule autrefois, je ne le fais plus désormais. Cette mutation est sans doute également liée à l'évolution de mes enseignements : je ne forme plus des linguistes capables de programmer, mais plutôt des utilisateurs avertis de logiciels, entre autres, mais pas seulement, de TAL (*cf*. volet 1, § 3.3.1). Selon moi, la meilleure façon pour apprendre est de devoir enseigner. D'une part, comme je n'enseigne plus la programmation, et que je n'ai donc plus à pratiquer cette dimension au quotidien, et, d'autre part, comme les informaticiens sont *a fortiori* plus à même de suivre les évolutions techniques en matière de programmation, je préfère être désormais positionnée dans ce rôle de dialogue, située quelque part *entre linguistique et informatique*.

Dans ce troisième volet, les *outils de traitement automatique du langage naturel écrit (TALNE)* englobent l'*analyse du discours électronique/numérique médié (DEM/DNM)*. Ce travail de recherche permet, me semble-t-il, d'établir un lien direct entre mes recherches précédentes fondamentales en traitement automatique du langage (volet 1, § 3.3.1) et le domaine des TICE et de la FOAD (volet 2, § 3.3.2). Les données deviennent authentiques, recueillies soit dans un contexte d'enseignement supérieur et de recherche, soit auprès du grand public, et les évolutions, les mutations langagières, notamment, des pratiques scripturales, sont explorées et analysées.

3.3.3.1 Débats terminologiques : CMO

Revenons à 1996, année de parution de l'ouvrage coordonné par Susan Herring, *Computer-Mediated Communication* (Herring 1996), qu'elle définit ainsi :

> Computer-mediated communication (CMC) is communication that takes place between human beings via the instrumentality of computers.
>
> (Herring 1996, p. 1)

Cette même année, je démarrais les formations pour les personnels de l'université (*cf*. Volet 2, § 3.3.2), et je fournissais également des adresses électroniques à mes étudiants. Grâce au contenu des messages estudiantins reçus et ceux émanant des enseignants participant à mes formations, j'avais accès à des données électroniques « nativement numériques » et je pouvais donc observer les pratiques des usagers qui communiquaient par ce biais. Je résume ces premières pratiques voire, pour certains, ces réticences, et les raisons qui m'ont amenée à m'intéresser à ce sujet de recherche, dans un article bilan (Panckhurst 2006a) :

> Les raisons du choix de cette recherche sont triples :
>
> 1. Au début de ces recherches, en 1996, les étudiants ne disposaient pas d'adresses électroniques dans le cadre de leurs études ; nous leur avons donc fourni cette possibilité et nous avons par la suite rapidement constaté qu'ils se servaient de l'outil pour communiquer avec leur enseignante dans un contexte périphérique à celui du cours [64] :
>
>   J'aurais préféré m'adresser à vous de vive voix pour vous présenter mes excuses, mais constatant que vous étiez très demandée par les élèves en fin de cours, je n'ai donc pas oser insister. Aussi, je profite de ce mode de communication pour vous informer de ma situation actuelle…J'élève seule ma fille de deux ans et demi et j'ai beaucoup de difficultés à la réveiller tôt en ce moment, ceci expliquant mes retards systématiques aux T.D. du mardi matin. (étudiant, corpus 1996).
>
>   Si je vous écris, c'est parce que je préfère ne pas être en face de vous pour vous entendre me répondre définitivement « non » ; au moins le courrier électronique m'aura donné « cet avantage »…je ne sais même pas pourquoi je continue à espérer que pour 0,5 points les choses pourraient changer. (étudiant, corpus 1997).

64. Les fautes éventuelles contenues dans les messages apparaissent telles quelles.

2. En parallèle, nous assurions des formations de formateurs enseignants, pour les TICE ; certains collègues étaient réticents et nous voulions comprendre pourquoi :

   Je n'arrive pas à m'habituer au courrier électronique. C'est un moyen de communication qui me paraît distant, froid. (enseignant, corpus 1997).

   J'espère que ce moyen de communication ne remplacera jamais la communication inter-personnelle. (enseignant, corpus 1997).

3. L'accès aux documents électroniques était quasi-immédiat ; même si les recherches en traitement automatique des langues sur des vastes corpus foisonnaient déjà, l'idée de pouvoir disposer de manière quasi-immédiate de corpus de taille raisonnable était très séduisante.

   Enfin, le choix du contexte de l'enseignement supérieur français, n'était pas uniquement dû à l'accès simplifié aux données textuelles. Nous voulions étudier l'évolution du langage, et nous pensions — à tort, à cette époque — que les discours contenus dans ces messages électroniques ne seraient pas fondamentalement différents d'autres moyens écrits plus classiques. Nous avons très vite observé que ce n'était pas le cas. (Panckhurst 2006a, p. 349-350).

Ces pratiques m'interpellaient et je voulais en savoir plus. Assez rapidement, j'ai constaté que l'utilisation de l'ordinateur semblait provoquer un changement comportemental des usagers. Dès 1996, j'ai donc décidé d'orienter mes recherches dans cette direction. (Herring 1996) avait constaté que le choix d'étudier l'évolution du langage dans le cadre d'espaces communicationnels contemporains n'avait rien d'évident et que la recherche dans ce domaine en était à ses premiers balbutiements :

> Surprisingly, although text-based CMC is constructed almost exclusively from linguistic signs, linguists have been slow to consider computer-mediated language a legitimate object of inquiry. [...] Research on CMC is still in its infancy.
> (Herring 1996, p. 3, p. 5).

Les variations comportementales lorsque l'ordinateur (puis, plus tard, le téléphone) est utilisé dans un cadre de communication interpersonnelle, existaient bel et bien. Je reviendrai sur ces caractéristiques (§ 3.3.3.2). Mais tout d'abord, je voulais proposer une expression en français. Est né, en 1997, mon néologisme

pour le français, la *communication médiée par ordinateur*, suite à (Herring 1996), et non la *communication médiatisée par ordinateur* (Panckhurst 1997a).

> En français, le verbe *médier* ne semble pas exister (officiellement) [...] ; cela étant, la morphologie dérivationnelle et flexionnelle semble le permettre a partir du nom *médiation*, qui, lui, existe. Mais pourquoi ne pas choisir *mediatiser* (en rapprochement avec *médiatisation*), verbe officiellement reconnu, au sens de : « diffuser par les médias » (*Petit Robert*) ? Rappelons un des sens de *diffuser* : « répandre dans toutes les directions » (*Petit Robert*). Il semble que le verbe *médiatiser* soit précisément trop connote en direction des médias pour convenir dans le cadre d'un échange de courrier électronique entre (le plus souvent) deux personnes [...].
> Par ailleurs, en italien et en anglais, par exemple, le verbe existe : respectivement *mediare* (emploi rare) et *mediate* (du latin *mediatus*). En français, le nom *médiation*, et l'adjectif/nom *médian*, *médiane* existent, ainsi que le verbe *remédier* a partir du latin *remediare* (mais ce dernier ne peut être retenu, puisqu'il constitue une racine distincte).
> Si le verbe *médier* n'est pas officiellement reconnu, il est d'ores et déjà utilise sur le réseau Internet [...]. Je propose donc que soit adopte le verbe néologique *médier* en français, et ainsi l'expression en néologie terminologique : *la communication médiée par ordinateur*. L'acronyme serait alors CMO.
>
> (Panckhurst 1997a, p. 57).

Mon acronyme néologique communication *médiée* par ordinateur est bien moins utilisée que communication *médiatisée* par ordinateur en français [65], et les appellations pour décrire ce type de communication foisonnent, comme je le signalais dans (Panckhurst 2009) [66] :

65. Recherche Google Scholar, 10/1/2017 « Communication médiée par ordinateur » (417 résultats), « Communication médiatisée par ordinateur » (1 170 résultats).

66. Cette même année, Vold Lexander, proposait *communication mediatisee par les technologies de l'information et de la communication* (CMTIC,). Mais, comme Cougnon, je trouve cette appellation trop imprécise (Cougnon 2015, p. 18, note 3). (Cougnon 2015, p. 3), tout en acceptant l'appellation CMO, apporte une précision supplémentaire en choisissant de cibler la dimension écrite : *communication écrite médiée par ordinateur* (Cémo). On peut également ajouter, de manière non-exhaustive, d'autres appellations de chercheurs français : « communication médiée par les réseaux », CoMéRé, (Chanier et al. 2014), « communication médiatisée par les technologies de l'information et de la communication », CMT, (Panckhurst 2007), « écriture électronique » « communication électronique écrite », « écrilecte », (Liénard 2007, 2012), « discours numérique », « écriture numérique » (Marcoccia 2016), « textualité numérique » (Panckhurst *et al*. Rapport

> D'autres chercheurs proposent des variantes terminologiques pour décrire les phénomènes d'analyse linguistico-informatique de ces types de discours. Entre autres : « communication médiatisée par ordinateur » (Marcoccia 2000), « communication électronique scripturale » (Anis et al. 2004), « nouvelles formes de communication écrite » (Guimier de Neef et Véronis 2004), « communication électronique » (Anis et al. 2004) et, pour l'anglais : « Netspeak » (Crystal 2001), « computer-mediated communication » (Herring 1996). Les termes retenus dans un contexte (journalistique) plus vaste (comprenant également la publication électronique) incluent notamment : « cyberl@ngue », « cyberlangage »
>
> (Dejond, 2002, 2006). ((Panckhurst 2009), p. 34).

Figure 3.8 – Terminologie CMO/DEM, 2009

En revanche, mon appellation a été parfois préférée à d'autres, entre autres, par (Anis 1999) [67] et (Develotte et al. 2011) :

> Dans le domaine de la terminologie savante, nous adoptons le terme de *Communication Médiée par Ordinateur*, **proposé par un des membres de notre équipe. (Panckhurst 1997a) écrit pour justifier cette option :**

scientifique DGLFLF, 2013), « technologie discursive » « discours (socio)numérique » (Paveau 2013), « français tchaté » (Pierozak 2007), etc.

67. (Anis 1999) a décidé d'intituler la première partie de son ouvrage, qui en compte trois, *Communication médiée par ordinateur*.

L'utilisation de la machine modifie notre discours et ainsi notre façon de communiquer avec autrui. Je pense que le verbe néologique *médier* serait plus approprié que celui qui existe en français, *médiatiser*, car la communication par ordinateur est véritablement « médiée » (au sens de la médiation de Vygotsky), et non pas simplement « médiatisée ». [...]

(Panckhurst 1997a, p. 56), in (Panckhurst 1999a, p. 9), je souligne.

L'émergence récente des formes de communication médiées par ordinateur a généré un lexique abondant et non stabilisé pour l'heure. Les désignations sont multiples, d'une part, pour rendre compte de modes de communication différenciés mais, également, pour renvoyer au même référent sous des termes différents [...]. [U]ne précision terminologique s'impose. Contrairement à l'anglais qui ne connaît que le terme *mediation*, le français dispose de deux mots que leur proximité de forme a placés au cœur d'un numéro de Notions en questions en 2003 : « Médiation, médiatisation et apprentissages ». Dans ce numéro, Jacquinot-Delaunay, au terme d'un développement étymologique détaillé (p. 128), explique que *médiation* renvoie au fait de servir d'intermédiaire, avec une notion implicite d'acteur humain, alors que *médiatisation* évoque un intermédiaire technologique. On aboutirait ainsi à une opposition entre médiation humaine et médiation par la technologie (cette dernière appelée aussi médiatisation ; par exemple une conversation en présentiel sera médiatisée lorsqu'elle sera transformée en un enregistrement). Or le passage en ligne introduit la technologie non plus seulement comme support mais comme acteur de l'échange (Hewling 2009 ; Mondada 2007), ce qui amène à revisiter la portée du terme médiation. Par exemple si une conversation peut être dite médiatisée parce qu'elle est matérialisée sur un support appelé *forum*, elle est par ailleurs au centre d'une médiation. En effet, la technologie des forums permettant des visualisations par chronologie, par auteur ou par thème, c'est la modalité de visualisation choisie par l'utilisateur humain qui va conditionner la façon dont les contenus du forum vont faire sens pour lui. La technologie n'est plus ici le simple support d'une médiatisation, elle est aussi l'agent d'une médiation. **Les coordonnateurs font donc le choix d'adopter, à la suite de (Panckhurst 1997a, 1999a), la désignation de communication médiée par ordinateur et non médiatisée** (lexème néanmoins privilégié par d'autres auteurs dans ce livre, comme par exemple dans la contribution de Michel Marcoccia).

(Develotte et al. 2011, p. 9-10), je souligne).

D'autres chercheurs (Véronis et Guimier de Neef 2006) mentionnent mon néologisme :

> Reconnaissant cette difficulté, **(Panckhurst 1997a) propose le néologisme communication médiée par ordinateur**. Ce néologisme est d'ailleurs utilise indépendamment dans d'autres sciences, par exemple en biologie (« La tolérance immunologique médiée par une molécule », etc.).
>
> (Véronis et Guimier de Neef 2006, p. 3) in (Sabah 2006), je souligne).

mais ne l'adoptent pas, préférant une autre appellation : *nouvelles formes de communication ecrite* (NFCE). Celle-ci me gêne quelque peu dans la mesure où les formes décrites ne demeurent jamais nouvelles très longtemps (*cf*. l'exemple des NTIC — *nouvelles technologies de l'information et de la communication* — ensuite devenues *technologies de l'information et de la communication (éducatives)*, TIC(E)).

3.3.3.2 DEM/DNM

Dès (Panckhurst 1998a, p. 47-48, p. 52, p. 56-58) (Panckhurst 1999a, p. 63) [68] puis de manière plus appuyée quelques années plus tard, (Panckhurst 2006a,b) j'ai préféré le terme *discours électronique médié* (DEM). Le terme est en quelque sorte devenu ma *marque de fabrique* pour le français — en effet, j'étais la première à pointer les spécificités du DEM, et ce dès 1998-1999. Dans un article bilan (Panckhurst 2006a), je rappelle le terme et j'en décris ses marques spécifiques.

> Quand l'ordinateur est utilisé pour le courriel, les forums de discussion et les chats, en tant qu'outil permettant la communication entre individus, il devient un véritable *médiateur ;* son utilisation modifie notre discours et ainsi notre façon de communiquer avec autrui. Émerge alors un nouveau « genre de discours », le « discours électronique médié (DEM) ». Le DEM contient des marques linguistiques et extra-linguistiques qui lui sont propres et il entre dans le cadre plus global de la « communication médiée par ordinateur » (CMO).
>
> (Panckhurst 2006a, p. 345).

68. J'ai également utilisé *discours médié par ordinateur* (Panckhurst 1999b, p. 317), mais pour éviter une confusion terminologique, j'ai préféré rapidement trancher en CMO et DEM.

Je trouvais que l'appellation *communication* médiée par ordinateur était trop vaste pour mon sujet d'étude, m'intéressant plus précisément au *discours* présent au sein des espaces communicationnels. J'avais déjà décidé de cibler un sous-ensemble de la CMO, à mon sens, le DEM (Panckhurst 1998a, 1999a). Cela rejoint également (Herring 1997, 2001) qui adopte l'appellation *Computer-mediated discourse* pour l'anglais :

> *Computer-mediated discourse* is the communication produced when human beings interact with one another by transmitting messages via networked computers. The study of computer-mediated discourse (henceforth CMD) is a specialization within the broader interdisciplinary study of computer-mediated communication (CMC), distinguished by its focus on *language and language use* in computer networked environments, and by its use of methods of *discourse analysis* to address that focus. (Herring 2001, p. 612), in (Schiffrin et al. 2001)

En revanche, au contraire de Herring, j'estimais que l'inclusion du mot *ordinateur* dans l'intitulé pouvait paraître quasi-obsolète, tout au moins dans le cadre d'une recherche bientôt menée, pour ma part, sur les SMS, donc à partir de téléphones mobiles. Je justifiais mon évolution de la manière suivante :

> Notre terme, CMO [...] pourrait paraître désormais quasi-obsolète, surtout si on souhaite élargir l'analyse des discours au-delà des trois moyens de communication asynchrone et (quasi) synchrone utilisés couramment, à savoir : le courriel, le forum et le chat, afin d'inclure, par exemple, les messages courts (SMS). Cependant, même si nous dirigeons actuellement des travaux de recherche étudiants sur les SMS (français-bulgare et français-russe), nos propres recherches se situent exclusivement dans un cadre d'enseignement supérieur et de recherche, et, de ce fait, seuls les SMS qui apparaîtraient dans le cadre des outils utilisés par ordinateur nous interpellent [69]. Depuis très peu de temps, tout au moins dans le contexte de nos corpus recueillis depuis une décennie, les SMS font leur apparition à la fois dans les chats et les forums universitaires ; c'est encore très rare de les rencontrer dans le cadre de courriels entre enseignants et étudiants. Mais, en tout état de cause, il est plus approprié d'utiliser le terme « discours électronique médié »

69. À l'époque, je n'avais pas encore proposé le terme eSMS pour *écriture SMS* (Panckhurst 2009). L'élargissement du champ de recueil des données incluant celles émanant du grand public n'a commencé que 4 ans plus tard, en 2011.

> (DEM), qui constitue un sous-ensemble de la CMO, pour les types d'analyses linguistico-communicationnelles que nous menons.
>
> (Panckhurst 2007, p. 121-122).

D'autres chercheurs m'ont suivie avant et après cette mutation terminologique : (Bevilacqua 2012) cite CMO, (Lazar 2012, 2013, 2017), (Poirier 2011, p. 45) utilisent DEM, incluant parfois jusque dans l'intitulé de leurs articles, le terme DEM (Lazar 2012, 2013, 2014) :

> D'après nous, le terme le plus précis nous est apporté par (Panckhurst 1998a, 1999a), *discours électronique médié*, qui met davantage l'accent sur la notion d'une analyse du discours électronique médié apparaissant au sein de ces espaces communicationnels. (Lazar 2013).

(Cougnon 2015, p. 18, note 3) estime de son côté que DEM est trop precis, justement, à cause du fait que le terme cible uniquement le discours.

Resterait donc la dichotomie *communication / discours*. S'il est vrai qu'un flottement existe entre *communication médiée par ordinateur, communication(s) électronique(s), discours électronique, discours médié, discours électronique médié*, dans mes premiers écrits (Panckhurst 1998a,c, 1999a,b ; Panckhurst et Bouguerra 2003), à partir de 2006, j'assume le choix de DEM et les titres des publications l'attestent (Panckhurst 2006a,b, 2007). En tout état de cause, c'est cette dernière appellation que les autres chercheurs retiennent, lorsqu'ils me citent :

> On peut tout d'abord observer l'utilisation d'une très grande variété de dénominations pour désigner la communication numérique, sans doute parce que le champ de recherche est encore assez jeune et que tout n'y est pas stabilisé. [...] Murray utilise « electronic communication », **Panckhurst préfère « discours électronique médié »**, Anis adopte « communication électronique scripturale », et Marcoccia préfère « écriture numérique » [...] (Marcoccia 2016, p. 21), je souligne).

Je ne vois nullement un inconvénient à utiliser l'un ou l'autre aujourd'hui, le premier se situant en effet dans un contexte disciplinaire plus large, le second étant plus ciblé :

> Le DEM contient des marques linguistiques et extra-linguistiques qui lui sont propres et il entre dans le cadre plus global de la CMO.
>
> (Panckhurst 2006a, p. 345)).

Tout dépend du positionnement et du degré d'interdisciplinarité recherché, et des points de rencontre qui conviennent à l'ensemble des chercheurs travaillant dans un projet.

Dans son ouvrage, qui offre un excellent panorama du champ disciplinaire pour un public estudiantin, (Marcoccia 2016) propose d'incorporer un terme en vogue : « numérique » [70]. Il a également évolué du point de vue terminologique, préférant opter désormais pour la « communication numérique écrite » (au lieu de « communication (écrite) médiatisée par ordinateur » qu'il utilisait précédemment). Il définit cette appellation ainsi :

> La communication numérique renvoie à toute forme d'échange communicatif dont les messages sont véhiculés par des réseaux télématiques, c'est-à-dire basés sur la combinaison de l'informatique et des télécommunications, du minitel à la téléphonie mobile, en passant par l'internet. La communication numérique est donc le terme générique englobant divers types de situations de communication interpersonnelle (privée ou publique) par courrier électronique, messagerie instantanée, forums, tchats, plateformes de réseaux sociaux, etc.
>
> (Marcoccia 2016, p. 16).

À vrai dire, c'est plutôt qu'il situe la *communication numérique écrite* au sein d'un sous-ensemble, constitué par la *communication médiée/médiatisée par ordinateur* (CMO) ou la *communication médiée/médiatisée par téléphone* (CMT) (Marcoccia 2016, p. 16), reprenant ainsi la distinction — qui s'estompe de plus en plus entre moyens techniques utilisés — de (Liénard 2012) :

> Les ordinateurs (favorisant la CMO) sont de plus en plus portables, mobiles rendant les usages proches de ceux produits en CMT et [...] les téléphones (permettant la CMT) se rapprochent de plus en plus des ordinateurs transformant les usages (qui se rapprochent de la CMO). (Liénard 2012, p. 145).

70. Cependant, son utilisation du terme « numérique » ne date pas de 2016, mais figure dans des publications plus anciennes, par exemple, (Marcoccia 2012). D'autres travaux, notamment en anglais, incorporaient l'équivalent : *Digital discourse* (Thurlow et Mroczek 2011).

Par ailleurs, (Cougnon 2015) justifie le maintien du terme CMO (ou plutôt, Cémo — *communication écrite médiée par ordinateur*, pour ce qui la concerne), même pour les recherches ciblant les données émanant de la téléphonie mobile :

> Pour ce qui est de l'exclusion du sms du domaine de l'informatique, le téléphone portable, y compris lors de l'emploi de sms, est une technologie qui présente des composantes électroniques, permettant des fonctions semblables à celles des ordinateurs ; le téléphone portable de base est en quelque sorte un mini ordinateur qui permet la rédaction et l'envoi du sms, la suggestion orthographique, l'introduction de signes de ponctuation, voire meme de smileys-images. Nous pensons donc, a la suite de Bieswanger (2007), que l'ordinateur en est bien le medium. (Cougnon 2015, p. 19).

La terminologie continue et continuera à évoluer au gré des pratiques et des usages technologiques. Se situant dans le champ des sciences de l'Information et de la communication, il est naturel que Marcoccia utilise *communication* dans l'intitulé de son ouvrage. Par ailleurs, comme (Cougnon 2015), il précise la dimension *écrite* de ses recherches. Pour ma part, bien que je travaille également sur des *données saisies à l'aide d'un clavier*, donc plutôt sur de l'écrit, par opposition à l'oral, je préfère ne pas introduire le terme au sein même de l'acronyme, pour ne pas exclure des aspects sémiologiques, de type emoji. Puis, si *numérique* peut en effet remplacer *électronique* — je me laisse alors volontiers embarquer par l'évolution terminologique, en m'alignant sur les pratiques actuelles [71] — je trouve important de maintenir la dimension de *médiation*. Je propose alors ma propre appellation évoluée : *discours numérique médié* (DNM), qui est reflétée dans le titre même de cette habilitation.

Maintenant, comment définir les caractéristiques du DEM/DNM [72] ? Je ne reprends pas l'ensemble de mes publications, mais je proposerai plutôt un condensé ci-dessous. Les seules incursions supplémentaires que je me permettrai seront faites afin d'évoquer des points précis qui permettent de mieux comprendre ma démarche.

71. Même si l'effet de mode n'a pas eu d'impact sur l'appellation « courrier électronique » qui aurait muté en « courrier numérique », ou « electronic mail » en « digital mail ».

72. Je maintiendrai parfois l'appellation DEM dans le contexte de ma synthèse pour l'habilitation, ici, puisque je réfère constamment à des publications précédentes employant ce terme. Pour toute publication ultérieure à 2017, j'emploierai DNM.

Je rappelle tout d'abord ma classification du DEM. En effet, j'ai été la première à pointer les spécificités du DEM, dès 2006 (Panckhurst 2006a), suite à l'évolution terminologique depuis CMO (Panckhurst 1997a). *Classification*

> Le discours électronique médié, au sens large (incluant donc courriels, forums, chats, blogs, dans lesquels apparaissent ou non des eSMS [73]), se caractérise par un ensemble de phénomènes qui nous rappelons ci-dessous (*cf.* (Panckhurst 2006a, 2007), (Véronis et Guimier de Neef 2006) in (Sabah 2006)) :
>
> 1. les *didascalies électroniques* (*cf.* (Mourlhon-Dallies et Colin 1999) in (Panckhurst 1999a)), par exemple, les *binettes* (ou *smileys*, [emoji] [74]) permettant d'introduire des aspects sémiologiques non-verbaux ;) ^^ 💗 😊😉😜😰 ; les signes typographiques spécifiques : majuscules (exprimant colère ou agacement ou encore angoisse : *C'est pourquoi je vous lance un S.O.S. HELP ME !!!!!*), allongement, répétition de caractères (permettant de simuler, dans certains cas, la communication para-verbale : *sssssuuuuupppppeeeerrrr !!!!!*) ; l'usage des chevrons « > » ou barres verticales « | » (permettant une répétition du discours entre l'énonciateur et son destinataire, dans le cas du courriel) ;
> 2. les erreurs ou « ratages » orthographiques, typographiques et grammaticaux et l'absence (ou la diminution) de ponctuation (*cf.* (Panckhurst 1998c), (Véronis et Guimier de Neef 2006) in (Sabah 2006), pour une synthèse) ; (Véronis 1988) distingue les erreurs de *compétence* et de *performance* ; dans (Panckhurst 1998c) nous distinguons les erreurs *machinales*, *discriminantes* (probablement engendrées uniquement par l'utilisation de l'ordinateur) des erreurs *floues*, *non-discriminantes* (plus difficiles à déterminer, provenant soit d'une méconnaissance de règle, soit d'une erreur due au moyen utilisé, par exemple) ;
> 3. la *néologie* et la *néographie* [75] (*cf.* (Véronis et Guimier de Neef 2006), in (Sabah 2006)), entre autres, par exemple, les emprunts linguistiques et les abréviations, troncations, notations *sémio-phonologiques* (*cf.* (Liénard 2005, 2007) ou graphies phonétisées typiques des SMS.
>
> (Compilation de (Panckhurst 2006a, 2009).

73. eSMS=écriture SMS. Je développe ce point plus bas.
74. À l'époque de la rédaction de cet article, les emoji n'étaient ni disponibles pour les ordinateurs ni pour les téléphones portables en Europe. Je reviendrai sur ce point (§ 3.4).
75. Comme (Anis 1998), je désigne par néographie des variantes de graphie qui s'éloignent de la langue standardisée, souvent de manière délibérée, ludique, et qui sont très présentes et instables dans l'écriture SMS.

Écrit/oral

Le DEM/DNM peut être parfois comparé à de l'**oral** :

> Le message écrit est quasi dépourvu de marqueurs para-verbaux et non-verbaux (intonation, regard, proximité, posture, gestualité, etc.) qui permettent, dans une situation d'échange verbal conversationnel, une certaine régulation. Tout naturellement, en situation de CMO, les internautes essaient de reproduire ces fonctionnalités. Certaines utilisations rapprochent, en effet, le DEM/DNM d'une situation de communication orale. (Panckhurst 2006a, p. 350).

En parallèle, les **marques écrites** apparaissent fréquemment dans le DEM/DNM et semblent le situer du côté de l'**écrit** (des écrits) aussi :

> De manière générale, le DEM/DNM semble ressembler à d'autres formes de l'écrit, car certaines marques sont omniprésentes :
>
> — l'usage des noms est plus important que l'usage des verbes ;
> — les formes interrogatives sont majoritairement des formes classiques qui excluent l'intonation ;
> — la négation « ne…pas » est tenace.
>
> (Pour plus de détails, Panckhurst 1998c, 1999a,b Panckhurst 2006a, p. 351.)

(Gadet 1996, p. 17-19) rappelle les divergences entre oral et écrit que j'ai résumées dans (Panckhurst 1999a, p. 63) :

— hétérogénéité, variabilité de l'oral (face à une relative homogénéité de l'écrit : codifié, fixé, stabilisé, normé) ;
— « scories » : l'oral reste « truffé de pauses, hésitations, reprises, recherches de mots, incomplétudes, redites, anticipations, auto-interruptions… » ;
— intonation (et les facteurs prosodiques).

Dans une analyse portant sur l'anglais, (Halliday 1989, p. 61-62) opposait la « densité lexicale » très élevée à l'écrit, de la « densité grammaticale », privilégiée à l'oral. (Gadet 1996) appliquait cette analyse au français et constatait « une préférence de l'oral pour les verbes, et de l'écrit pour les noms » (Gadet 1996, p. 23), in (Panckhurst 1999a, p. 66).

Certains chercheurs, entre autres (Véronis et Guimier de Neef 2006), Marcoccia in (Pierozak 2007), réfutent le débat écrit/oral, voire l'opposition éventuelle entre les deux, arguant, pour leur part, que les fréquences/marques dépendent de la situation de communication, des registres nécessaires à l'échange, par exemple, un niveau de langue spécifique, en fonction du destinataire d'un message. S'il est indéniable que ces phénomènes ont une influence sur les échanges (entre autres) numériques, je maintiens que dans certains cas, il est possible de plutôt situer ceux-ci du côté de l'écrit ou de l'oral, en fonction de l'outil de communication utilisé (Panckhurst 2006b). Autrement dit, choisir un outil de communication précis, notamment en situation pédagogique, peut avoir certaines conséquences pour les usagers. Je reviendrai sur ce point *infra*.

(Anis 1998, p. 122) évoque l'écrit communicationnel comme étant « un hybride entre l'écrit et l'oral » in (Panckhurst 1999a, p. 64). Non seulement le DEM/DNM peut tantôt ressembler à l'écrit tantôt à l'oral, ou se situer *somewhere between*, il peut, selon moi, également contenir **ses propres spécificités** :

Marques spécifiques

> Des marques plus spécifiquement syntaxiques sont apparentes également :
>
> (a) une utilisation prédominante du présent de l'indicatif ;
>
> (b) une utilisation importante de déictiques, essentiellement des pronoms à la 1re et à la 2e personne ;
>
> (c) un pourcentage moins élevé de verbes par rapport à d'autres formes de l'écrit plus normé ;
>
> (d) l'« ellipse » (ou la « non-expression », terme préféré par certains linguistes) fréquente de certains morphèmes (*ne* par exemple) ou de formes pronominales, entre autres.
>
> (Panckhurst 2009, 2016b)

Puis sur un plan *extra-linguistique* — ou peut-être dirais-je plus facilement aujourd'hui — sur un plan *(con)textuel*, et bien que les caractéristiques soient diversifiées en fonction des supports, on peut, entre autres, citer, pour le DEM/DNM :

(a) la promotion du relationnel par rapport aux autres objets du discours (divers sentiments et attitudes ayant trait à la rapidité de l'échange, angoisse d'un nouveau type, accusé de réception (« *je ne sais pas si vous avez bien reçu*

mon courrier »), impulsivité, agressivité (« *Tu me pompes le peu d'énergie qui me reste…* » ; « *Quand on a été élu […], il faut être un minimum responsable* »), réajustements ultérieurs (« *j'espère que je ne t'ai pas blessée avec certaines de mes paroles. Avec l'email c'est toujours le danger. On n'a pas la patience d'écrire des pages, une bonne discussion c'est quand-même mieux* »), accoutumance, incapacité à assumer des rencontres en face-à-face) ;

(b) une prise en compte spécifique de la situation de communication (réduction, voire absence de formules d'ouverture et de clôture, bouleversement des tours de parole, de l'ordre, de la séquentialité, etc.)
(Compilation de Panckhurst 2006a, 2009)

Certains **aspects inattendus** ont surgi lors de mes recherches sur le DEM/DNM, que j'extrais de (Panckhurst 2006a) :

1. *Erreurs* : L'analyse des courriels d'étudiants (corpus 1996-1997) a montré que ceux-ci comportaient des erreurs. Les enseignants trouvaient cela normal : « les étudiants ne savent plus écrire correctement ». À titre comparatif, nous avons analysé également des messages en provenance de collègues. Ceux-ci contenaient également certaines erreurs grammaticales « graves » qui ne seraient sans doute pas apparues dans un autre contexte, par exemple dans des lettres manuscrites.
2. *Appropriation de l'outil* : Nous avions fait l'hypothèse que les étudiants, se servant d'un forum électronique pour la première fois en situation pédagogique (corpus 2001), resteraient dans un cadre formel circonscrit, en respectant, par exemple, des règles fournies par l'enseignant. Nous avons ainsi donné quelques indications concernant les ouvertures et les clôtures de message. Cependant, les étudiants se sont approprié extrêmement rapidement l'outil en fonction de leurs propres usages et besoins, même en situation pédagogique et même après avoir reçu des consignes strictes.
3. *Situation de communication* : Au début de ces recherches, nous nous attendions à ce que les étudiants, précisément, parce qu'ils étaient en situation pédagogique, utilisent les outils électroniques (le courriel uniquement à cette époque) pour des échanges relevant du contexte pédagogique/didactique. En fait, les étudiants se sont servis aussi de l'outil pour faire part

> d'information hors contexte stricte pédagogique.
>
> (Panckhurst 2006a, p. 351).

Comme annoncé précédemment, suite à ma définition terminologique (Panckhurst 1997a), mes analyses en linguistique-informatique ont porté sur les discours contenus dans des corpus de **courriels** en provenance d'étudiants et de collègues (Panckhurst 1998a,c, 1999a,b ; Panckhurst et Bouguerra 2003), avant de me focaliser sur des recherches comparatives entre **courriels, forums** et **chats** (Panckhurst 2003a, 2006b). Ci-dessous, je les résume rapidement. Pour un survol synthétique, je conseille directement la lecture de (Panckhurst 2006a, 2007).

Outils de TAL

Dans (Panckhurst 1998c), j'ai utilisé des outils de TAL et de linguistique de corpus, disponibles à l'époque (le *Correcteur 101* — vérification lexicale, typographique et partiellement syntaxique ; *Conc* — repérage de concordances ; *Nomino* — lemmatisation et analyse syntaxique) afin d'analyser les fautes/ratages dans l'écriture saisie, figurant au sein des messages étudiés. Puis, j'ai déterminé quels types de fautes étaient utilisés en les divisant en catégories (*cf.* également la page 153) :

> On peut diviser les erreurs [76] lexicales, typographiques et syntaxiques en deux catégories majeures :
>
> — machinales, discriminantes [77] ;
>
> — floues, non-discriminantes [78].
>
> Les erreurs « machinales » sont probablement engendrées uniquement par l'utilisation de l'ordinateur [79], tandis que les erreurs « floues » sont plus difficiles à déterminer : elles proviennent soit d'une méconnaissance de règle, soit d'une erreur due au moyen utilisé. L'utilisation de l'ordinateur implique une certaine surcharge cognitive ; on réfléchit en écrivant, mais ce moyen exige une certaine rapidité, immédiateté, d'où l'acceptation par autrui d'« erreurs permises ».
>
> (Panckhurst 1998c), p. 33.

76. Maintenant, je préfère utiliser le terme « faute » plutôt qu'« erreur », ce dernier étant connoté de manière plus négative à mon sens.
77. Fautes de saisie, de répétition ou d'omission de caractères, d'espacements typographiques,
78. Fautes d'accord, pronoms/noms/déterminants manquants, etc.
79. Ma citation date de 1998. On peut bien sûr ajouter, de nos jours, le téléphone portable, les tablettes, voire tout outil d'écriture numérique.

Si je reviens sur ces travaux concernant les « ratages », c'est précisément car le type de faute rencontré dans l'exemple cité ci-dessous m'avait interpellée à l'époque et m'avait motivée à continuer à creuser mes recherches dans le domaine :

> Nous sommes convaincue que l'utilisation de l'ordinateur pourrait réduire nos capacités à appréhender le champ cognitif du contexte environnant de notre production écrite. Hypothèse saugrenue ? […]
>
> Je vous envoie un message afin de vous transmettre *mon* nouvelle adresse electronique (Message d'une étudiante) [80]
>
> L'utilisation non corrigée du déterminant possessif *mon* à la place de *ma* dans une séquence écrite par une francophone est particulièrement pertinente. Il semble inconcevable qu'une francophone ignore le genre du nom *adresse* ; ainsi, a-t-elle écrit de prime abord « mon adresse » avant de revenir sur sa production et d'insérer « nouvelle » ? La touche d'effacement ne nous permettra pas de le déterminer, mais l'hypothèse paraît plausible, tel un « ratage » dans une interaction orale : « mon ad…, ma nouvelle adresse… » (Panckhurst 1998c, p. 36).

J'ai poursuivi la comparaison : entre courrier électronique, traitement de texte, écrit manuscrit (dans le cadre d'examens d'étudiants) (Panckhurst 1998a). Par ailleurs, je découvrais que les fautes n'étaient pas circonscrites au seul public estudiantin ; les collègues, dont je qualifierais la grammaire d'« irréprochable » en d'autres circonstances, en faisaient également, ce qui m'intriguait. Je commençais à aborder également la situation du DEM/DNM, puis à étudier ses marques : déictiques, verbes et temps verbaux, aspects interdialogiques, négation, interrogation, etc.

Ensuite, j'ai évoqué la dimension du discours perçu comme étant « autre », et, plus précisément, les éventuels « dérapages » pouvant découler de l'utilisation du courrier électronique : rapidité/angoisse, agressivité/malentendus, choix du moyen de communication *vs*. une conversation en face-à-face, etc. (Panckhurst 1999b). La situation de communication, les outils de médiation (Vygotsky) et les genres de discours (Bakhtine), ainsi que les aspects pragmatiques et énonciatifs liés à l'utilisation du courrier électronique ont été également étudiés.

80. Tous les exemples figurent tels quels avec leurs « ratages » éventuels.

Puis, j'ai élargi le corpus d'étude — portant toujours sur l'analyse du courrier électronique — à un peu plus de 1500 messages de courrier électronique, que j'ai analysés à l'aide de deux outils : *Nomino* et l'étiqueteur et l'analyseur syntaxique, *Cordial* (dans sa version universités) (Panckhurst 1999a). Ils m'ont permis d'évaluer : la répartition des catégories grammaticales (noms (N) vs. verbes (V)), les types de verbes (auxiliaires : *être, avoir* ; semi-auxiliaires de modalité (*sembler, paraître, devoir, pouvoir*) auxquels j'ai ajouté *savoir, croire, vouloir* (selon Leeman, 1996)) et les temps verbaux, les pronoms personnels, etc. Même si le DEM/DNM se rapproche parfois de l'oral/des oraux et parfois de l'écrit/des écrits, les taux de fréquence concernant la répartition N/V au sein du corpus de courriels concordent avec (Gadet 1996) et (Halliday 1989), le plaçant nettement, pour ce point spécifique, du côté de l'écrit, avec un pourcentage élevé de noms.

Dans (Panckhurst et Bouguerra 2003), mon co-auteur a proposé une réflexion autour de l'utilisation didactique en français langue étrangère (FLE) de la communication électronique. Pour ma part, j'ai mené une recherche sur l'analyse des messages électroniques reçus sur les listes officielles en provenance de la présidente de notre université (de 2000 à 2002). Cette étude était partie d'un constat : j'entendais des collègues se plaindre du surplus de l'information. L'idée de vider leur boîte aux lettres électronique devenait un véritable fardeau, alors qu'ils ne se plaignaient jamais à propos de leur boîte aux lettres dans la salle du courrier. Je voulais regarder cela de plus près. Dans un contexte de mutation des pratiques au sein de l'université, il s'agissait de comprendre si l'utilisation du courrier électronique soit de nature à améliorer (ou non) la communication interne à l'université [81]. En plus des logiciels utilisés précédemment (*Nomino* et *Cordial*) pour l'extraction des informations morpho-syntaxiques, j'ai programmé une petite application à l'aide de *Perl*, afin de repérer des combinatoires de catégories grammaticales (notamment N+Adj), pour en étudier l'impact sur les usagers de la communauté universitaire. Parmi les qualificatifs utilisés, celle-ci a permis d'extraire une quantité semblable d'exemples positifs et négatifs : « voie innovante », « action ambitieuse », « image internationale », « service ma-

81. Il est en globalement ressorti que la communication était améliorée mais que certains aspects devaient être modifiés (formation minimale des personnes, impulsivité des écrits, impact différencié en fonction de la présence quotidienne ou non des personnels, hiérarchisation des sujets, problèmes récurrents de la rétention d'information…).

jeur », « dysfonctionnement grave », « invasion dangereuse », « nuisance actuelle », « charge épuisante », etc.

Courriels, fourms, chats

Grâce aux cours fournis en formation ouverte et à distance (FOAD), j'ai également étudié le contenu des forums et des chats étudiants (Panckhurst 2003a, 2006b). Ces résultats ont ensuite été confrontés au corpus de courriel (Panckhurst 1999a).

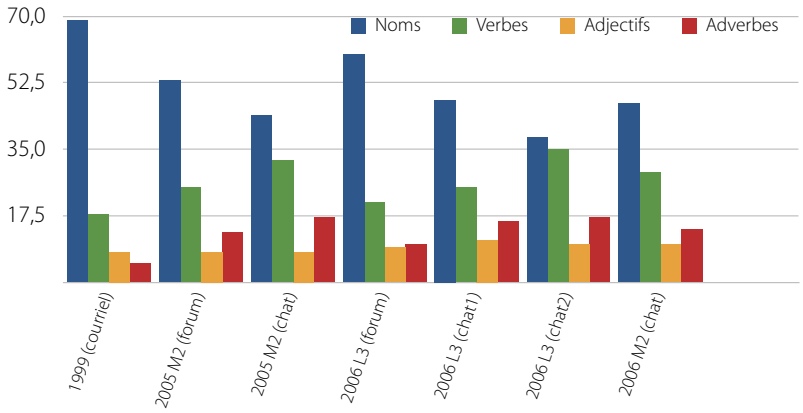


Figure 3.9 – Catégories syntaxiques (occurrences) utilisées (courriel, forums, chats) entre 1999 et 2006 (Panckhurst 2006b)*

*Remarque : J'ai légèrement modifié l'ordre d'apparition des données ici, par rapport à la publication initiale, afin de respecter l'ordre chronologique.

J'ai exposé ces résultats dans (Panckhurst 2009), pour la version française :

> Jusqu'en 2005, nos corpus montraient une similitude entre les catégories syntaxiques utilisées dans le DEM/DNM et celles apparaissant au sein d'autres formes écrites plus conventionnelles (c'est-à-dire une utilisation importante de noms et une quantité relativement faible de verbes). À partir de nos corpus de 2005 (et le corpus de 2006 le confirme), un changement notoire s'est effectué : figurent une réduction de noms, une stabilité adjective et une augmentation de verbes ainsi que d'adverbes modalisateurs. Cela démontre une évolution linguistique radicale ; les résultats de l'analyse des catégories syntaxiques indiquent que le DEM/DNM peut se rapprocher de l'oral ou de l'écrit, et cela varie selon l'outil de communication adopté. Dans la figure 3.9, le chat apparaît comme correspondant à un moyen

> de communication plus « oral », tandis que les forums et le courriel semblent être davantage rattachés à un moyen plus « écrit ». (Panckhurst 2009, p. 38).

L'extrait suivant de ma conclusion de (Panckhurst 2006b) souligne l'importance de l'utilisation d'outils de TAL pour l'analyse du DEM/DNM, afin de nous aider à percevoir certains changements, certaines mutations dans les pratiques langagières. Par ailleurs, cette étude a montré que le choix d'un moyen de communication peut avoir des conséquences importantes sur l'efficacité pédagogique, notamment : le chat — qui contient un pourcentage global élevé de verbes et d'adverbes vs. un pourcentage réduit de noms — serait plus approprié pour une situation sociale, orale ; le forum et le courriel — qui, *a contrario*, contiennent un pourcentage global élevé de noms et un pourcentage réduit de verbes et d'adverbes — seraient plus utiles pour des échanges d'information :

> In our study using automatic analysis with Cordial, syntactic features indicate that chat sessions, which contain an overall higher percentage of verbs and adverbs and a lower percentage of nouns, may be more appropriate for oral, social communication, whereas forums and emails, which contain an overall higher percentage of nouns and a lower percentage of verbs and adverbs, may be more readily used for exchanging information (Crystal 2001, p. 168). Many of our colleagues choose chat sessions (rather than forums) with distance-education students because they see them as an important tool for creating a virtual « community » instead of simply a « group » of students, as well as for maintaining links and reducing student dropout. However chat sessions are not the right tool to use in all pedagogical situations (as forums may be more appropriate in specific contexts) and teachers may not necessarily perceive this. The choice and use of communication methods needs to be thought through carefully. Choosing the right communication tool for a particular pedagogical context is important for coherent learning ; linguistic analysis which demonstrates features related to various communication methods (i.e. syntactic indications highlighting oral/written, informal/formal usage, etc.) can help in making the most effective choice, but broader comparative linguistic, extra-linguistic (Panckhurst et Bouguerra 2003) and cross-disciplinary research is necessary in order to understand more about current language and communication situations. MED has changed over the past ten years, according to communication methods chosen and as a result of overall language evolution. (Panckhurst 2007, p. 168).

En février 2005, un *événement* dans le cadre pédagogique m'a fortement interpellée. Le contexte était le suivant : les étudiants effectuaient un travail en groupe pour un cours de compétences informatiques (C2i), à l'aide de forums exclusivement entre eux, entre pairs. Ils étaient prévenus que j'accéderais à leurs discussions, voire même que celles-ci entreraient dans mon évaluation finale de leurs travaux. Que s'est-il passé ? Un étudiant a posté un message au sein d'un forum, contenant une forme *d'écriture de type SMS* (désormais *eSMS*) :

> je c ke c chian de devoir se connecter mé en fait on é noté ossi sur le nombre d'échange et ce qu'on dira dc si vous avez un ordi à disposition ca seré cool d'envoyer des infos sur d sites par ex pr le dossier ou se genre de truc pr le site je pense ke ca devré allé vite je vou envéré un message ds les jours suivants sur chaque page k on peu faire et chacun poura dire ce ki lui plé le plus de faire moi je m en fou voila. (forum entre étudiants, travail en groupe, 7/2/05)
>
> Différents phénomènes y apparaissent comme ceux relevés par Guimier de Neef (in Guimier de Neef & Véronis, 2004). De manière non-exhaustive, nous en exposons quelques-uns ici :
>
> — modifications de type phonétique (*c, mé, é, ossi, seré, envéré*...),
> — suppressions de la lettre finale d'un mot (*chian, vou, fou, peu*...),
> — des troncations intermédiaires, des abréviations, des réductions aux squelettes consonnantiques, (*dc, ordi, pr, ds*) etc.
>
> L'extrait analysé présente également une absence totale de ponctuation (aucun point, aucune virgule, absence d'apostrophes, etc.). Ces usages s'accompagnent d'un problème de traitement automatique du langage.
>
> (Panckhurst 2006a, p. 361).

Depuis 1996, c'était la toute première fois — replaçons-nous dans le contexte relativement formel de l'enseignement supérieur, dans lequel ce type de découverte pouvait paraître étonnant — que j'ai recueilli un exemple de cette nature, que je qualifierais d'extrêmement riche.

En 2006, l'heure était au bilan. L'ouvrage de Piolat (ed, 2006) dans lequel s'inscrit (Panckhurst 2006a) m'avait permis de collaborer avec des collègues dans un contexte pluridisciplinaire : *linguistes, informaticiens, linguistes-informaticiens, psychologues, neurobiologistes, spécialistes de l'information et de la communication*.

Les sujets étaient très variés, et concernaient des espaces de lecture, d'écriture, de communication, et d'apprentissage à partir d'Internet, allant de l'accessibilité numérique pour les aveugles, à l'esthétique dans la conception et l'utilisation de sites Web, aux traitements de l'écriture SMS par des outils de synthèse de la parole, en passant par des aides de lecture pour les sourds par le biais de personnages de synthèse, etc. Cela m'a énormément apporté.

La même année, au mois de mai, j'ai participé au colloque *La langue de la communication médiatisée par les technologies de l'information et de la communication* (*CMT*) à Bordeaux, où j'ai présenté mes travaux comparatifs du DEM/DNM entre courriels, forums et chats dans le contexte de l'enseignement supérieur (Panckhurst 2007). Entre autres collègues que j'y ai rencontrées, Sébastien Paumier a fait une présentation du projet belge « Faites don de vos SMS à la science », *SMS4science*, (Fairon et al. 2007, 2006b,c), que j'évoquerai plus bas. Il venait de participer, entre autres, au protocole de transcription et à la transcription manuelle d'une (grande) partie du corpus de SMS « bruts » recueillis en français standardisé. Pour l'anecdote, il m'a confié : « je ne veux plus jamais voir un autre SMS de ma vie ! ». Quelques années plus tard, j'ai mieux compris cette surcharge cognitive suite à des lectures intensives de SMS. J'y reviendrai *infra* (§ 3.3.3.3).

Par ailleurs, au colloque *ascilite*, qui s'est tenu à Sydney en décembre 2006, j'avais présenté deux recherches : l'une qui ciblait le DEM/DNM et les analyses possibles de messages de forums/chats à l'aide d'outils de TAL (Panckhurst 2006b), et l'autre à propos des réseaux d'échanges en eLearning (Volet 2, Panckhurst et Marsh 2006). Le croisement des recherches entre les volets 2 et 3 apparaissait alors. On le comprend aussi en lisant la conclusion du chapitre bilan (Panckhurst 2006a) :

> Il est important de poursuivre ces recherches, dans une double direction. Sur un plan plus pragmatique, comprendre comment se forment des communautés en ligne. Les étudiants échangent en tant que groupe et forment ensuite des communautés. Comment un individu réagit-il et évolue-t-il au sein de ces communautés ? Il est important de comprendre davantage ces processus, en tant qu'enseignants et tuteurs pour en tirer un meilleur usage didactique. Il faudrait, également, effectuer plus de recherches sur des aspects mêlant les techniques de TAL et l'étude des textos. Ces derniers commencent effectivement à envahir

l'espace communicationnel virtuel dans l'enseignement supérieur.
Afin de poursuivre le travail dans ces deux directions, élaborer d'autres corpus s'avère nécessaire. En fonction de l'évolution des usagers, il devrait alors être possible de confirmer que les textos font leur apparition dans le contexte pédagogique asynchrone dans lequel l'enseignant est également impliqué. Il y a quelques années cet usage n'aurait pu être envisagé. Penser que le langage écrit des SMS ferait son apparition de manière de plus en plus importante dans les espaces asynchrones était inimaginable. On peut le comprendre pour le chat, mais plus difficilement dans un forum de discussion universitaire. Il est vrai que le corpus C2i présenté ici contenait uniquement des discussions entre pairs. Mais nous sommes sûre que nous verrons pointer ce type de discours prochainement dans les forums entre étudiants et enseignants. Nous n'avons sans doute pas pris encore la mesure de cette révolution langagière.
En tout état de cause, comme le souligne (Herring 1996), les études doivent être menées de manière pluridisciplinaire, afin de mieux comprendre cette évolution langagière moderne. Selon elle : « Ce secteur [...] continuera à l'avenir à rassembler plusieurs disciplines universitaires et à faire en sorte qu'elles apprennent les unes des autres. Parallèlement, l'émergence de sous-spécialisations sera favorisée par l'expertise disciplinaire appliquée aux questions portant sur la CMO. Un approfondissement multidisciplinaire est nécessaire pour nous aider à mieux comprendre ce phénomène technologique aux implications profondément humaines. » (p. 10, notre traduction).
Par ailleurs, il est remarquable que les usages langagiers évoluent toujours de manière aussi créative. Nous pensons que ce domaine constituera un terrain de recherches fructueuses pour bien des années futures. Comme le souligne Crystal : « Il est vraiment remarquable qu'autant de personnes se soient si rapidement adaptées aux exigences linguistiques de ces situations nouvelles et aient su exploiter tout le potentiel du nouveau medium de manière suffisamment créative pour favoriser l'émergence de nouvelles formes d'expression. Cela s'est produit en quelques décennies seulement. J'en conclus que la faculté linguistique de l'être humain est en bonne forme. L'apparition du parler Net nous révèle ce que l'*Homo loquens* a de meilleur. » (notre traduction,

(Crystal 2001), in (Panckhurst 2006a, p. 362-364).

3.3.3.3 SMS

Cette première apparition de texto, qui a surgi dans mon espace d'enseignement supérieur et de recherche, a marqué le début d'un grand tournant dans mes recherches — ou plus modestement, d'un élargissement de mon champ — afin de prendre en considération les écritures de type SMS, dans un premier temps au sein de l'espace communicationnel de l'enseignement supérieur, mais assez rapidement, on le verra, dans un espace plus ouvert en direction du grand public. Dans la citation incluse ci-dessus, on peut déjà percevoir les orientations qui allaient me passionner par la suite : techniques de TAL et étude des textos, nécessité de pluridisciplinarité dans les recherches, créativité scripturale…

Tout en continuant mes recherches en binôme avec Debra Marsh autour de nos expérimentations pédagogiques en eLearning (*cf.* Volet 2, § 3.3.2), je m'intéressais de plus en plus à l'idée de travailler sur l'écriture SMS. En mars 2008, j'ai été invitée à Louvain-la-Neuve, en Belgique, afin de participer à une réunion d'organisation et de partenariat avec le CENTAL (Centre de traitement automatique du langage) de l'université Catholique de Louvain (UCL).

Quelques années auparavant, les collègues belges (Fairon et al. 2006b,c) avaient procédé à une vaste collecte de SMS auprès du grand public, dont le slogan était « Faites don de vos SMS à la science ».

> En 2004, un groupe d'universitaires belges a démarré un projet international, intitulé *sms4science*, afin de recueillir, organiser (en une base de données mondiale), et analyser des SMS authentiques (www.sms4science.org, (Fairon et al. 2006b,c), (Cougnon 2015)). S'en sont suivies d'autres collectes de SMS : l'île de la Réunion (Cougnon et Ledegen 2010), la Suisse (Dürscheid et Stark 2011), le Québec (Langlais et al. 2012), la région Rhône-Alpes en France (Antoniadis et al. 2011) [82]

82. Par ordre chronologique, après le recueil belge initial, les pays et régions suivants ont récolté des SMS authentiques, toujours dans le cadre du projet international *sms4science* : Île de la Réunion : http://www.lareunion4science.org/ (20 000 SMS, 2008), (Cougnon et Ledegen 2010) ; Suisse : http://www.sms4science.ch/ (24 000 SMS, 2009-2010), (Dürscheid et Stark 2011) ; Québec : http://www.texto4science.ca/ (5 000 SMS, 2010), (Langlais et al. 2012) ; Grenoble : http://www.alpes4science.org/ (22,000 SMS, 2010), (Antoniadis et al. 2011) ; Colombie Britannique (14,300 SMS, 2012, http ://www. text4science.ca/ ; (Drouin et Guilbault 2016). L'initiative la plus récente pour le français est le projet *sud4science LR* (www.sud4science.org).

> En trois mois, (15/9/11 au 15/12/11), plus de 90 000 SMS authentiques ont été recueillis auprès du grand public par un groupe de chercheurs dans la région Languedoc-Roussillon. [83] Accorsi et al. 2014 ; Panckhurst et Moïse 2014 ; Patel et al. 2013 ;Panckhurst et al. 2013, p. 108-109.

Je reviendrai plus bas sur le projet *sud4science LR* (www.sud4science.org) et le recueil de SMS à Montpellier qui s'en est suivi (§ 3.3.3.4).

Le fruit de la réunion de Louvain en 2008 était une convention de partenariat tripartite d'une durée de 5 ans, entre l'UCL, le CNRS et l'université Paul-Valéry Montpellier 3, par le biais du laboratoire Praxiling. Étant donné l'expertise acquise par le CENTAL (UCL) dans « le développement et l'exploitation de corpora et de base de données contenant des messages textuels-SMS notamment par la réalisation du projet SMS4science », et étant donné l'objectif du prolongement du projet, qui était de « contribuer à l'étude de la communication par SMS et à l'étude du langage qu'elle véhicule en établissant des partenariats avec les autres universités », la convention avait pour but, « la constitution de vastes sous-corpus nationaux et, par consolidation, d'un corpus international de SMS de langues diverses pour la recherche scientifique » [84].

Nous avions donc une jolie convention. Mais j'ai lutté pendant près de trois ans pour essayer de trouver un opérateur téléphonique français intéressé par un projet de recherche universitaire de recueil et d'analyse de SMS. Après tout, des

83. *Pourquoi avoir constitué deux collectes distinctes de SMS en France métropolitaine, Grenoble-Alpes et Languedoc-Roussillon ?* Pour deux raisons : 1. Dans un premier temps, il est vrai que nous voulions étudier des régionalismes variés, suite aux comparaisons effectuées par Louise-Amélie Cougnon entre la Belgique, la Réunion, le Québec, la Suisse (*cf.* (Cougnon 2015, 3e partie, « De la variation des données »). Restreindre la collecte à une étude régionale en Languedoc-Roussillon aurait donc pu être intéressant. Mais comme nous avions oublié de le stipuler dans les renseignements légaux aux participants, nous avons dû accepter des SMS en provenance d'autres régions. Si besoin est, une extraction pourrait être envisagée via le code postal qui était demandé dans le questionnaire. 2. Il nous semblait pertinent d'effectuer une nouvelle récolte en 2011 — après celle effectuée dans la région Rhône-Alpes — suite à l'inclusion massive des SMS illimités dans les forfaits téléphoniques mensuels, puis grâce à l'adoption d'une méthode de récolte novatrice (*cf.* la note 103).

84. Cette convention a été signée le 23 avril 2008 par le Délégué Régional du CNRS, B. Jollans, le Président de l'université Paul-Valéry Montpellier 3 de l'époque, J.-M. Miossec, le Recteur de l'université Catholique de Louvain, B. Coulie.

collectes avaient déjà eu lieu dans d'autres pays, pourquoi pas nous ? [85] À l'époque (avant 2011, donc), la plupart des forfaits téléphoniques n'incluaient pas les SMS illimités, et, afin d'inciter le grand public à faire don de leurs SMS à un projet, mieux valait afficher la gratuité de l'envoi de leurs données. C'était le nœud de l'affaire. Je suis allée à la rencontre de plusieurs grands opérateurs téléphoniques en France métropolitaine, mais il n'y avait rien à faire ; décidément, ils ne comprenaient absolument pas l'intérêt en matière de retombées économiques pour eux, ni même l'idée « saugrenue » que quelqu'un voudrait s'intéresser à cette écriture qui « massacrait la langue française ».

Jusqu'en 2010, les collectes dans les 4 pays avaient transité par un opérateur téléphonique, via, le plus souvent, un système de numéro court : les donateurs de SMS redirigeaient les SMS qu'ils envoyaient (ou avaient déjà envoyé) à autrui, à un numéro court, lié à un opérateur téléphonique, et les messages ainsi recueillis étaient ensuite redirigés aux chercheurs du projet. Certains problèmes ont émergé : afin de distinguer un projet d'un autre — car on partage souvent les numéros courts, qui sont relativement coûteux, entre projets de recherche — il fallait que le donateur de SMS saisisse un code en début d'envoi, qui permettait à l'opérateur téléphonique d'identifier le projet, afin de correctement rediriger ensuite les données aux chercheurs. Mais, comme les SMS étaient découpés en séries de 160 caractères, dans le cas où le SMS dépassait la limite, la partie figurant au-delà de la taille maximale était tronquée, car n'ayant pas de code correspondant au projet au début de la suite du SMS. Je voulais absolument éviter cela, afin que le corpus ne soit (trop) biaisé.

J'ai pris mon mal en patience et j'ai continué à effectuer des recherches avec mes étudiants sur l'écriture SMS, qui me passionnait de plus en plus, en français et en d'autres langues, dans des domaines divers. Quelques années auparavant, Filipov (2006) avait effectué une étude comparative linguistico-communicationnelle entre le français et le bulgare, Svarinska (2006) sur le russe, Fayada (2007) sur le « langage SMS des jeunes », et Caumont (2007) sur une approche phonologique des SMS. Puis, Gaussuron (2009) sur les « SMS et déficiences visuelles », Catapano

85. Deux collectes ont néanmoins eu lieu en France avant la nôtre, à la Réunion en 2008 (Cougnon et Ledegen 2010), et dans la Région Rhône-Alpes, en 2010 (Antoniadis et al. 2011). Dans le sud de la France, en revanche, la situation était rédhibitoire.

(2009) sur « conversations et SMS » et Fontaine (2010) sur une analyse comparée des outils asynchrones : courriels, forums, SMS [86].

En 2009, j'ai écrit un chapitre dans ouvrage en mémoire à une collègue récemment décédée, Michelle Lanvin, phonéticienne. Suite à mes travaux sur les UVPL, et parce qu'une approche phonétique, entre autres, de l'écriture SMS m'interpellait, j'estimais que c'était pertinent de lui rendre hommage via une typologie de l'écriture SMS, incluant une dimension qui touchait directement sa spécialité. (Panckhurst 2009, p. 40). fournit une typologie néographique (et non néologique, *cf*. aussi (Roche et al. 2016)), divisée en 4 grands phénomènes (*substitution*, *réduction*, *suppression*, *ajout* de caractères), légèrement remaniée dans (Panckhurst et al. 2013, p. 125-126)). Je ne propose pas la typologie d'origine ci-dessous, mais plutôt la nouvelle variante (Roche et al. 2016), qui incorpore une modification dans la partie *ajout*, suite à une discussion fructueuse avec Frédéric André, alors doctorant à Paris-Sorbonne sous la direction de Gilles Siouffi (André 2017), et ayant été stagiaire dans le cadre de notre projet *sud4science LR* lorsqu'il était étudiant à Montpellier — la preuve que mes étudiants ont une influence importante sur mes recherches ! Je reprends l'explication synthétique de ces 4 phénomènes ci-dessous, et j'explique la nécessité de ces changements dans l'annexe de (Roche et al. 2016).

> L'écriture SMS (eSMS) [87] est riche, innovante, créative. (Panckhurst 2009) a proposé une typologie (uniquement *néographique* et non *néologique*) des SMS pour le français, suite à d'autres chercheurs (Anis et al. 2004 ; Fairon et al. 2006b,c ; Liénard 2005 ; Véronis et Guimier de Neef 2006). Pour comprendre la complexité de l'eSMS, il faut distinguer les cas de phénomènes *simples* où un seul phénomène par segment ou item lexical [88] est constaté et les cas de phénomènes *complexes* où plusieurs phénomènes simples figurent simultanément. Par exemple, lorsque l'orthographe d'un lexème (*eau* ou *au*) est totalement modifiée en une

86. Pour les intitulés exacts de ces travaux universitaires, *cf*. l'encadrement de la recherche (cf. tableaux 3.1, 3.21) et mon Curriculum Vitæ. En bibliographie générale, je n'indique que les références des dossiers et des mémoires que j'ai (co-)encadrés, ou pour lesquels j'ai participé à soutenance, ou qui ont un lien privilégié avec le projet *sud4science LR*, et ce depuis 2011.

87. Nous préférons *écriture SMS* (eSMS) à *écrit SMS* (Cougnon 2015). Nous refusons l'appellation *langage* SMS.

88. Dans ce contexte, un segment, ou item lexical, sera considéré comme une suite de caractères compris entre deux espaces. Par exemple, « 12c4 » correspond à un seul item lexical.

> lettre (*o*), nous sommes face à une *substitution phonétisée entière simple*, car une seule modification est effectuée. De même, cela peut concerner des digrammes ou trigrammes, qui transcrivent un phonème. Dans ce cas, l'orthographe du lexème est partiellement modifiée (*jamé* pour « jamais », *boC* pour « bosser », etc.) et il s'agit d'une *substitution phonétisée partielle simple*. En revanche, un exemple comme « *2m1* » (demain) comporte deux substitutions au sein d'un même item lexical : « de » [də et « m1 » [mɛ̃ = « demain » [dəmɛ̃. Cela correspond effectivement à une situation de *phénomène complexe*.En effet, au sein des SMS, les phénomènes *complexes* abondent. Par exemple *12c4* (« un de ces quatre »), *6T, 2manD* (*substitutions multiples*, dans ce cas), mais d'autres cas traversent les catégorisations, par exemple, des *substitutions* + des *réductions* + des *suppressions* : *7éta* constitue à la fois une réduction graphique en agglutination, une suppression de fin de mot muette, une substitution phonétisée entière. Pour *chui, chais, yora, kestufé*, on a des agglutinations + du compactage + des écrasements. On peut rencontrer des cas d'ajouts/de variations, etc. : *moua*, *suuuupppeer*.Quelques précisions sur cette typologie essentiellement morpho-lexicale [89] : la *substitution* correspond à un remplacement total ou partiel de la graphie standardisée ; cela n'exige pas nécessairement une réduction du nombre de caractères (par exemple, *nan* pour « non », *mwa* pour « moi » *m en* pour « m'en » — où l'apostrophe est remplacée par l'espace). En revanche, la *réduction* correspond systématiquement à un retrait de caractères, donc cela mène à une diminution du nombre total de caractères. De même pour la *suppression* qui correspond à un enlèvement graphique total (signes de ponctuation, signes diacritiques, marques typographiques).La figure 3.10 présente une typologie de tous les phénomènes rencontrés au sein des SMS : *substitution, réduction, remplacement, ajout*. Les 20 exemples authentiques [90] qui le suivent correspondent chacun à un cas distinct de la figure.
>
> (Roche et al. 2016)

Pour l'anecdote, (Panckhurst 2009) est ma publication la plus citée sur *Google Scholar*. Je pense que cela est dû à sa dimension de recherche appliquée, proposant une (nouvelle) typologie des SMS, pouvant être testée par des chercheurs de différentes disciplines (voir par exemple, en psychologie, Bernicot et al. 2014b,

89. Dans ce cadre, je ne traite pas des questions (morpho) syntaxiques ou discursives : changements de catégorie grammaticale, ellipses, rafales de questions/réponses, etc.

90. Ces exemples ont été extraits du corpus *88milSMS* par Frédéric André et Rachel Panckhurst.

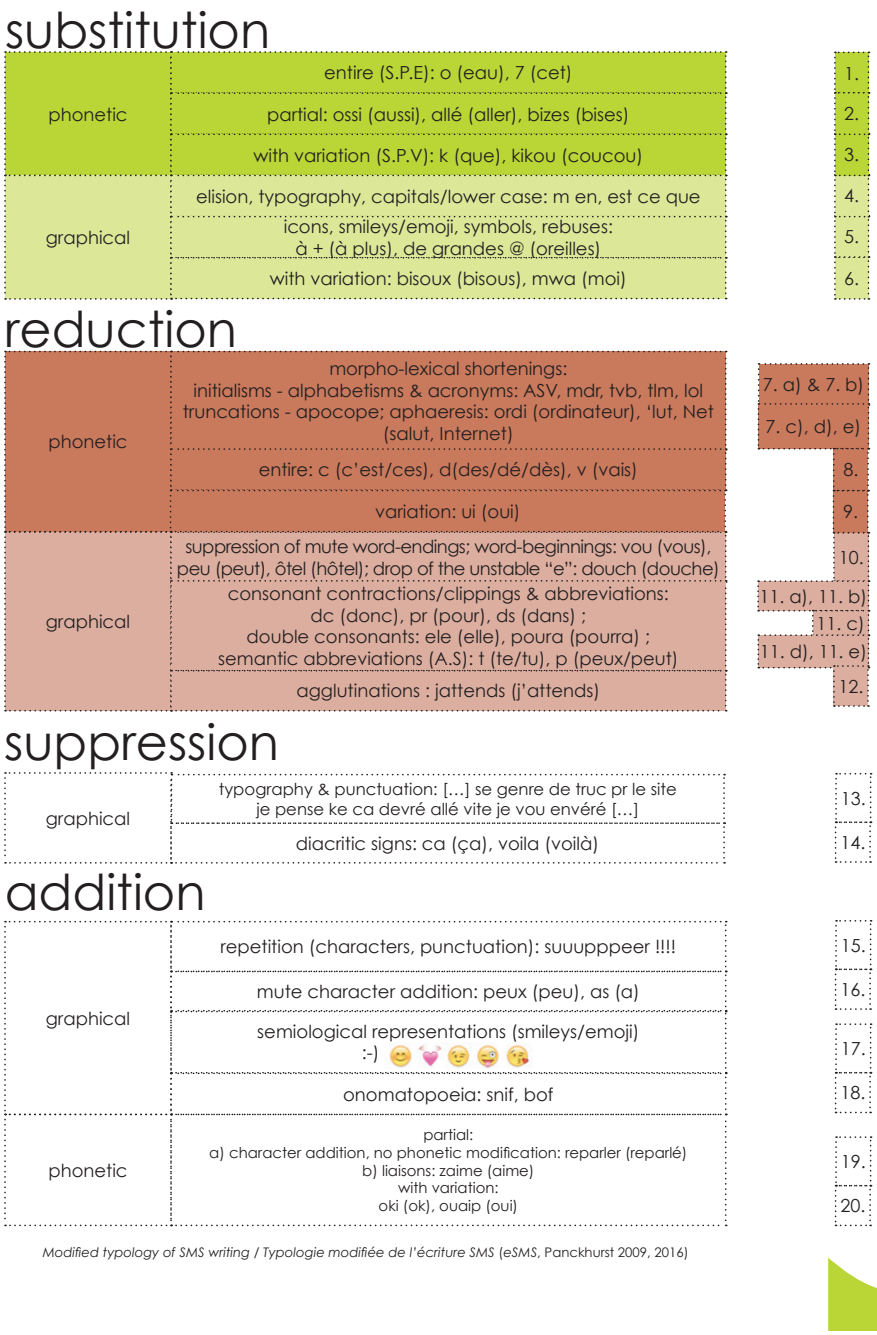


Figure 3.10 – Typologie modifiée des phénomènes simples de l'écriture SMS, présentée à CINEO, Salamanque, 2015, à l'ENS, Lyon, 2016, et à l'LIGM-UPEM, Paris, le 20 janvier 2017.

Exemples correspondant à la typologie modifiée.

1. 30168, J ai **u** <PRE_6> **o** tel hier soir (pdt que monsieur recevai...)elle est ravie...tu as assure sur ce coup la,@+

2. 3577, Cc <PRE_1> !! lèr soir g pa insisté com ta pu le voir, **voulé** pa **ke** <SUR_7> sénerv a koz 2 moi !!! 1é cour 2 robin d boi ce soir à 20 h, tjs **Dcidé** a yalé, si le **fé** pa ce soir le **feré jamé**. Pr l€patatoes 2 vend., c tjs pa si on pt chanté «orevoir Présiden é J-<SUR_4>» !! Kèl suspens !!! Ds le doute, préférable kon continu à **boC**. Repriz 2min (12 h) é jeudi matin pui stop pr 7 sem. T sur le piton ou tu te **bala2** ? Bon courage

3. 30657, Logiquement, si on s'embrasse, c'est que ça signifie quelque chose, mais dans la position où nous somme je pense que la réponse est **nan** :)

4. 39406, Ouais mais je **m en** doutais à moitié, il **m en** avait deja parlé. Le salaud il **m a** tjr pas répondu!

5. 53770, Bon bah si déjà tu as trouvé ton plan ca va moi je trouve que c'est le **+** dur

6. 70343, Dit **mwa** pk tfai sa a **mwa** jt rien fait **mwa**

7. a) 71936, Ouais, faut que j'ramène à **tlm** des trucs --' mais j'sais pas quoi **lol**

7. b) 89810, Aaaaahhhh :p tu vient pas en cour demain ? Sinon **tfk** ? Et **qdb** ?

7. c) 92754,Oui! J'etais trop **deg**!

7. d) 31023, Regarde sur le **net**

7. e) 47607, Bouh ɑ.ɑ bon allez espère que ta flemme s'est arrangée un peu.. Un **zou***

8. 5226, Vendredi promis. Je viendrai en scoot. Je **v** aller faire mon certificat.

9. 60286, Slt miss juste un **ptit** mot pour avoir de t news vu ke la dernière fois t'avais pas trop le moral. Bizoo

10. 43283, Pleins 2 bisous a vs 3. On pense a vs et on vs oublie pas.**just** 1 peu trop overbooké en ce moment. je **fini** le taf a 17h demain.j.essairai 2 vs tel.bisous a vs 3.2 ns 4.

11. a) 92836,Ben je vien **pr** 12h ... Dsl ...

11. b) 42632,17 oct. 2011 11:03:39,162,,Bon je pars de la fac la ^^ **dc** a tte ^^

11. c) 60840,9 nov. 2011 11:06:52,120,,Lol on **vera** dan kelk anè lol

11. d) 23631, tu **f** koi

11. e) 40789, **C** np koi **c** jeu lol

12. 44069, Oh tu me crois **jviens darriver jouvre** la porte mon portable vibre :P ; 43315 : C'est bon, j'ai fini. **J't'attends** là bas. Bisous

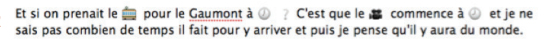
13. 67645, Mais non c est pas une embrouille pour une histoire de carte grise qui a une solution mais cela ne change rien a votre amour elle t aime allez je te sers ton jaune je t attends on va passer une bonne soiree

14. 92 752, **A** chaud... Purr**e** moi j'ai eu un contr**o**le sur stat, primitive et **e**tude de fonction par rapport au co**u**t marginal et d'autres truc comme **c**a, laisse tomber je me suis ramasser, alors que j'avais trop bosser, attends j'avais refait tt les exos j'avais tt bon et l**a** au contr**o**le c'est que des truc hyper dur genre 2 fois plus dur qu'au contr**o**le ! Avec genre on voit pour **a** la puissance 2 et la c'etait puissance 3 on savais pas comment faire pour l'exo ca faisait bizarre

15. 8427, A**hhh** t'es tr**oooooooooo**p f**ooo**rte !! :D :D j'y avais pas pensee md**rrrrr**

16. 32081, […] Pour info, elle m'a**s** dit qu'il y avait plusieurs filles qui voulait sortir avec lui, mais qu'il gardait ses distances. Si t'as besoins de précisions sur un passage, demande, je t'expliquerais mieux. Je te l'accorde, elle peux etre un peux chiante...»...

17. 986, J viens de me faire virer par sms :(

35642 

18. 90944, Ben si jrencontre personne **snif**

19. 53487, Non on en a pas reparle**r**

4360, Cc <PRE_1> ! Ça boum ? <PRE_5> na k bien se tenir ! Oui, j sui **zalé** ! Pensé pa ke c t oçi fisik, v avoir lé **zépol** en Bton ! Ambiance tré 5pa. Si Papi ne cri pa o scandal, je continu. Pense pa mé fo ke lui dde 1 certif é me dira si c bon, en tt k pa mové, pr ce ke g ! 12 h à lépad bien o cho é nouri ki + zé é ceriz sur le gato : av 2-2 : 2 koi je me plin ?!! Ta fé koi ièr soir ? 1 tètàtèt av le PC ? Bone journé. Biz

20. 86258, **Oki oki** Beh bon courage ^^

R. Panckhurst, LIGM-UPEM, 20 janvier 2017

p. 36, Combes et al. 2014, p. 55-56, en information et communication, Liénard, (Marcoccia 2016, p. 82)).

Étude trilingue

Mais je reviens à la période pré-2011. Plusieurs dossiers d'étudiants ont été effectués sur une étude comparative de langues. Par exemple, dans le cadre de la validation de leur cours, (Boudrique et al. 2008) ont comparé les SMS français et espagnols en se servant de ma typologie, afin de la tester sur d'autres langues que le français. (Stabile et Tortorella 2008) ont comparé l'italien (et également le napolitain) et le français. Suite à leurs travaux, en creusant quelque peu leurs recherches et en les citant, j'ai proposé une typologie trilingue que j'ai présentée au colloque i-Mean, *Meaning and Interaction*, à Bristol en 2009.

La figure 3.11 montre quelques exemples plurilingues (*cf.* (Panckhurst 2010) pour plus de détails).

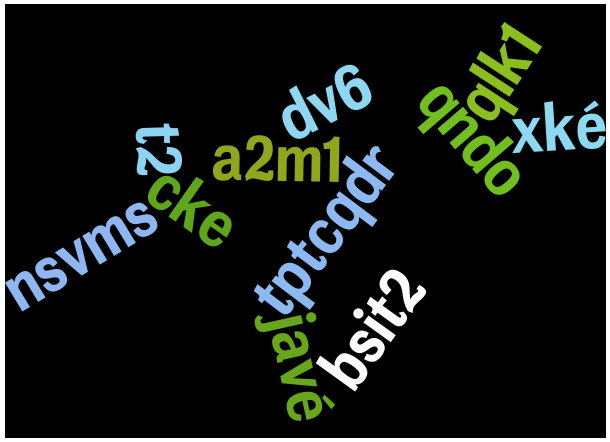


Figure 3.11 – Exemples français, espagnols, italiens, d'écriture SMS [91]

Certains phénomènes sont non-applicables pour l'espagnol et l'italien : la suppression de fins de mots muettes (*peu (peut)*), ou la substitution de trigrammes (*bo (beau)*) ; en espagnol, l'inexistence de l'apostrophe — qui permettrait l'élision, voire l'agglutination (*me acuerdo, *m'acuerdo*). La contraction de doubles

91. a2m1 (à demain), javé (j'avais), cke (c'est que), xké (perché) pourquoi ?, dv6 (dove sei) Où êtes-vous ? qlk1 (qualcuno) quelqu'un, t2 (todos) tous, qndo (cuando) quand, bsit2 (besitos) bises nsvms (nos vamos) On y va, ¿tptcqdr ? (¿te apetece quedar ?) Veux-tu me rencontrer ?

consonnes (*ele (elle), poura (pourra)*) ne figure pas dans les corpus espagnol/italien. En revanche, la variation phonétique en espagnol/italien semblait plus fréquente qu'en français (*weno (bueno), poxo (posso)*). Puis, la ponctuation initiale (¿¡), qui n'existe qu'en espagnol, est systématiquement absente dans les corpus SMS recueillis par (Boudrique et al. 2008).

Substitution graphique en arabe

Si cette étude trilingue a montré quelques variations, elle n'a pas révélé un phénomène qui n'existerait pas dans la typologie proposée pour le français. Mais à Bristol, j'ai rencontré une jeune arabophone — que je remercie de manière anonyme — qui m'a parlé d'un exemple en arabe qui permettrait d'étendre ma typologie. J'étais tout ouïe. Le prénom *Ayesha/Aïcha* (عائشة) pouvait s'écrire avec le chiffre 3, en écriture SMS. Pourquoi ? Car la lettre initiale du prénom est ع, qui graphiquement, ressemble à un 3 en écriture miroir. Grâce à quelques locuteurs arabes de mon entourage professionnel et personnel [92], j'ai pu creuser et vérifier l'existence de ce phénomène, qui n'existe évidemment pas en français : le chiffre qui correspond à un remplacement graphique.

		phon	nom	finale	médiane	initiale	SMS			
2	ا	aː	alif	ـا	ـا	ا	L2nha	لأنها	laynha	parce qu'elle...
3	ع	ʔˤ	ɛayn	ـع	ـعـ	عـ	3sel	عسل	asel	miel
3'	غ	ɣ	ghayn	ـغ	ـغـ	غـ	3'ayma	غيمة	ghayma	nuage
5	خ	x	kha	ـخ	ـخـ	خـ	5rouf	خروف	kharouf	mouton
6	ط	tˤ	Ṭa	ـط	ـطـ	طـ	6bib	طبيب	tabib	médecin
6'	ظ	zˤ , ðˤ	Ẓa	ـظ	ـظـ	ظـ	6'hri	ظهري	zahri	mon dos
7	ح	ħ	Ḥa	ـح	ـحـ	حـ	7bibi	حبيبي	habibi	mon amour
8 (ou parfois 9)	ق	q	qaf	ـق	ـقـ	قـ	8wi	قوي	qawi	fort
9	ص	sˤ	Ṣad	ـص	ـصـ	صـ	9yaad	صيّاد	sayaad	pêcheur
9'	ض	dˤ , ðˤ	Ḍad	ـض	ـضـ	ضـ	9'rsi	ضرسي	dersi	ma dent

Figure 3.12 – Substitution graphique en arabe.

En français, lorsqu'un chiffre est utilisé dans un SMS, il s'agit systématiquement

92. Je remercie Djazia Kenzi, Hani Qotb, Ismahan Brakta, Yahya Al-rahmah, pour leur aide. Les erreurs éventuelles sont les miennes.

d'une substitution phonétisée (*7 (cet)*). En arabe[93], dans les exemples de la figure 3.12, on peut constater que les chiffres de 2 à 9 sont utilisés dans l'écriture SMS.

Plus étonnant, bien que l'écriture en arabe soit modifiée en fonction de la position (initiale, médiane ou finale) des lettres, la substitution graphique demeure (*cf.* figure 3.13). Dans l'exemple suivant, les chiffres 3 et 7 ne correspondent pas systématiquement à la graphie arabe tels quels, à cause de la position, mais l'emprunt des chiffres demeure dans les SMS.

عندش شي العصر؟ ودش نروح كافي؟

3endech shy el *3aser* ? weddech *nroo7* cafe ?

Etes-vous occupé dans l'après-midi ? Souhaitez-vous prendre un café ?

		phon	nom	finale	médiane	initiale	SMS			
3'	غ	ɣ	ghayn	غ	ـغـ	غ	3'rfa	غرفة	gherfa	chambre
							sa3'r	صغير	saghir	petit
5	خ	x	kha	خ	ـخـ	خ	5yma	خيمة	khayma	tente
							ba5l	بخيل	bakhil	avare
							waba5	وبّخ	wabakha	punir
6	ط	tˤ	ṭa	ط	ـطـ	ط	6wil	طويل	tawil	long
							bassi6	بسيط	bassit	simple
7	ح	ħ	ḥa	ح	ـحـ	ح	7riya	حرّية	heriya	liberté
							la7m	لحم	lahm	viande
9	ق	q	qaf	ق	ـقـ	ق	9lb	قلب	kalb	cœur
							fa9r	فقير	fakir	pauvre

Figure 3.13 – Positions (initiale, médiane, finale) des lettres en arabe.

J'ai présenté ces tableaux/exemples dans un diaporama à une conférence invitée[94], mais je n'ai pas creusé cette recherche initiale — des chercheurs arabo-

93. Les exemples fournis en arabe relèvent de locuteurs, de pays, et de variations d'arabe (littéraire, dialectal) différents : Algérie, Égypte, Koweït.

94. Conférence invitée (par Olga Volckaert-Legrier), « SMS et eSMS : une autre forme de communication électronique médiée ? », au Laboratoire « Psychologie du Développement et Processus de Socialisation » (EA 1687), université de Toulouse 2, 17 novembre 2010.

phones l'auront très certainement déjà fait, en l'approfondissant. En tous les cas, si ma typologie (Panckhurst 2009 ; Roche et al. 2016) pouvait servir à d'autres chercheurs pour l'étudier et la confronter avec d'autres langues, cela me ferait très plaisir.

Appel MSH-M

Retour chronologique. Fin 2010. La recherche continuait à petit pas. N'ayant toujours pas trouvé d'opérateur téléphonique enthousiaste, j'ai contacté notre service de valorisation de la recherche. Sabine Cotreaux (Responsable Service Partenariat Recherche) — que je remercie de cette heureuse initiative — m'a conseillé de répondre à l'appel de financement à projet de la *Maison des Sciences de l'Homme de Montpellier* (MSH-M). Paul Pandolfi (alors directeur de la MSH-M) et Florian Pascual (son secrétaire général) nous ont reçues pour nous expliquer le montage. En novembre 2010, j'ai répondu à l'appel de la MSH-M avec la proposition de programme suivante :

> L'objectif du programme *sud4science Languedoc-Roussillon (LR). Mutation des pratiques scripturales en communication électronique médiée,* http://www.msh-m.fr/programmes-2011/sud4science-lr/, est de contribuer à l'étude de la communication médiée par SMS et à l'étude du/des « genre(s) langagier(s) » qu'elle véhicule, à partir de téléphones mobiles. Il s'agit de récolter 30 000 SMS de la région LR (phase 1) puis, une fois le corpus constitué, de l'anonymiser, le transcoder, et l'annoter (phase 2). Un travail d'analyse ultérieur (phase 3) sera entrepris par des chercheurs pluridisciplinaires, afin d'étudier ces fonctionnements langagiers mouvants et dynamiques. (Panckhurst, 2011).

Nous avons obtenu le financement du projet *sud4science LR* pour l'année 2011, puis pour une année de reconduction supplémentaire. Dès janvier 2011, nous avons démarré une série de séminaires scientifiques et de journées d'étude, car il s'agissait essentiellement de « subvention à recherche entrante », pour des chercheurs se déplaçant à Montpellier (*cf.* § 3.4 Réseaux, diffusion et valorisation). De janvier 2011 à décembre 2012, nous avons donc consolidé un réseau de chercheurs, accueilli des chercheurs à Montpellier et tissé de nouveaux liens scientifiques (http://www.msh-m.fr/programmes-2011/sud4science-lr/ ; vidéos

consultables en ligne de certains de nos séminaires scientifiques : MSH-M.TV : http://msh-m.tv/spip.php?rubrique138 itunes : https://itunes.apple.com/fr/itunes-u/sud4science-languedoc-roussillon/id594173977?mt=10).

Pendant cette première année du projet *sud4science LR*, j'ai présenté une étude menée en parallèle avec Claudine Moïse, sur un corpus d'échanges de SMS, recueillis par deux étudiantes de Master (L. Fabre et Ravel 2011), dans une session intitulée : « Textos : dimensions linguistiques et pragmatiques » au 79e colloque de l'*Acfas (Association francophone pour le savoir) : curiosité, diversité, responsabilité, qui s'est tenu* à Sherbrooke, Canada, en mai 2011. Cette communication (Panckhurst et Moïse, 2011), exposait une étude exploratoire sur la dimension interactionnelle des SMS. Des grandes récoltes mondiales de SMS, que j'intitule, « isolés » (Fairon et al. 2006b,c), http://www.sms4science.org, fournissaient un formidable terrain d'investigation. Mais, de notre côté, nous voulions aussi étudier les SMS, que je nomme « conversationnels », à savoir ceux constitués de séquences d'échanges, parfois nombreuses, entre deux (ou plusieurs) interlocuteurs.

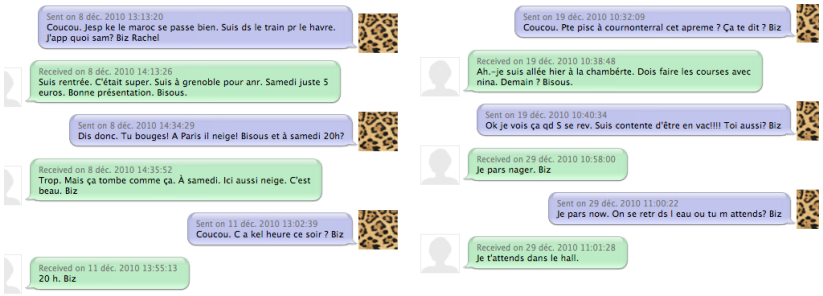


Figure 3.14 – SMS « isolés » et « conversationnels » [95]

Il s'agissait de repérer les marques interactionnelles et pragmatiques (ouverture et clôture, relance, mots du discours, notamment) des SMS et d'en décrire les marques propres « d'expressivité » (Moïse et Schultz-Romain 2010), dans un cadre théorique s'étendant de Bakhtine 1929 à (Develotte et al. 2011). [96].

95. La figure 3.14 est publiée dans (Panckhurst et Moïse 2014), p. 157.
96. Le cadre théorique adopté était celui en collaboration avec Claudine Moïse : une conception

Les récoltes officielles auprès du grand public, dans le cadre du projet *sms4science*, avaient permis de recueillir des SMS envoyés par les donateurs, mais pas ceux reçus par les donateurs, et ce pour des raisons de droit [97]. Le corpus recueilli par (L. Fabre et Ravel 2011) a été constitué à partir d'échanges entre les étudiantes et leurs entourages utilisant (majoritairement) des téléphones intelligents dotés de forfaits SMS illimités. Je résume quelques points de notre étude initiale ci-dessous, car, d'une part, les actes de la conférence (Panckhurst et Moïse, 2011) ne sont pas parus [98], puis, d'autre part, mes recherches ultérieures n'ont pas porté sur ces aspects interactionnels (à l'exception d'une section dans (Panckhurst et Moïse 2014, p. 157-160), qui reprend et approfondit une partie de cette étude). Dans un corpus de cette nature, le cadre de l'interaction (contextes temporel et spatial, rituels conversationnels, enjeux ou stratégies des locuteurs, émotions), puis l'analyse des interactions nous interpellent (*cf.* figures 3.15 et 3.16.)

Le contexte permet évidemment de situer l'échange (Marcoccia 2011) :

> **Référence au lieu : thème de l'échange.**
> A [99]. Je suis en bas de chez toi :)
> B. Ok
> **Référence au lieu : repère interactionnel**
> C. J'espère que t'auras encore de quoi manger tout à l'heure parce que là, y a déjà pas grand-chose !: s
> C. Je suis au fond de la salle (toujours la même)... À tout'

Les téléphones intelligents permettent de voir les échanges sous forme de fil de discussions. Ainsi le locuteur ne conçoit pas obligatoirement son message comme

actionnelle du langage et sociopragmatique (Austin 1962, Searle 1972, 1982) ; la construction interactive du sens (Bakhtine [1929] 1977) ; l'analyse conversationnelle (Sacks, Schegloff et Jefferson, 1974 [1978] ; Kerbrat-Orecchioni 1990, 1992, 1995, 2005 ; (Develotte et al. 2011)) ; les théories de la politesse (Brown et Levinson 1987) ; la préservation des faces (Goffman 1959). À l'exception de (Develotte et al. 2011), ces références ne sont pas incluses dans la bibliographie générale de fin de volume.

97. *Cf.* la CNIL. Il n'est pas possible qu'une personne fasse don des SMS qu'elle a reçus ; comme ces SMS ne lui appartiennent pas, elle ne peut pas décider d'en faire don.

98. Tous les exemples suivants que je cite et analyse sont extraits de (L. Fabre et Ravel 2011). Je remercie Claudine Moïse d'avoir accepté que je fasse figurer des extraits de ce travail mené en commun ici.

99. Les locuteurs sont indiqués/différenciés par des lettres.

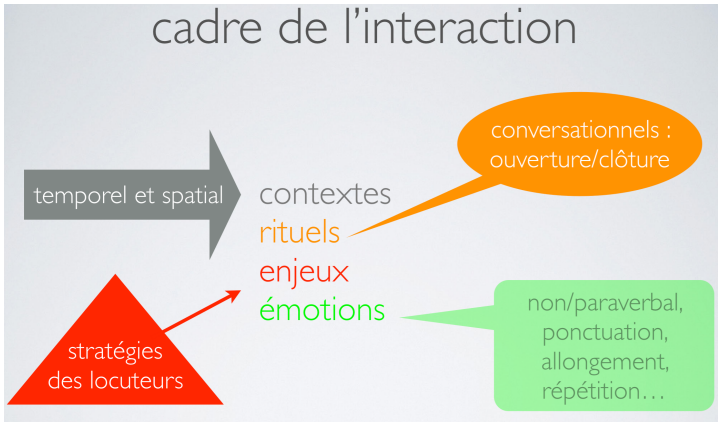


Figure 3.15 – Cadre de l'interaction

une nouvelle prise de contact mais plutôt comme une suite à la discussion précédente. Les interlocuteurs peuvent donc avoir l'impression que la conversation est toujours ouverte, et de ce fait les échanges de politesse ne sont pas obligatoirement nécessaires pour réenclencher le contact. Les ouvertures/clôtures au sein des rituels montrent ceci. Les clôtures (*Gracias la madre, bonne semaine*) sont beaucoup plus fréquentes que les ouvertures [100].

Les enjeux (stratégies) des locuteurs (histoire interactionnelle, thème interactionnel ou *topic*, contrat interactionnel, position hiérarchique) sont importants à cerner. L'histoire interactionnelle correspond à ce qui lie les locuteurs entre eux (implicites, sous-entendus, etc.). L'échange n° 1 ci-dessous est un exemple de ce lien entre locuteurs. Les SMS se passent ici dans l'espace privé. Ainsi, comme une conversation en présentiel, l'histoire interactionnelle est forte et il y a moult implicites, car on connaît l'interlocuteur.

Échange n° 1 : appréciatif

1. **A.** Félicitation ma jolie ! Meme pas tu me le dirais !! Jsui vmt contente pr toi. D gbisous !! Et à bientôt !

100. Cela est également vrai au sein du corpus *88milSMS*. Des clôtures, dont la variation est riche, (*à demain, à toute, bisous, bonne soirée…*) sont beaucoup plus présentes que les ouvertures (*Bonjour, bjr, hey, coucou, cc, ça va…*) : 75% de clôtures vs. 25 % d'ouvertures.

2. **B.** Ouiii j'allais t'envoyer un SMS !!! merci ma belle ! Tu vois tout est possible faut pas te décourager ! Prend ton tps pour te praparer ! Gros bisoux a vous deux a bientôt
3. **A.** Je plaisantai !mais tt se sait a leucate ! Et ki était l'inspecteur ?! mdr Gbisous a vs 2 aussi !
4. **B.** Comme d'hab !! Mais Elle était de meilleure humeur ! J'ai fait la route de gruissan ! Bonne soirée !!!! Bizzzz

Pour ce qui concerne les thèmes ou *topics*, soit le thème sera unique (la question sera suivie d'une réponse de l'interlocuteur : *Als ça à donné koi l'AG ?*), soit on aura des thèmes/interrogations « en rafale » (Fairon et al. 2006b,c) : *Coucou L :)) cmt tu vas ? quoi de beau ? Cmt se passe les cours ? Faudra se faire une soirée ! F est la ? Au fait scoop jsui en coupe lol te raknterai ! Dis moi taurai le tel a C ? G a pas sn tel et elle connaît pas sn nim ! Biyoux*), ce qui va jouer un rôle important dans la grammaire des interactions [101].

Les émotions sont « injectées » dans les échanges par SMS, notamment par les aspects sémiologiques non-verbaux (binettes/émoticônes/smileys/emoji), des répétitions de signes de ponctuation *(!!!)*, des reprises de phonèmes (*ouiii*) qui simulent l'intonation, donc de l'infomation paraverbale, des interjections (*hey*) aussi, des mots du discours (*lol*), des onomatopées (*grrr !*), etc.

Dans le passage des SMS isolés, aux SMS conversationnels, il s'agit de comprendre ce qui est exprimé dans les SMS précisément en fonction des échanges. Au niveau de la grammaire de l'interaction, cela implique l'étude des tours de parole, ou d'*écriture* (Marcoccia), des thèmes, et des catégorisations (qualifications). Dans cette étude, ma collègue et moi avons focalisé notre attention sur les deux premiers. À chaque fois, il s'agissait de comprendre comment se mettent en place la *validation* (*ok*), la concession (*je suis d'accord, mais…*), la *réfutation* (*non*) et la *réparation* (*excuse-moi*) au sein des échanges. Au niveau des types discursifs, les informatifs et les appréciatifs semblent majoritaires. Pour les effets pragmatiques, il s'agit surtout de la façon d'interpeller (l'adresse), ainsi que

101. Je n'évoquerai pas le contrat interactionnel (c'est à dire le cadre interactionnel, qui est contraint par le contexte, car il s'agit, dans ce corpus, d'échanges entre amis ou entre membres d'une même famille). Le positionnement hiérarchique, dans un contexte professionnel, par exemple, n'entrera pas en ligne de compte non plus ici, étant donné le contexte évoqué.

l'utilisation (ou l'absence) de modalisateurs et les actes perlocutoires (menaçants directs, par exemple). Enfin, est mentionné « Lol », en tant que mot du discours, dans les procédés discursifs.

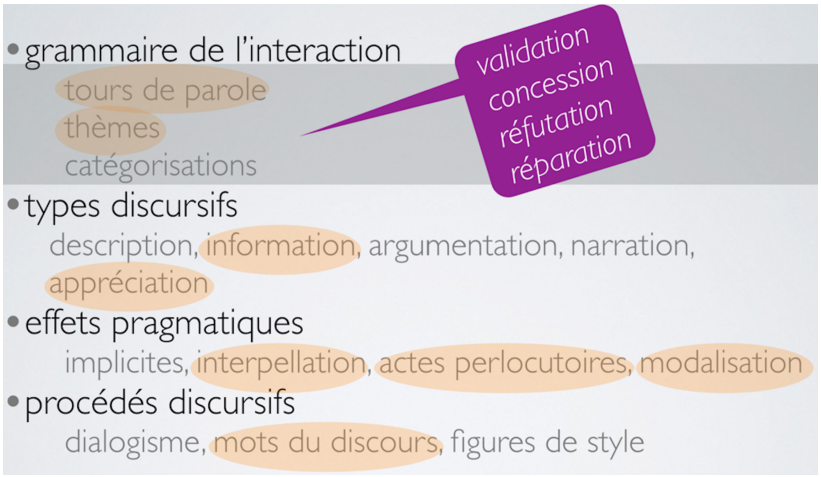


Figure 3.16 – Analyse des interactions

Si l'étude des SMS permet aux différents domaines relevant de la linguistique théorique et au-delà (phonétique/phonologie, morphologie, lexique, syntaxe, morphosyntaxe, sémantique, stylistique, pragmatique…) de se côtoyer, (Marty 2005), citant (Anis et al. 2004), rappelle les contraintes imposées et le dépassement du cadre purement linguistique :

> Le modèle proposé [par Anis] part des contraintes imposées : *linguistiques* (concision ; rapport avec l'oral ; dialogisme ; rapidité ; jeu avec les normes) mais aussi *techniques* (taille de l'écran ; nombre de caractères), *communicationnelles* (communication rapide et discrète), *psychosociales* (intimité ; langage du groupe) et *économiques* (coût : moins cher qu'un appel vocal).
>
> (*cf.* aussi (Marty 2005, p. 65), in (Panckhurst 2009, p. 38)).

À cause, entre autres, de ces contraintes, des bouleversements existent, si on compare cela à des situations de conversation orale. Dans l'exemple de SMS conversationnel fourni *supra* (*cf.* Échange n° 1,), les chevauchements sont absents. Contrairement à des chats synchrones, dans lesquels figurent typiquement

des problèmes de séquentialité (décalage entre le moment où on écrit, et le moment où le message apparaît à l'écran), en général, ce problème ne se pose pas pour des SMS asynchrones. La gestion des tours de paroles/écriture, *floor-taking*, (Crystal 2001, p. 33), n'est pas non plus problématique. En revanche, on rencontre beaucoup de situations dans les SMS où les paires adjacentes (*adjacency pairing*) seront absentes, ou vraiment bouleversées : questions/réponses au sein d'un même SMS (*t'a été à la foir ? moi oui*...), absences de réponse, de reprise, etc.

Afin d'étudier les types discursifs, on a distingué trois exemples : l'échange n° 2 qui est plutôt de type discursif *informatif*, l'échange n° 1, *appréciatif* et l'échange n° 4, qui constitue un mélange des deux.

Dans l'échange n° 2, figure un thème unique initial (le travail de la mère), suivi d'un deuxième thème unique (le magasin de jeu) puis un dernier thème (l'imprimante). Cet échange est donc séquentiel, ce qui est loin d'être systématiquement le cas.

Échange n° 2 : *informatif*

12h49

5. **B**. Tu sais si maman travaille cet aprem ?
6. **C**. Oui elle taff

13h17

7. **B.** Ok
8. **C.** Ya un magasin de jeu d'occasion à polygone ?
9. **B**. Yes Game & micromania
10. **C.** Tu sais si ces chère ou pas ?
11. **B**. Ça dépend de ce que tu veux Moi j'ai acheté des sims a 5 – 10€
12. **C**. ok on ira voir samedi :)
13. **B**. Ils prennent les jeux aussi Si tu veux en vendre
14. **C.** Cool je voulais en vendre en plus :)
15. **B**. ok regarde sur Internet pour voir leur site s'il donne des prix
16. **C**. Ces quoi le nom du magasin ?
17. **B**. Game Micromania
18. **C.** Ok merci :)

19. **B**. tu rentres a quelle h ?

    13h31

20. **C.** Je suis à la maison la j'ai pas cours cette aprem.
21. **B.** Ok bein F va passer pour déposer une imprimante qu'il faudra me porter samedi Essaye la Pour voir si les cartouches fonctionne
22. **C.** Imprimante ordi ?
23. **B.** Oui Une comme la notre
24. **C.** ok pas de souci

L'échange n° 3 suivant contient des changements marqués, intercalés : « allez moi jme bouge au volley », et « ah au fait faut que… ».

Échange n° 3

33. **D**. Oki.Pense à faire la perspective de la feuille pr l'étape du triangle dure à expliquer.Allez moi jme bouge au volley.Bisous à demain. […]
45. **D** Mdr oki.Tu es notre œil ^^. Ah au fait faut que je vous envoie les polycop de Mme B.Je fais ça après.Bisous.Passe une bne soirée. […]

Dans l'échange n° 2, des injonctions non modalisées sont utilisées, contrairement à l'oral, qui préfère souvent les modalisateurs : *« regarde sur Internet » =>« si tu veux bien regarder sur Internet » ; « Essaye la » =>« tu peux l'essayer »*. Les interactions y sont directes, sans implicites. Ce sont des routines conversationnelles établies, ici, entre frère et sœur.

En revanche, les échanges n° 1 et 4 sont beaucoup plus appréciatifs : au niveau du cadre de l'interaction, il y a de nombreuses variations en clôture (*1A. D gbisous !! Et à bientôt / 2B. Gros bisoux a vous deux a bientôt / 3A. Gbisous à vs 2 aussi / 4B. Bonne soirée !!!!! Bizzzzz*), et l'histoire interactionnelle est bien établie (*1. A. Félicitation ma jolie ! Meme pas tu me le dirais !! Jsui vmt contente pr toi. D gbisous !! Et à bientôt !*). On y trouve, au niveau des effets pragmatiques, un acte de félicitation et des termes d'adresse (*1A. Félicitations ma jolie ! 2B. Merci ma belle !*). Les actes menaçants renvoient à une réparation (explication-justification) (*Meme pas tu me le dirais !!, j'allais t'envoyer un SMS !!!, Je plaisantai*). Cela est également présent ci-dessous entre 106B et 107F.

Échange n° 4 : *mélange informatif & appréciatif*

105. **F.** Coucou L :)) cmt tu vas ? quoi de beau ? Cmt se passe les cours ? Faudra se faire une soirée ! F est la ? Au fait scoop jsui en coupe lol te raknterai ! Dis moi taurai le tel a C ? G a pas sn tel et elle connaît pas sn nim ! Biyoux
106. **B.** Salut ça va bien. Non je suis seule. Je croyais que vous m'avez oubliée !!! Cool je vois qu'on a des trucs a se raconter !!!! Non j'ai pas le num a C ! Dsl bonne journée, a bientôt ! Bizz
107. **F.** Lol nan tkt on ta pas oublié c juste quentre les cour et ma nouvelle relation jai pas trop trep de temps lol ^^ et ui c clair on a plein dtruc a se dire lol :)) demain jai pas cours on peut se voir si tu veux :))
108. **B.** J'ai cours moi c'est ma journée CM ! LOL je rentre ce w-end mais F vient la semaine prochaine, on se fait une soirée !!!! Good aprem ! Bizzz

Dans la continuation de cet échange n° 4, au sein de 113B, ci-dessous, des adoucisseurs apparaissent : « LOL » amorce le message et « mdr » le clôture, alors qu'en 107F et 108B, on est plutôt en présence de modalisateurs et de ponctuants (connecteurs).

113. **B**. LOL laisse tomber les stages avec les grèves ils vont pas te prendre ! Mdr je rigole Je sais pas peu être faire un tour vers 15h sur la com pour voir l'ambiance
114. **F**. Lol cmt t mechante mdr jme vengerais : p Jte tiens au courant si je change davis :))
115. **B**. Ça marche ! Je rigole tu as raison ! Moi je suis au chomage des cours donc ! LOL Je suis a l'AG ! C'est fun tout ces cris ! Mdr

D'autres types d'échanges figurent bien entendu au sein de ce corpus étudiant, notamment au niveau *métalinguistique* :

166. **D**. Bon je v faire population et jarive
167. **I**. Tu va koi ?
168. **D**. Faire pipi mdr.C'est à cause du correcteur d'orthographe.

ou simplement, la conversation « ordinaire » :

57. **B**. J'arrive ce soir
58. **E**. Et tu repar ?
59. **B**. Jeudi proch

Dans la conclusion de l'étude, Claudine Moïse et moi avons annoncé vouloir approfondir, au sein de la grammaire des interactions/du discours, les questions/réponses en rafale, ou en décalage, puis tout l'enchaînement apparaissant dans la figure 3.17 (*cf.* aussi (Panckhurst et Moïse 2014, p. 158)).

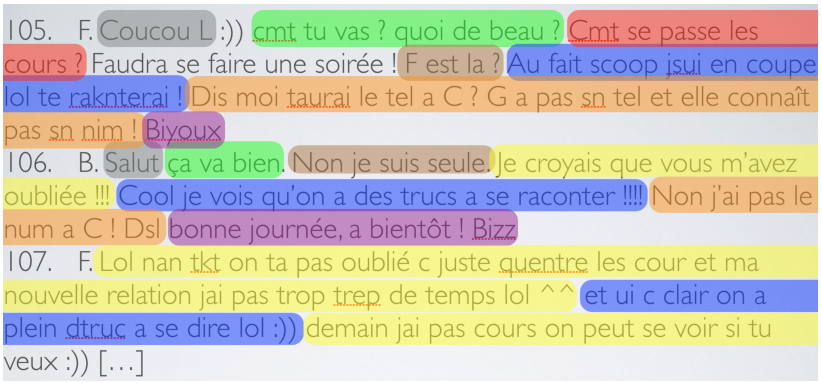


Figure 3.17 – Questions/réponses « en rafale » dans le corpus de (L. Fabre et Ravel 2011).

Jusqu'il y a peu, une collecte conversationnelle auprès du grand public n'avait pas encore eu lieu, donc je n'avais pas creusé ces aspects. Une initiative récente de collecte, actuellement en cours, par le CENTAL (http://www.vospouces.org/) permet de renouveler les espoirs. [102].

102. Le descriptif de ce projet, très prometteur, se trouve ici : http://www.vospouces.org/about « *Vos Pouces pour la Science (thumbs4science)* est un projet de collecte chapeauté en 2016 par l'université catholique de Louvain (UCL) et financé par le Fonds national de la recherche scientifique (FNRS). Il est issu d'une collaboration entre 3 domaines scientifiques : la sociolinguistique, le traitement automatique du langage et la psychologie. Il se situe dans la prolongation du projet *Faites don de vos sms à la science* (sms4science, http://www.sms4science.org/) qui avait pour but de collecter des sms qui ont servi de matériau de base pour la recherche scientifique.

Le projet *Vos Pouces pour la Science* vise à étudier les compétences de la population au contact des nouveaux médias de communication en général (réseaux sociaux, messages électroniques...).

En attendant que la collecte soit terminée et que les analyses puissent démarrer, je me suis orientée vers des applications en TAL avec mes collègues informaticiens, Mathieu Roche et Cédric Lopez, que j'évoque plus bas, (§ 3.3.3.11).

3.3.3.4 Projet SMS montpelliérain

Quant au problème de la collecte de SMS « isolés » à Montpellier : comment pouvait-on faire pour lancer notre propre collecte ? Afin de nous sortir de l'impasse avec les téléphones opérateurs, j'ai décidé qu'à Montpellier nous procéderions autrement. Les forfaits en France évoluaient et, en 2011, beaucoup incluaient les SMS illimités ; une aubaine pour l'équipe de chercheurs que nous étions car les donateurs de SMS étaient en mesure de nous les faire parvenir sans surcoût. Donc, j'ai pensé que nous pourrions effectuer le recueil directement à partir d'un smartphone et j'ai pris contact avec une entreprise informatique montpelliéraine qui était enthousiasmée par l'idée.

Contrairement aux idées reçues et véhiculées dans la presse et dans les milieux de l'éducation, nous partons de l'hypothèse que l'utilisation des nouveaux médias chez les classes d'âge de plus en plus jeunes ne mène pas à une incompétence linguistique (baisse du niveau de l'orthographe, méconnaissance des règles grammaticales…), mais plutôt à une "pluricompétence" qui amènerait chaque locuteur à jongler avec son code à chaque changement de situation, d'interlocuteur et de médium de communication. Et de telles pratiques sont-elles liées au "potentiel créatif" de chaque locuteur ?

Nous enquêtons également sur les compétences sociales et psychologiques de la population au contact de ces nouveaux médias. Les nouveaux médias nous ont-ils contraints à réinventer nos modes de communication (comment gère-t-on les silences - les réponses qui n'arrivent jamais -, les personnes en présence physique, les personnes en présence médiatique) ? La société MoodWalk, spécialiste dans le domaine de la science comportementale nous aide à répondre à ces questions, via un test de quotient émotionnel qui dresse une carte personnelle pour chaque participant.

Sommes-nous en train de réinventer notre mode de communication au travail ou en classe ? Existe-t-il une compétence émotionnelle qui nous permet de mieux gérer verbalement nos rapports aux autres, notre "popularité médiatique" ainsi que des situations embarrassantes ou fâcheuses sur les réseaux sociaux ? Quelles sont les raisons qui poussent les personnes à utiliser ces messageries dans le cadre de leur travail ? Dans quelles situations les messageries privées sont-elles utilisées au travail ? En quoi l'outil facilite-t-il ou contraint-il les interactions ?

Autant de questions auxquelles vous nous aidez à répondre ! »

> Le projet montpelliérain s'est distingué des collectes précédentes par la méthode de récolte. Après inscription et consentement légal en ligne, les participants donateurs de SMS, au moment de l'envoi de leur texto à autrui, pouvaient l'envoyer en copie aux chercheurs. Il était également possible de réexpédier des SMS (précédemment envoyés et contenus dans la mémoire du téléphone du scripteur) aux chercheurs. Le moyen utilisé ? Un smartphone [103], grâce auquel l'ensemble des textos a été recueilli pendant 13 semaines. Ce dispositif a été un véritable pari technique, car personne ne savait à l'avance si l'iPhone allait permettre un recueil de SMS très important, sans défaillir. En définitive, aucun problème n'est survenu. Chaque semaine, les SMS ont été copiés sur un disque dur externe, déconnecté d'Internet (pour des raisons juridiques). La grande base de données (BD) en constitution devait rester dans son intégralité sur le téléphone également jusqu'à la fin de la récolte, afin d'assurer que la BD soit entière et homogène, avant transfert final. Depuis le début des collectes de SMS en 2004, il était très important que la méthode de recueil ne passe pas par une (re)saisie des données : seul ce type de transfert technologique était possible afin d'assurer que les données demeurent réellement authentiques. (Extraits de (Panckhurst et al. 2014b, p. 22-25)).

Grâce à notre mise en place de réseau scientifique (*cf.* § 3.4 Réseaux, diffusion et valorisation) et aux liens scientifiques solidement tissés avec mes collègues linguistes et informaticiens : Catherine Détrie (Praxiling, UM3), Cédric Lopez (Viseo, Grenoble), Claudine Moïse (Lidilem, université Grenoble-Alpes), Mathieu Roche (Tétis, Cirad), Bertrand Verine (Praxiling, UM3), nous avons décidé d'embarquer ensemble dans l'aventure de la collecte, de l'anonymisation et de l'analyse de SMS authentiques.

Ci-dessous, j'indique quelques points importants extraits des publications autour de nos projets SMS : anonymisation, transcodage-alignement [104], annotation, analyses (socio)linguistiques et en TAL, mais pour ne pas alourdir cet écrit,

103. L'entreprise iTribu (http://www.itribustore.fr/), très enthousiaste à l'idée de participer à un projet de recherche universitaire, a prêté un iPhone aux chercheurs pour la durée de la collecte.

104. D'un commun accord, nous avons choisi, pour le projet *sud4science LR*, le terme « trans-

j'inviterai le lecteur à se reporter aux références indiquées pour consulter le travail plus approfondi [105].

Descriptif des projets autour du corpus *88milSMS*. Le lecteur pourra se reporter à http://88milsms.huma-num.fr/references.html pour toutes les publications concernant le corpus *88milSMS*.
La référence originale du corpus *88milSMS* (http://88milsms.huma-num.fr/) est :
(Panckhurst et al. 2014a) Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014), « *88milSMS*. A corpus of authentic text messages in French » produit par l'université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8
Suite à sa mise sur *Ortolang* en 2016, une deuxième version du corpus est référencée comme suit :
(Panckhurst et al. 2016a) Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2016) « 88milSMS. A corpus of authentic text messages in French », (version nouvelle du corpus ISLRN 024-713-187-947-8). In (Chanier 2016) Chanier T. (ed) Banque de corpus CoMeRe. Ortolang : Nancy. [cmr-88milsms-tei-v1 ; https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1]
Les références générales (publiées) concernant le projet sont : (Panckhurst 2013, 2016a ; Panckhurst et al. 2013 ; Panckhurst et Moïse 2014 ; Panckhurst et al. 2016b).
À cela s'ajoutent deux articles d'« explication » : (Panckhurst et al. 2014b,c)

codage » qui semblait mieux convenir que « transcription » habituellement réservé pour un traitement de l'oral.

105. Je remercie mes co-auteurs (Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, Bertrand Verine) de m'avoir donné leur accord pour que je reproduise des extraits (parfois conséquents) de nos publications dans le cadre de cette habilitation à diriger des recherches.

> Une équipe pluridisciplinaire de linguistes et d'informaticiens (Rachel Panckhurst, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, Bertrand Verine [...]) a recueilli plus de 88 000 SMS authentiques en français à Montpellier, en 2011. Cette collecte a été effectuée dans le cadre du projet *sud4science LR* (http://sud4science.org, *Sud4science Languedoc Roussillon. Mutation des pratiques scripturales en communication électronique médiée* (financement principal : MSH-M)), lui-même faisant partie du projet international *sms4science* (http://sms4science.org), coordonné par le CENTAL à l'université catholique de Louvain (UCL) en Belgique. Lors du recueil des SMS, un questionnaire sociolinguistique a également été proposé aux participants. Les SMS du projet *sud4science LR* ont été ensuite *anonymisés* de manière semi-automatique (en collaboration avec des étudiants stagiaires et un juriste-CIL, Nicolas Hvoinsky, DAJI-université Paul-Valéry), puis partiellement *transcodés* (en français standardisé) et *annotés* (*cf*. Panckhurst *et al*. 2013).
> Les analyses portent sur l'écriture, la textualité numérique, les fonctionnements langagiers, les pratiques et usages des donateurs de SMS ont été menées dans le cadre de *sud4science LR* et de deux projets supplémentaires : *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS* (financement principal DGLFLF) et *Analyse contrastive des**é**motions **co**ntenues dans les**mess**ages courts* (PEPS CNRS ECOMESS (HuMaIn [106])).
> Cette recherche pluridisciplinaire a permis de réaliser un très grand nombre de travaux et de publications et d'encadrer 8 stagiaires étudiants. Les résultats de recherche ont été diffusés en France et dans 9 pays étrangers (Belgique, Canada, Espagne, Finlande, Grèce, [Islande], Maroc, [Pays Bas], Suisse) [107]. Les chercheurs ont pu offrir l'occasion aux étudiants-stagiaires de *présenter* leurs propres travaux et d'écrire, pour deux d'entre eux, en tant que *premier auteur*, [un chapitre dans ouvrage et] un article dans une revue internationale (*cf*. (Accorsi et al. 2014 ; Patel et al. 2013). (Extrait du fichier intitulé « 88milSMS.pdf », consultable et téléchargeable sur http://88milsms.huma-num.fr/)

Afin de bien comprendre le déroulement des deux projets principaux (*sud4science* et DGLFLF), voici un tableau récapitulatif des principaux points saillants :

Tout au long des projets, l'équipe que nous formions a été soutenue et accompa-

106. Je n'évoque pas ce projet (porté par Mathieu Roche, Cirad), dans le cadre de cette habilitation.

107. *Cf*. § 3.4 Réseaux, diffusion et valorisation, pour une étude plus approfondie sur la diffusion de *88milSMS*, depuis la mise à disposition du corpus.

Tableau 3.7 – *Points saillants de chaque projet*

sud4science LR (1/1/2011-31/12/2012)	DGLFLF (1/7/2012-30/6/2013)
documents légaux/site Web/questionnaire	*anonymisation* (suite et fin : 30/9/2013)
recueil (de données SMS)	déplacements (colloques, réunions)
accueil (de chercheurs à Montpellier)	séminaires scientifiques étudiants
séminaires scientifiques filmés	*transcodage* : projet 6 étudiants + 1 stage d'un mois (juin 2013)
journées d'étude	analyses (socio)linguistiques
visioconférences	médias (presse écrite)
médias (télévision, presse écrite, radio)	publications (revues, Actes)
conception d'un programme informatique	préparation du cadre légal pour la diffusion du corpus
anonymisation (début)	17 mois stagiaires étudiants gratifiés*
13 mois stagiaires étudiants gratifiés*	

*8 mois stagiaires sont partagés entre les deux projets (entre juillet et décembre 2012). Du 1/1/2011 au 30/6/2013, 22 mois stagiaires étudiants au total ont été nécessaires afin de mener à bien les étapes de lecture, d'anonymisation (corpus total) et de transcodage partiel (1 000 SMS). (Extrait du rapport scientifique remis à la DGLFLF, juin 2013, p. 4).

gnée à bon escient par notre *Direction des Affaires Juridiques et Institutionnelles* (DAJI, Directrice, Stéphanie Delaurany), et, plus précisément, par notre juriste-CIL, Nicolas Hvoinsky [108], et ce malgré les aspects légaux particulièrement sensibles, à cause du type de corpus recueilli et des données relevant de la sphère privée, à anonymiser.

> Après inscription et consentement légal en ligne (www.sud4science.org), les participants à la collecte — que nous appelions des *donateurs* de SMS, à la suite du slogan initial belge « Faites don de vos SMS à la Science » — envoyaient leur texto à autrui, en mettant les chercheurs en copie, ou bien transféraient aux chercheurs un texto (précédemment envoyé à un tiers) qui était présent dans la mémoire de leur téléphone. Pour des raisons légales, seuls les SMS envoyés (et non reçus) ont été recueillis. Un questionnaire sociolinguistique pouvait également être rempli. [L]'aspect juridique est primordial. Une collecte de données qui passe outre la règlementation en vigueur risque de produire des données inexploitables pour des raisons légales. Rappelons que les SMS constituent des données personnelles, sensibles. La vie privée doit donc être protégée. Nous avons choisi d'associer, dès le départ de notre projet, [la direction] des affaires juridiques et institutionnelles de l'université Paul-Valéry Montpellier 3 (DAJI) par l'entremise de sa directrice, la juriste Stéphanie Delaunay, et de son juriste correspondant informatique et libertés, CIL, Nicolas Hvoinsky.
>
> **Parfois, des universitaires recueillent des données, des années durant, pour**

108. Nicolas Hvoinsky a également assuré le lien avec le service juridique du CNRS.

> **se rendre compte trop tard que celles-ci sont inutilisables, car la collecte n'a pas respecté les normes juridiques.** (Je souligne, (Panckhurst et al. 2014c), consultable en ligne : http://www.bulletin.auf.org/index.php?id=1875)

De même, les chercheurs du CENTAL, UCL (notamment Louise-Amélie Cougnon, Cérick Fairon et Hubert Naets) nous ont soutenus — sans faille — tout au long de nos projets SMS, du double point de vue logistique et technique [109].

3.3.3.5 Anonymisation

Ce paragraphe est extrait de (Panckhurst et al. 2013) et du rapport scientifique remis à la DGLFLF en juin 2013. Pour une étude plus approfondie, concernant notamment les traitements en TAL de l'anonymisation, le lecteur pourra également se reporter à (Accorsi et al. 2014 ; Patel et al. 2013).

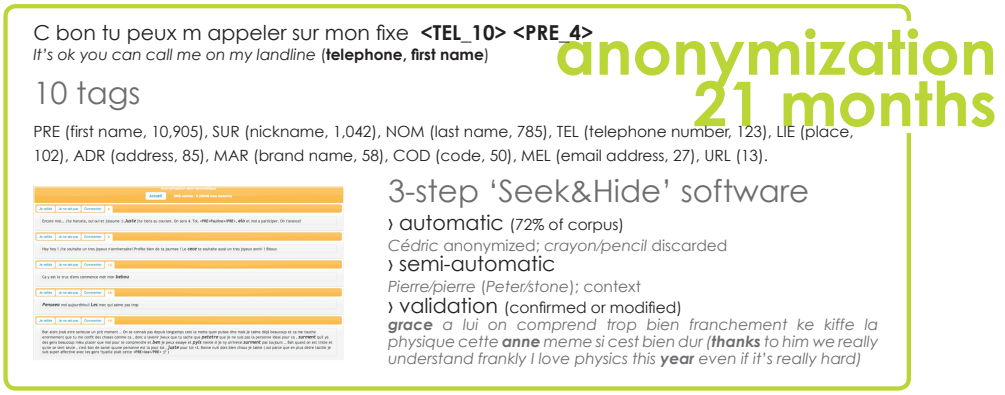


Figure 3.18 – Anonymisation de *88milSMS* (21 mois)

Dans le but de masquer l'identité d'un individu, l'anonymisation se révèle une tâche indispensable, par exemple dans le domaine juridique (Plamondon *et al.*,

109. Notre site web http://www.sud4science.org/ est hébergé sur un serveur du CENTAL, à l'UCL, en Belgique.

2004) ou médical (Grouin *et al.* 2009). Dans ces domaines, les systèmes reposent principalement sur la reconnaissance automatique des noms, des dates, des lieux et d'autres éléments qui peuvent conduire à l'identification des personnes. Généralement les méthodes de reconnaissance de ces types d'entités nommées [110] s'appuient sur des règles spécifiques et l'utilisation de dictionnaires. De plus, des méthodes d'apprentissage supervisé peuvent être appliquées. Par exemple, (Szarvas *et al.*, 2007) ont entraîné plusieurs classifieurs afin de proposer une fonction de prédiction combinant les résultats pour une tâche d'anonymisation. Une limite essentielle de ces méthodes est liée à la nécessité de disposer d'une quantité importante de données étiquetées.

Nous considérons qu'un tel processus d'anonymisation ne peut être entièrement automatique. Suivant cette même hypothèse, les travaux de (Reffay *et al.*, 2012) se focalisent sur la création d'une interface par laquelle l'expert peut identifier les données personnelles et décider si elles nécessitent d'être anonymisées. Le logiciel d'anonymisation de SMS que nous proposons repose sur le même principe, en sachant que nous cherchons à faciliter le travail de l'annotateur par une procédure automatique. Dans de telles situations, les marqueurs identitaires des SMS doivent être anonymisés [111]. Par exemple, le SMS « G pas encore de rep de sab ! [...] », nécessite l'anonymisation du mot « *sab* ». Ici, la difficulté majeure réside dans le fait que ce marqueur est atypique et ne représente pas un prénom trivial, c'est-à-dire appartenant à un dictionnaire de prénoms et/ou de surnoms.

Le logiciel d'anonymisation *Seek&Hide,* développé par deux étudiants en informatique [112], s'appuie sur des méthodes de TALN (Traitement Automatique du Langage Naturel). Il propose une page web sécurisée accessible pour les annotateurs. Le but du logiciel est de faciliter l'expertise et de traiter une quantité importante de données. L'approche développée se décline en trois phases :

Phase automatique : traitement automatique des données (mots) qui ne présentent *a priori* aucune ambiguïté quant à leur interprétation (à anonymiser ou non).

110. Pour une approche multilingue récente sur les entités nommées, on pourra se référer, entre autres, à *NERosetta* (Krstev et al. 2016).

111. Notre [juriste-CIL] a rempli une déclaration exigeant que l'anonymisation de la totalité du corpus de SMS*sud4science* ainsi que des données émanant du questionnaire soit effectuée avant le 30/9/13.

112. Pierre Accorsi et Namrata Patel (étudiants en Master Informatique) ont développé le logiciel d'anonymisation *Seek&Hide* pendant un stage de deux mois en 2012 (Accorsi et al. 2014 ; Patel et al. 2013).

Phase semi-automatique : traitement manuel de l'information nécessaire pour les SMS qui présentent des mots ambigus ou inconnus. Ceci s'effectue à travers un système qui met en relief les éléments nécessitant une expertise. Cette mise en valeur facilite significativement le travail de l'annotateur.
Phase de validation : relecture et validation des SMS anonymisés automatiquement ou suppression d'une anonymisation incorrectement appliquée par l'outil lors de la phase automatique.
La phase automatique
De manière concrète, le traitement des données textuelles se décline en trois étapes successives que nous synthétisons ci-dessous :
A) Pré-traitement des données. La première étape consiste à segmenter le corpus en mots.
B) Identification des mots susceptibles ou non d'être anonymisés. Dans cette phase du processus automatique, chaque mot d'un texte peut demander d'être anonymisé (AA : à anonymiser) ou non (NPA : ne pas anonymiser). Pour cela, nous utilisons deux types de dictionnaires : un « Dictionnaire » qui contient des mots qui doivent être rendus anonymes. Le dictionnaire que nous utilisons est constitué d'une liste de prénoms ; un « Anti-dictionnaire » qui contient des mots qui ne nécessitent pas d'anonymisation. Cet anti-dictionnaire est issu de la fusion de différentes ressources : lexique des formes fléchies du français (Lefff [113], (Sagot 2010)), dictionnaire de certaines formes récurrentes utilisées dans l'écriture SMS (par exemple, les binettes, certaines abréviations, etc.), dictionnaire de noms de lieux.
C) Traitement des mots à anonymiser. Chaque mot est traité en vérifiant son appartenance aux différents dictionnaires et anti-dictionnaires. Quatre situations sont rencontrées comme l'illustre le tableau 3.8.
Le mot *Rachel* doit être anonymisé car il apparaît dans le seul dictionnaire des prénoms ; le mot *crayon* est ignoré car il apparaît uniquement dans l'anti-dictionnaire (LEFFF) ; le mot *Pierre* est *ambigu* car il est présent dans les deux dictionnaires ; enfin, *Namrata* est *inconnu* car il est absent des deux dictionnaires. Dans ces deux derniers cas, le mot est surligné par le logiciel et sera à traiter dans la phase d'anonymisation semi-automatique.
Un exemple d'anonymisation de prénoms complétée figure dans le tableau 3.9.
Cette section se concentre sur la phase d'anonymisation la plus complexe du processus, à savoir l'anonymisation des prénoms. Dans le cadre de nos travaux,

113. Lexique des Formes Fléchies du Français, LEFFF, http://alpage.inria.fr/~sagot/lefff.html

Tableau 3.8 – *Identification automatique des mots à traiter (*Accorsi *et al. 2014)*

Traitement du mot	Dans le dictionnaire ?	Dans l'anti-dictionnaire ?	Type	Traitement
Rachel	Oui	non	Dictionnaire	Automatiquement anonymisé
crayon	Non	oui	Anti-dictionnaire	Ignoré (ne pas anonymiser, NPA)
Pierre	Oui	oui	Ambigu	Surligné (candidat pour la phase semi-automatique)
Namrata	Non	non	Inconnu	Surligné (candidat pour la phase semi-automatique)

d'autres types d'anonymisation [114] ont été réalisés qui s'appuient sur la mise en place d'expressions régulières pour identifier quelques éléments spécifiques (adresses de courriel, numéros de téléphone, adresses URL, etc.)
La partie la plus délicate du traitement automatique de ce type de corpus réside dans la mise en correspondance des dictionnaires compte tenu des spécificités lexicales des SMS. Ainsi, nous avons identifié les cas ci-dessous :

— Mots orthographiés, de manière non standard, par exemple : *surment* (à la place de *sûrement*)

— Mots écrits sans les accents, par exemple : *desole* (à la place de *désolé*)

— Mots avec des accents non standards, par exemple : *dèsolè* (à la place de *désolé*)

— Prénoms avec ou sans majuscules : *cédric* (à la place de *Cédric*)

Tableau 3.9 – Du SMS « brut » au SMS anonymisé. Les chiffres renvoient au nombre de caractères du prénom dans le SMS brut.

Coco est pas la ! Éva non plus ! Tanpis ! Lol J'irai aux journée du patrimoine ! Éva m'a dit que tu venais cette semaine peut etre ! Bisous !!
<PRE_4> est pas la ! **<PRE_3>** non plus ! Tanpis ! Lol J'irai aux journée du patrimoine ! **<PRE_3>** m'a dit que tu venais cette semaine peut etre ! Bisous !!

114. Les 10 étiquettes d'anonymisation, par ordre de fréquence d'utilisation dans le corpus sont : PREnom (10 905 étiquettes), SURnom (1 042), NOM (785), TELéphone (123), LIEu (102), ADResse (85), MARque (58), CODe (50), MEL (adresse électronique, 27), URL (13).

— Répétition de lettres, par exemple : *nicooooolllaassss* (à la place de *Nicolas*)

— Formes abrégées, diminutifs : *Nico* (à la place de *Nicolas*)

— Onomatopées, par exemple : *mouhahaha*

— Élision sans apostrophe, par exemple : *jexplique* (à la place de *j'explique*)

— Agglutination, par exemple : *jtaime* (à la place de *je t'aime*)

Le tableau 3.10 résume les heuristiques (programmes informatiques) conçues pour traiter les cas cités ci-dessus. Notons que les différentes heuristiques développées donnent des résultats tout à fait satisfaisants car les situations décrites sont correctement reconnues de manière automatique dans plus de 96,9% des cas (Accorsi et al. 2014).

À partir des 88 683 SMS du corpus, *Seek&Hide* en a anonymisé 63 728 (soit 72 %) ; les 24 955 SMS restants (28 %) sont soumis à la phase semi-automatique suivante.

La phase semi-automatique

Seek&Hide propose une interface web sécurisée permettant aux annotateurs-experts linguistes [115] de mener à bien la phase suivante, qui permet de désambiguïser les SMS et de décider si l'anonymisation doit ou non être effectuée.

La figure 3.19 montre les mots qui ont déjà été anonymisés pendant la phase automatique précédente (*Pauline* et *Lea*). Les autres mots surlignés requièrent une intervention humaine : *Juste* est NPA dans ce cas, mais le mot est surligné car il est potentiellement ambigu (tout comme *pierre*), entre un prénom et un mot

Tableau 3.10 – Différents algorithmes pour le traitement de mots spécifiques aux SMS.

Nom	Nom détaillé	Description
WWoutA	Mots écrits sans signes diacritiques (*desole*)	Effectuer une désambiguïsation au moment de la recherche
WWithA	Mots écrits avec des signes diacritiques non standards (*dèsolè*)	Effectuer une désambiguïsation au moment de la recherche
OmiA	Élision (*jexplique*) et agglutination (*jtaime*)	Identifier et éliminer les préfixes tels que *jt, jl, j,* etc., puis effectuer la recherche
SRepet	Onomatopées (*mouhahaha*)	Détecter les répétitions de sous-chaînes telles que *ha, hé*
LRepet	Répétitions de caractères (*nicooooolllaassss*)	Supprimer les lettres identiques consécutives puis effectuer la recherche

115. Deux étudiants de Master en Sciences du Langage à l'époque, Camille Lagarde Belleville et Michel Otell, ont effectué cette phase, pendant trois mois (octobre-décembre, 2012).

FIGURE 3.19 – Capture d'écran du logiciel *Seek&Hide*

figurant dans le dictionnaire LEFFF ; *elo* et *cece* sont AA, mais parce que ce sont des diminutifs, ils n'apparaissent pas dans le dictionnaire des prénoms ; *bebou* est anonymisé, car comme ce surnom n'est peut-être pas utilisé très souvent, il est potentiellement facile à reconnaître par des destinataires ou des tierces personnes ; *Penseea* montre les types de problèmes qui se posent et qui sont difficiles à traiter de manière automatisée, puisque l'espace est absente et correspond à : « Pense à » ; *Lec* est inconnu de tous les dictionnaires et dans ce cas est surligné ; cela est probablement une faute de saisie à la place de *Le*. Le dernier SMS montre d'autres mots modifiés, absents des dictionnaires : *petetre* à la place de *peut-être* (NPA), *surment* à la place de *sûrement*, (NPA), *pyis* à la place de *puis* (NPA), et *Juste*, encore une fois. [...] *Ben* présente en effet une ambiguïté (entre le prénom et l'interjection *bien*, abrégée dans ce contexte en *ben*) et doit être résolu grâce à un traitement humain : « ya des gens beaucoup mieu placer que moi pour te comprendre et *ben* je peux essaye » ; [ici] *ben* est NPA.

Précisons que des approches d'apprentissage automatique issues du domaine de l'Intelligence Artificielle ont été développées et ont été combinées avec le système décrit dans cet article (PATEL et al. 2013).

Phase de validation
La troisième phase [116] consiste en la lecture des SMS (72% du corpus) ayant été anonymisés de manière automatique par *Seek&Hide*, afin de vérifier si tous les textos ont bel et bien été correctement anonymisés. Trois cas de figure ont été repérés par les annotateurs :
Cas 1 : *anonymisation automatique à enlever* :
grace a lui on comprend trop bien franchement ke kiffe la physique cette *anne* meme si cest bien dur
Dans cet exemple, *grace* et *anne* ont été anonymisés, mais ce n'est pas une erreur du logiciel. Si le scripteur avait ajouté l'accent circonflexe, *Seek&Hide* n'aurait pas procédé à l'anonymisation en prénom pour *grâce* ; l'autre occurrence est *anne* au lieu d'*année*, qui n'est donc pas un prénom dans ce contexte.
Cas 2 : *anonymisation manquante à insérer* :
Excuse pour c texto si tard c'était pour t dire q **mat** a u l permis bisous bisous
Mat est absent du dictionnaire de prénoms.
balises d'anonymisation à remplacer :
Cas 3 : Une **clio** noir phase 2 vendue par une amie dune collègue de boulot.
Clio est ici un nom de voiture de la marque Renault, et non un prénom.
Les annotateurs humains peuvent donc retirer, ajouter, modifier les étiquettes précédemment insérées de manière automatique par le logiciel (*cf*. figetiquettes). À ce stade, ils peuvent également décider de noter certains SMS comme devant être supprimés du corpus si ceux-ci contiennent des propos légalement inacceptables.
(Panckhurst et al. 2013, p. 113-121).

Récapitulatif

Le logiciel *Seek&Hide* a permis d'anonymiser très majoritairement le corpus *88milSMS* de manière automatique, avant le travail semi-automatique et la relecture par les annotateurs humains (figure 3.20).

À l'issue des 21 mois-stagiaires consacrés à l'anonymisation, j'ai voulu récapituler la répartition des étiquettes utilisées en majorité et vérifier l'efficacité du logiciel.

Comme on le constate dans la figure 3.21, les deux étiquettes *prénom* (PRE) (Prénom) et *surnom* (SUR) constituent à elles seules plus de 90 % de l'utilisation

116. Deux étudiants de Master en Sciences du Langage à l'époque, Frédéric André et Yosra Ghliss, ont effectué cette phase, pendant trois mois (février-avril, 2013).

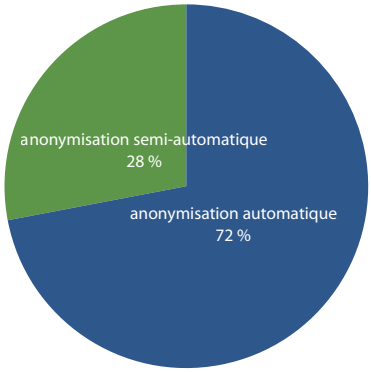


FIGURE 3.20 – Traitement de l'anonymisation du corpus *88milSMS*.

totale (82,7 % et 7,9 % respectivement). Viennent ensuite *nom* (NOM), à 6 %, puis les 3 % restants correspondent respectivement aux étiquettes *téléphone* (TEL, 0,9 %), *lieu* (LIE, 0,8 %), *adresse* (ADR, 0,6 %), *marque* (MAR, 0,4 %), *code* (COD, 0,4 %), *courriel* (MEL, 0,2 %) et *adresse web* (URL, 0,1 %).

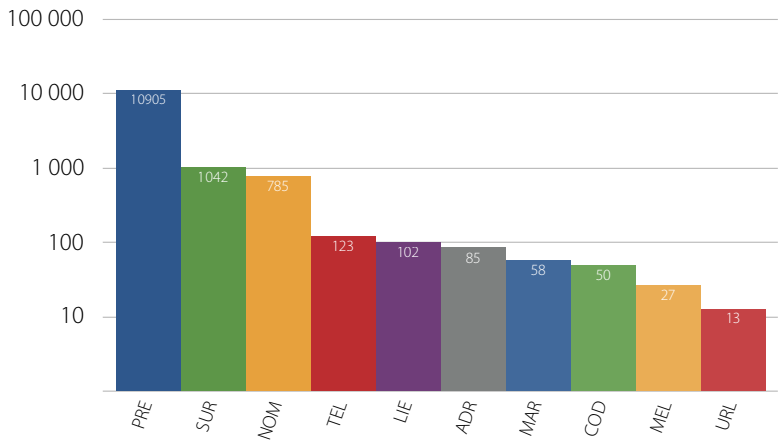


FIGURE 3.21 – 10 étiquettes d'anonymisation du corpus *88milSMS*, par fréquence d'utilisation décroissante.

Sur un échantillon représentatif de 20 000 SMS, nos étudiants stagiaires ont répertorié 358 cas où des modifications d'étiquetage de l'anonymisation ont dû être effectuées : cas 1, anonymisation automatique à enlever (66 %) ; cas 2, anonymisation manquante à insérer (29 %) ; cas 3, étiquette d'anonymisation à modifier (5 %) (*cf.* figure 3.22).

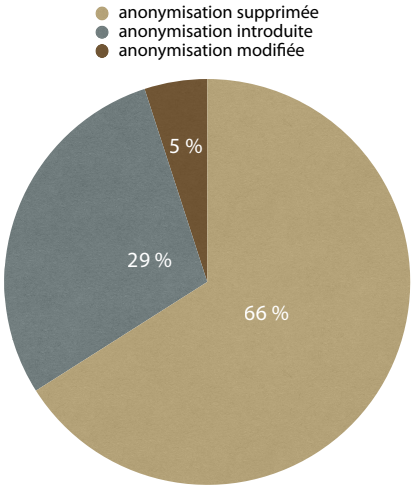


Figure 3.22 – Suppression/introduction/modification des étiquettes d'anonymisation, corpus *88milSMS*.

D'un point de vue juridique (Ghliss et André 2017), le cas 2 (figure 3.22) est le plus important à repérer (29 % des 358 modifications au total), d'où l'importance d'un travail rigoureux de la part d'annotateurs humains.

3.3.3.6 Corpus et questionnaire

N-grams fréquents : « je », « c est », « je t aime ».

En parallèle au travail sur l'anonymisation, et ce rapidement après la fin de la collecte en décembre 2011, j'ai travaillé (en collaboration avec Claudine Moïse) sur une extraction de résultats concernant le corpus et les réponses au questionnaire sociolinguistique. Les premiers résultats concernant le corpus *88milSMS* sont synthétisés dans la figure 3.23.

117. Pour une présentation moins schématique de ces résultats, *cf.* (Panckhurst et al. 2013, p. 109-111).

corpus
93 085 sms authentiques recueillis auprès du grand public (15/9/2011-15/12/2011)
+88 000 sms après épuration (doublons, messages publicitaires, etc.)
1 219 771 mots, 5 984 306 caractères avec espaces, 4 858 902 caractères sans espaces.
424 donateurs définitifs (509 inscrits au départ)
209 sms par personne, en moyenne
chaque sms fait 55 caractères, sans espaces, en moyenne
chaque sms fait 67 caractères, avec espaces, en moyenne
13,75 mots par sms en moyenne
14 677 SMS sont dotés de ≤ 15 caractères, soit 16,55 % du corpus 88milSMS
Sur les 30 000 binettes/smileys utilisés, seul 1 % = smileys « graphiques »/emoji
questionnaire
95 % des donateurs ont rempli un questionnaire sociolinguistique
+ 60 % de femmes
80 % ont moins de 30 ans
29 % ont moins de 18 ans
74 % ont un smartphone (dont 53% un iphone et 21% un blackberry)
88 % ont un forfait sms illimités
80 % utilisent les sms depuis + de 5 ans
65,4 % ont fait ou font des études supérieures
53 % élèves/étudiants
11 à 66 ans
24,1 ans en moyenne
71,4 % majeurs
28,6 % mineurs

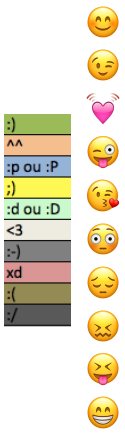


FIGURE 3.23 – Résultats Corpus/Questionnaire *88milSMS* [117]

Claudine Moïse et moi-même avons présenté (entre autres) ces résultats au colloque Vals-Asla à Lausanne en 2012, dans une communication (non-publiée, Panckhurst & Moïse, 2012) autour de la démarche et la méthode scientifiques de la collecte de SMS, en évoquant également le lien entre données et usages/pratiques [118]. En effet pendant la préparation du questionnaire sociolinguistique (*cf.* annexe au chapitre (PANCKHURST et MOÏSE 2014)), toute l'équipe a longuement discuté de la meilleure façon de formuler les questions — en sachant que le questionnaire s'adressait non seulement à des personnes majeures, mais également à des mineurs — afin de récolter des réponses pertinentes autour des thématiques suivantes :

Les utilisateurs de smartphones sont-ils toujours lexicalement créatifs, ou utilisent-

118. On pourra également consulter le diaporama de la communication en ligne (Panckhurst 2012), *Corpus Écrits* : https://groupes.renater.fr/wiki/corpus-ecrits-nouvcom/_media/public/reunion-31mai/nouv-com-panckhurst-120531.pdf, consulté le 11 janvier 2017. J'ai inséré plus bas certaines des diapositives de celles-ci.

> ils systématiquement l'écriture intuitive, désormais intégrée dans le téléphone ? Les textos sont-ils plus longs maintenant qu'il y a dix ans ? Les forfaits mensuels incluant des SMS illimités contribuent-ils à des mutations quelconques ? L'âge des scripteurs est-il systématiquement un critère concernant le style d'eSMS ? Les outils de reconnaissance vocale (Siri, Iris, etc.) modifient-ils de manière importante les usages ? [119] Les scripteurs multilingues font-ils souvent du code-switching [120] quand ils rédigent des textos ? Un étudiant [(Dos Santos 2013)] effectue actuellement une recherche sur les scripteurs qui se déclarent (dans le questionnaire) être bi ou trilingues, afin de comparer, d'une part, le lien entre leur(s) pratique(s) annoncée(s) et/ou leurs représentations de l'eSMS et leurs pratiques réelles, et, d'autre part, des différences éventuelles entre scripteurs monolingues et plurilingues. (Panckhurst et al. 2013, p. 130).

Langues

Les questions autour des langues ont été longuement débattues, également, y compris l'ordre dans lequel on les a posées dans le questionnaire final :

1. *Langue(s) pratiquée(s)*. Dans quelle(s) langue(s) vous sentez-vous le plus à l'aise ? Par exemple : français, espagnol, créole, arabe, berbère, wolof, LSF…)
2. *Langue(s) pratiquée(s)*. Quelle(s) langue(s) employez-vous le plus souvent ? Les réponses peuvent être différentes par rapport à la question précédente.
3. *Langue(s) maternelle(s)*. Quelle(s) est/sont votre/vos langue(s) maternelle(s) ? Les réponses peuvent être différentes par rapport aux deux questions précédentes.

Dans l'exemple de la figure 3.24, on constate qu'outre les deux langues maternelles, le malgache et le français, la personne se sent à l'aise dans trois langues (l'anglais en plus), mais utilise le plus souvent le français dans le quotidien. Cet apport d'information peut permettre de mieux comprendre les pratiques des usagers [121] (*cf*. également les tableaux suivants 3.11, 3.12, 3.13).

119. *Cf*. (Kaba 2014) pour un mémoire de Master sous ma direction.
120. *Cf*. également (Doehler 2011) et (Morel 2016) pour le contexte suisse.
121. Nous avons utilisé la norme ISO-639-2, dont les noms/abréviations sont répertoriés ici,

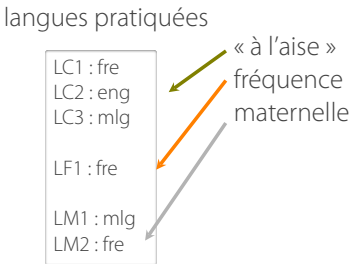


FIGURE 3.24 – Questionnaire : langues pratiquées d'un donateur. [122]

Tableau 3.11 – Langues courantes (LC1 à LC4).

LC1	LC2	LC3	LC4
fre	eng	spa	spa
eng	spa	eng	ita
spa	fre	ger	ger
ita	ger	ita	eng

http://www.loc.gov/standards/iso639-2/php/code_list.php. Trente-deux langues ont été citées dans les réponses au questionnaire : allemand, anglais, arabe classique, dialectes de l'arabe, bambara, langues berbères, catalan, chinois, créoles et pidgins basés sur le français, créoles et pidgins basés sur le portugais, espagnol/castillan, français, grec moderne (après 1453), hébreu, italien, japonais, kabyle, kazakh, langues des signes (LSF), latin, maori (nous pensons qu'il s'agit d'un remplacement effectué par les donateurs de SMS pour le mahorais/shimahorais/shimaore, Mayotte, qui n'est pas cité dans l'ISO-639-2), malgache, occitan (après 1500), polonais, portugais, roumain/moldave, russe, serbe, suédois, turc, vietnamien, wolof. Dans la version 2016 du corpus (PANCKHURST et al. 2016a), parmi les langues, seule la langue maternelle a été extraite des réponses. (*Cf.* également : (DOS SANTOS 2013), mémoire de Master sous ma direction concernant le plurilinguisme et les SMS).

122. LC = langue(s) courante(s), dans laquelle/lesquelles on se sent le plus à l'aise, LF = langue(s) fréquente(s), dans laquelle/lesquelles on parle le plus fréquemment dans le quotidien, LM = langue(s) maternelle(s) ; fre = français, eng = anglais, mlg = malgache. *Cf.* également (PANCKHURST et MOÏSE 2014).

Dans le tableau 3.11, le français est cité le plus souvent en première position (LC1), suivi de l'anglais, l'espagnol et l'italien.

Tableau 3.12 – Langues fréquentes (LF1 à LF4).

LF1	LF2	LF3	LF4
fre	eng	eng	spa
eng	spa	spa	chi
spa	fre	ger	ara_c, eng, fre, ger, ita, rus
ara_d, bam, fur, ger, ita, sgn, srp, vie, wol	ger	ita	ø

On constate, dans le tableau 3.12, que la LF peut varier, vis-à-vis de la LC. Ici, si le français, l'anglais et l'espagnol sont cités de prime abord, ensuite figurent l'arabe (dialectes), le bambara, le frioulan [123], l'allemand, l'italien, la langue des signes française, le serbe, le vietnamien et le wolof.

Tableau 3.13 – Langues maternelles (LM1 à LM4).

LM1	LM2	LM3	LM4
fre	fre	wol, eng	ø
eng	eng, spa	ø	ø
wol	cpf, ita	ø	ø
ara_d, ger, mlg, spa	ara_d, heb, vie	ø	ø

123. Cela correspond probablement à une erreur dans le choix de la langue du menu déroulant : français/frioulan.

Enfin, la langue maternelle (LM) n'est pas obligatoirement utilisée dans le quotidien, comme en témoignent les indications du tableau 3.13. En première position sont indiqués le français, puis l'anglais, le wolof, et l'arabe (dialectes), l'allemand, le malgache, l'espagnol.

Sélection de réponses du questionnaire

J'indique ci-dessous les réponses pour deux questions que nous avions posées, afin de comprendre les motivations des donateurs de SMS concernant le fait même d'emprunter ce moyen de communication, d'une part, et d'utiliser *l'écriture SMS* (eSMS), d'autre part :

> En réponse à la question « Pourquoi écrivez-vous des SMS (au lieu de téléphoner, par exemple ?) », les indications suivantes sont fournies, respectivement, en sachant que l'on pouvait cocher plusieurs cases :

— « parce que c'est moins cher ou inclus dans le forfait » (302 personnes) ;

— « pour aller plus vite » (294 personnes) ;

— « pour ne pas déranger » (211 personnes) ;

— « parce que je n'aime pas téléphoner » (143 personnes) ;

— « autre raison » (36 personnes) : « parce que ça laisse le choix au destinataire de répondre ou d'attendre suivant la situation dans laquelle il se trouve. » ; « [ce sont des] messages qui ne sont pas assez importants pour téléphoner à la personne » ; « pour pouvoir les relire » ; « le plaisir d'écrire ».

> À la question « Si vous rédigez en écriture SMS, pourquoi le faites-vous ? », voici les réponses obtenues :

— « parce que c'est plus rapide » (293 personnes) ;

— « parce que ça crée une complicité entre amis » (58 personnes) ;

— « parce que j'aime jouer avec la langue » (42 personnes) ;

— « autre » (4 personnes) : « j'écris souvent de long sms, dc ça raccourcis sans diminuer le contenu », « c'est amusant decrire certains mots d'une certaine maniere ; en plus d'être plus rapide on peut presque déterminer qui ecrit tel sms avec des mots particuliers ecrits d'une manière particulière et c'est amusant ^^ [124] ».

124. Les réponses sont retranscrites telles quelles.

> Notre questionnaire [...] montre que les SMS sont voués à un avenir encore prometteur, en incluant une dimension de *discrétion* ou de choix du *moment de communication* (« Ils sont très pratiques car ils me permettent de faire passer un message sans déranger la personne. On doit répondre tout de suite (plus ou moins) à un appel alors que nous pouvons répondre aux sms quand nous voulons. »), ou sans s'encombrer de *rituels conversationnels* (« L'envoi de sms permet de communiquer quasi instantanément en allant directement au but, contrairement à la conversation téléphonique qui induit un certain nombre de passages obligés avant d'entamer le sujet réel de conversation. »), ou encore en tant qu'*aide-mémoire* (« moi qui ai peu de mémoire, ça me permet de garder les infos (adresses, heure de rdv...) », ou en *se protégeant* derrière un écran (« écrire ce que l'on n'ose pas dire ») et enfin pour des publics ayant différentes spécificités (« je suis sourde, c'est plus simple ! »). Toutes ces caractéristiques s'appuient sur les spécificités de la conversation par SMS, temps différé, moment d'échange choisi, etc.
>
> (Panckhurst et al. 2013, p. 111-112).

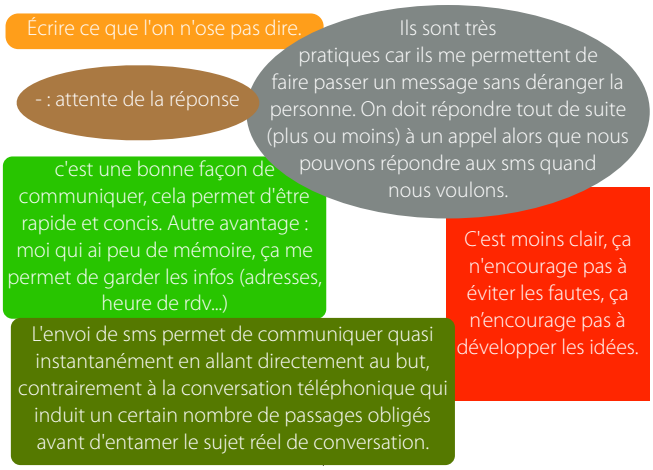


Figure 3.25 – Quelques réponses du questionnaire *sud4science LR*

A posteriori, nous aurions dû reformuler la question « Si vous rédigez en écriture SMS, pourquoi le faites-vous ? », car elle n'a pas toujours été correctement interprétée. Une autre question la précédait : « En général, quand vous rédigez vos SMS, utilisez-vous l'orthographe "standard" ou l'"écriture SMS" ? » (Panckhurst

et Moïse 2014) et on pensait qu'elle serait suffisamment claire pour saisir la nuance, mais certaines réponses ont démontré le contraire. Une formulation comme : « Si vous rédigez vos SMS en français non "standard", pourquoi le faites-vous ? » aurait peut-être été plus immédiatement comprise.

Nombre de caractères

Sur un échantillon total constitué d'un peu plus de 11 500 SMS, (André 2017) compare 5 sous-corpus (Belgique-2004, Réunion-2008, Suisse-2009, Québec-2010 et 88milSMS-2011 [125] en matière de nombre de caractères par SMS. Ces résultats apparaissent dans la figure 3.26.

> Alors que la limite des 160 caractères (espaces compris) apparait clairement dans les sous-corpus de 2004 et 2009 (notons également l'augmentation moins marquée dans le sous-corpus de 2008), elle disparaît totalement des autres sous-corpus. Ces derniers regroupent une majorité de messages plus courts (près de 20 % des messages des sous-corpus de 2010 et 2011 font moins de 20 caractères), mais également, bien que plus rares, beaucoup plus longs (jusqu'à 4 665 caractères dans le sous-corpus *88milSMS* (SMS n°7 375), soit près de 30 messages mis

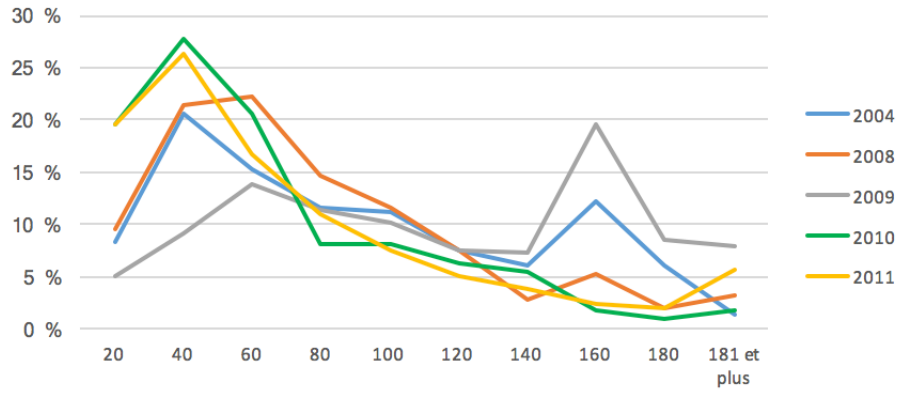


Figure 3.26 – Taux de SMS en fonction du nombre de caractères (André 2017, p. 240).

125. *Cf.* (Cougnon et François 2011) pour une comparaison détaillée des 4 corpus précédant notre collecte.

> bout à bout). Cette observation est corroborée par les commentaires produits par certains scripteurs lors du remplissage des questionnaires, notamment ceux répondant à la question « Pourriez-vous faire un commentaire sur la manière dont vous écrivez vos SMS ? » (Ledegen 2014), comme par exemple, pour le sous-corpus de 2008, ceux des profils suivants :
>
> — n°RE-003 : « Je racourcie les mots dans le but de n'evoyer qu'un seul sms. »
>
> — n°RE-046 : « selon la longueur du SMS, soit en texte complet si la place le permet, soit avec des abréviations courantes de prise de note, quelques mots en langage SMS. »
>
> — n°RE-063 : « J'utilise peu le langage sms. J'essaie de respecter au mieux l'orthographe. J'utilise néanmoins des abréviations personnelles, que je rajoute au fur et à mesure dans mon dictionnaire (j'utilise l'ecriture intuitive). En résumé, j'utilise le langage sms après coup, uniquement pour raccourcir mon sms quand je constate que ce que j'ai à dire ne "rentre" pas dans un seul texto. »
>
> (André 2017, p. 240-241).

En confrontant le corpus belge-2004 entier et notre corpus *88milSMS*, on constate une utilisation importante de SMS très courts (contenant moins de 15 caractères) : 17,3 % pour le corpus belge, contre 16,5 % pour *88milSMS* (*cf.* § 3.23). On aurait pu s'attendre à un taux inférieur pour le corpus de 2004, puisque les SMS n'étaient pas encore inclus dans les forfaits mensuels. Mais, comme le signale Frédéric André [126], à mon avis à juste titre : « les messages très courts restent [...] une des caractéristiques de l'écriture SMS, participant pleinement à la quasi-synchronie des échanges, simulant une conversation en coprésence ».

126. Échange par courriel, le 23 janvier 2017.

3.3.3.7 Transcodage/alignement

Ce paragraphe concernant les questions liées au transcodage et à l'alignement, est compilé à partir du fichier intitulé « 1000_SMS_transcodage.pdf » consultable et téléchargeable sur http://88milsms.huma-num.fr/, du rapport scientifique remis à la DGLFLF en juin 2013, et de l'article (Panckhurst et al. 2016b). Concernant les traitements en TAL pour le transcodage et l'alignement, le lecteur pourra également se reporter, entre autres, à : (Aw et al. 2006 ; Beaufort et al. 2010, 2008 ; Fairon et al. 2007 ; Guimier de Neef et Fessard 2007 ; Kobus et al. 2008 ; Kogkitsidou et Antoniadis 2016 ; Lopez et al. 2014).

anonymized sms (n° 11326, 88milSMS corpus) **transcoding?**
B, kèl intense réflexion ! Je c, t en week ! <SUR_5> a A C 2 matièr pr fèr son suG. Concer tré 5pa ièr. Bone soiré a toi é tte, bon week ? 2vé fèr gd bo ici : ke dal. Bisous.

anonymized & transcoded sms
Bon/Bien/Ben, quelle intense réflexion ! Je sais, **tu es** en week-end ! <SUR_5> a assez de matière pour faire son sujet. **Concert** très sympa hier. **Bonne soirée à toi et toute(s), bon week-end ? Il** devait faire grand beau ici : rien. Bisous.

Problems

B	semantic abreviation
t	oral/sociolinguists ('tes'/'tu es'), morpho-syntaxtic parsers ('tu t'es trompé')
le	'ellipsis', not injected before 'concert', additional interpretation?
Bone soirée…	ambiguity between 'to you' or 'see you later' (in French 'à toute')
il	mandatory for morpho-syntactic parser processing
que dalle	colloquial form, could be replaced by 'rien' ('nothing'), the standard form.

Figure 3.27 – Exemple et problèmes liés au transcodage de SMS

La « chaîne de traitement » nécessaire, depuis le SMS « brut » au résultat final du SMS « annoté », en passant par les phases d'anonymisation et de transcodage, peut être schématisée comme suit (*cf.* figure 3.28).

Suite à l'anonymisation, les SMS sont prêts à être transcodés en français standardisé afin de permettre d'éventuels traitements ultérieurs en linguistique-informatique (incluant des analyseurs morphosyntaxiques). L'idée est de restituer l'orthographe et la grammaire afin de faciliter la fouille et la compréhension, mais non d'« injecter » des éléments supplémentaires.

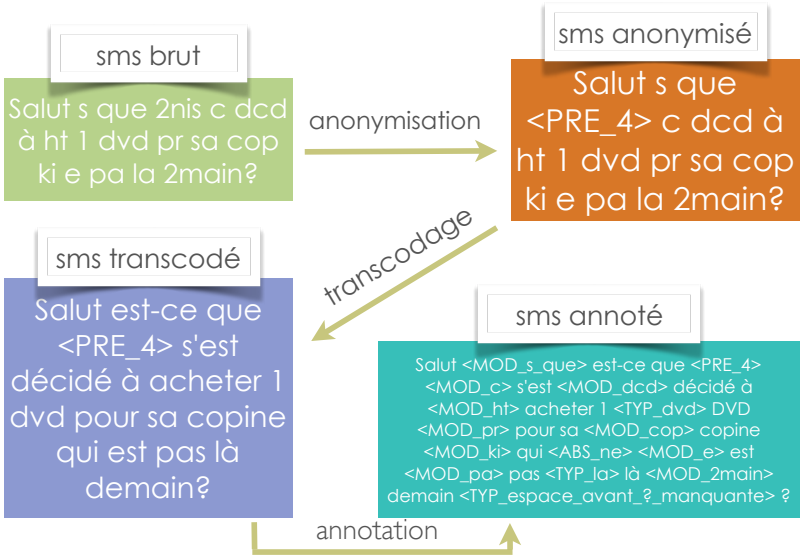


Figure 3.28 – Du SMS « brut » au SMS anonymisé, transcodé puis annoté

Le transcodage peut être utile pour le grand public, ou pour ceux qui veulent lire et comparer rapidement les SMS bruts anonymisés et transcodés, à des fins de recherche.

D'un point de vue linguistique, il est extrêmement difficile de procéder à un transcodage qui convienne à tous, car les interprétations sont nombreuses et variées. Prenons un exemple pour illustrer ce propos.

SMS brut anonymisé (n° 22446 du corpus 88milSMS) :
« En fait c rien de spécial, jprends juste un peu de recul et jcomprends pas ce que jfous là, fac, psycho, montpellier, pourquoi simplement je vis, enfin bref rien de grave. Qu'est ce qui cloche chez toi ? »
SMS anonymisé et transcodé en français standardisé :
En fait **c'est** rien de spécial, **je** prends juste un peu de recul et **je** comprends pas ce que **je** fous là, fac, **psychologie**, **Montpellier**, pourquoi simplement je vis, enfin bref rien de grave. Qu'est-ce qui cloche chez toi **?**

Exemple : transcodage

> Dans l'exemple ci-dessus, on n'ajoutera pas la particule de négation, ne/*n'*. On n'« injectera » pas non plus des éléments prépositionnels ou des déterminants (« *à la* fac », « *en* psychologie », « *à* Montpellier »), car le traitement automatisé demeure possible sans ces informations. En revanche, pour des formes abrégées, agglutinées, etc., on transcode en français standardisé (« c » =>« c'est » : ici, il s'agit d'une *abréviation sémantisée*, lorsqu'un mot est réduit à l'initiale et seul le co(n)texte permet de déterminer de quel mot il s'agit ; agglutinations : *jprends, jcomprends, jfous*) pour qu'un analyseur morphosyntaxique soit à même de traiter automatiquement la phrase. La forme en apocope « fac » demeure telle quelle dans la version transcodée, car nous avons décidé de valider le transcodage en lien avec les informations apparaissant au sein du *Petit Robert* en ligne (PR) : si une entrée dictionnairique existe, elle n'est pas transcodée dans sa forme entière (« fac » demeure intact, mais « psycho » sera transcodé en « psychologie », car si l'élément « psycho- » existe effectivement dans le PR, l'apocope qui renvoie à « psychologie » n'y figure pas). Par ailleurs, lorsque la ponctuation est présente, les normes typographiques sont rétablies pour le français, ici sont réintroduites la lettre « M » majuscule pour le nom de ville Montpellier et l'espace absente avant le point d'interrogation final. Cet exemple de transcodage donne un aperçu de la difficulté de cette opération. (Panckhurst et al. 2016b)

Un extrait du corpus belge de 30 000 SMS (SMS4science, http://www.sms4science.org/) a été manuellement transcodé (Fairon et al. 2007). Une fois cette opération effectuée, un système expert a été conçu pour aligner le corpus, caractère par caractère, et donc apprendre à partir des données, en comparant ainsi les SMS 'bruts' avec ceux qui avaient été transposés en français standardisé (Beaufort et al. 2010). À partir du travail belge, six étudiants (en Master d'informatique) ont travaillé sur le transcodage de *88milSMS* sous la direction de Mathieu Roche.

> [Les étudiants] ont étudié la faisabilité d'une méthode d'alignement des SMS pour faciliter le passage du SMS brut anonymisé au SMS transcodé en français standardisé, et ils ont proposé un modèle pour une interface en ligne afin de faciliter le travail de l'annotateur humain. Le modèle d'alignement incluant une interface s'intitule AlignSMS (Lopez et al. 2014 ; Panckhurst et al. 2016b)
> Plus précisément, leur travail se décline en deux phases.

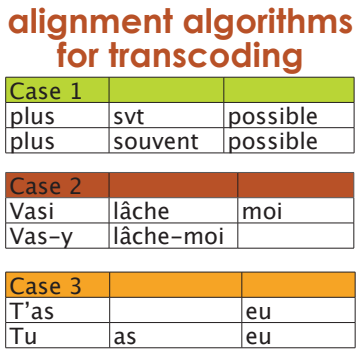


Figure 3.29 – Trois cas d'alignement pour le transcodage. (Lopez et al. 2014, 2016)

.

1. Dans un premier temps, une étape d'analyse a été menée afin d'identifier les erreurs de transcodage issues du programme de Louvain (Fairon et al. 2007). Une première évaluation quantitative sur la base des mesures de qualité issues du domaine de la fouille de textes (précision, rappel, F-mesure) a mis en avant qu'autour de 50% des mots étaient mal transcodés. L'évaluation qualitative associée a permis d'identifier les types d'erreurs. Cette étape a mis en avant que les erreurs de transcodage étaient souvent dues à la non prise en compte du contexte.
2. Ainsi, dans une deuxième phase, ils ont proposé un logiciel d'alignement afin que l'utilisateur (expert linguiste) puisse corriger les erreurs en prenant en compte le contexte. Leur algorithme d'alignement a donné des résultats tout à fait satisfaisants puisque plus de 99% des mots sont correctement alignés (évaluation menée sur un échantillon de 100 SMS représentant 2063 mots à traiter).

   Suite à ce travail, l'un des 6 étudiants l'a poursuivi à l'aide d'un stage d'un mois (juin 2013) en appliquant le travail d'analyse sur un échantillon de 1 000 SMS. (Panckhurst *et al.* 2013, Rapport scientifique DGLFLF. Pour les algorithmes et résultats détaillés, (Lopez et al. 2014).
   Lorsque des étudiants (en Master de Sciences du Langage) ont procédé à un test de transcodage sur un échantillon de 1000 SMS, ils ont conclu que l'écriture SMS étudiée à travers le travail de transcodage « contient une réelle forme de créativité [...] imprégnée des personnalités des émetteurs et récepteurs des SMS (de leur

> passé, de leur humour, de leur quotidien), des effets de modes, des contextes actuels. » (Dalle et al. 2013) in (Panckhurst et al. 2016b)

Taux d'abréviation

Concernant le taux de réduction de caractères entre SMS transcodés (en français standardisé) et SMS bruts (par exemple, *désolé/dsl*, passant de 6 à 3 caractères), on aurait pu s'attendre à un écart important reflétant ainsi les substitutions, abréviations, etc. omniprésentes au sein de l'écriture SMS. Cependant, il n'en est rien. Sur les 1 000 SMS transcodés de *88milSMS*, le taux d'abréviation est faible : 5 %. Cela est effectivement un pourcentage inférieur comparé au corpus *SMS4science* belge (9,5 %), mais peut s'expliquer par la taille de l'échantillon (1 000 SMS) qu'il faudrait accroître afin de comparer de manière plus fiable.

La figure 3.30 indique les taux d'abréviation dont les variations sont plus ou moins importantes d'une collecte et d'une région du monde à l'autre (Cougnon et François 2011, p. 30) [127].

Il serait intéressant d'effectuer de nouvelles collectes [128] afin de comparer avec l'évolution des pratiques scripturales et des technologies mouvantes (écriture intuitive, etc.).

127. Le pourcentage pour les SMS francophones de la collecte suisse m'a été signalé par Elisabeth Stark et Simone Ueberwasser dans un courriel du 24 janvier 2014.

128. Comme celle proposée par le CENTAL, en 2017 : http://www.vospouces.org/. *Cf.* également la note 102 pour le détail de ce projet.

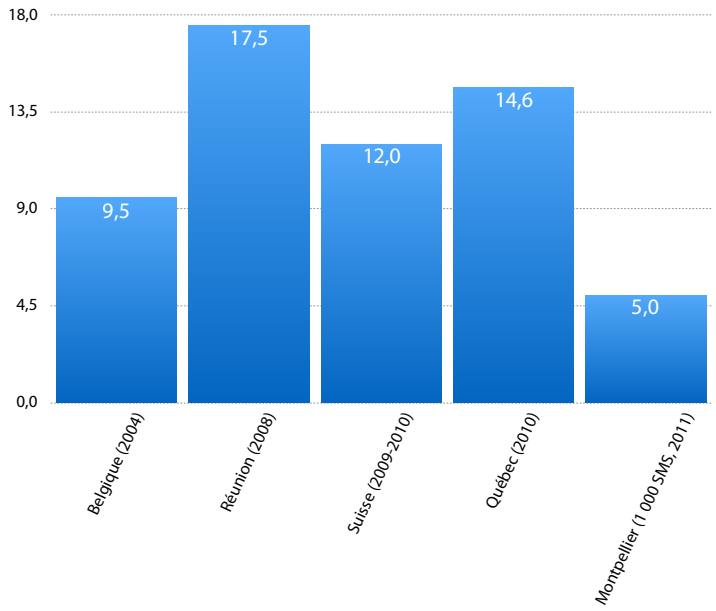


Figure 3.30 – Taux d'abréviation (%) entre SMS transcodés et SMS bruts.

3.3.3.8 Annotation

Ce paragraphe est compilé à partir du fichier intitulé « 100_SMS_annotations_balises.pdf » consultable et téléchargeable sur http://88milsms.huma-num.fr/, du rapport scientifique remis à la DGLFLF en juin 2013, et des articles (Panckhurst 2016a ; Panckhurst et al. 2016b). Ces deux articles situent également nos choix théoriques dans une approche pluridisciplinaire. Des journées d'études (14 et 15 novembre 2011) : « Harmonisation/standardisation des méthodes de traitement de corpus écrits de type SMS. Anonymisation, transcodage, annotation. » ont permis d'échanger avec les acteurs des différentes collectes *SMS4science* précédentes, afin de nous guider dans notre choix pour *88milSMS* : http://www.mshm.fr/programmes-2011/sud4science-lr/article/journees-des-14-et-15-novembre

La suite du processus mis en place concerne l'annotation linguistique, qui consiste

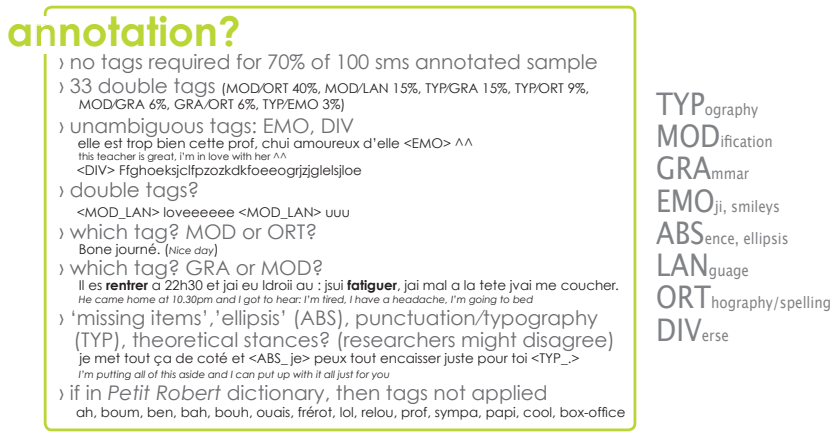


FIGURE 3.31 – Balises et problèmes évoqués pour l'annotation de SMS

> à apposer des balises associées à certains mots contenus dans un SMS, permettant de retrouver facilement une information. Cette annotation peut être utile pour des chercheurs qui souhaitent étudier certains aspects, en accédant rapidement aux données qui les intéressent. Lors du projet *sud4science LR*, nous avons invité les acteurs des collectes précédentes, dans le cadre de *SMS4science*, à présenter les balises utilisées pour l'annotation de leurs corpus de SMS. Une harmonisation générale nous a ensuite permis de réduire le nombre de balises précédemment utilisées. Huit balises ont été retenues : TYPographie, MODification, GRAmmaire, BINettes, ABSence, LANgue, ORThographe, DIVers. Les exemples figurant dans le tableau 3.14 correspondent chacun à une balise spécifique qui peut être appliquée manuellement. Nous n'indiquons pas l'annotation complète pour la totalité du SMS, mais simplement la balise concernée, afin d'éclairer la lecture.
>
> (PANCKHURST et al. 2016b).

Afin de tester la faisabilité de l'annotation linguistique [129] sur notre corpus *88milSMS*, j'ai annoté un extrait de 100 SMS (*cf.* tableau 3.15). Le détail de ce travail se trouve dans le fichier « 100_SMS_annotations_chercheurs.pdf », consultable et téléchargeable sur http://88milsms.huma-num.fr.

129. *Cf.* également (IDE et PUSTEJOVSKY 2017).

Tableau 3.14 – Exemples de SMS annotés (Panckhurst et al. 2016b)

SMS	Balises	Avant balisage	Après balisage
no 6 885	TYP*	Zorro est arrive, sans s'presse [...]	Zorro est <TYP_arrivé>arrive, sans s'presse [...]
no 4 360	MOD*	[...] Oui, j sui zalé ! [...]	[...] Oui, <MOD_j'y>j <MOD_suis>sui <MOD_allé>zalé ! [...]
no 5 536	GRA	Cc tu va mieux. Mam ma dis ke tèté retmbè malade. Et bb ? Bisx	Cc tu <GRA_vas>va mieux. Mam ma <GRA_dit>dis ke tèté retmbè malade. Et bb ? Bisx
no 6 887	BIN*	Au dos d'son beau tornado elle est trop bien cette prof, chui amoureux d'elle ^^	Au dos d'son beau tornado <BIN> elle est trop bien cette prof, chui amoureux d'elle <BIN>^^
no 19 621	ABS*	[...] je met tout ça de coté et peux tout encaisser juste pour toi. [...]	[...] je met tout ça de coté et <ABS_je>peux tout encaisser juste pour toi. [...]
no 43 133	LAN*	if(ce_soir == film) {get_commande ;} else {set_tagueule ;} return « bisous »	<LAN>if(ce_soir == film) { <LAN>get_commande ;} <LAN>else { <LAN>set_tagueule ;} <LAN>return « bisous »
no 19 621	ORT*	[...] notre couple sera tel un rosau à jamais se casser. [...]	[...] notre couple sera tel un <ORT_roseau>rosau à jamais se casser [...]
no 4 671	DIV*	Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe	<DIV>Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe

*TYP. Typographie : ponctuation, symboles mathématiques, signes diacritiques (accents, etc.) nombres, format des heures, ponctuations ou symboles inattendus, signe : &, chevrons, parenthèses, respect de la casse (majuscules/minuscules), mise en page.
*MOD. Modification (soit en réduction, soit en augmentation, soit en remplacement de caractères, abrègements et abréviations, acronymes, sigles et abréviations, répétition de lettres, transformations phonétiques, interjections et onomatopées...) : *ht* (acheter), *pr* (pour), *c* (s'est, c'est, ces...), *dcd* (décidé)...
*GRA. Accords : *il viens* (il vient), syntaxe : *si j'aurais su, je serais pas venu* (si j'avais su, je ne serais pas venu), etc.
*BIN. Binettes/frimousses/émoticônes/smileys :) ^^ : p ;) : d <3 :-) xd : (:/
*ABS. Absence/ellipse : négation, pronoms, éléments manquants faciles/difficiles à identifier, etc.).
*LAN. Contact de langues, emprunts, régionalismes, néologismes, verlan, argot, etc.)
*ORT. Uniquement l'orthographe lexicale : erreurs de saisie, interversion de lettres, etc.)
*DIV. Dans le cas où aucune autre balise ne semble convenir).

Tableau 3.15 – Synthèse chercheurs de l'utilisation des balises, pour un extrait de 100 SMS

Synthèse chercheurs			
TYP	1	449	**43,59%**
MOD	2	294	**28,54%**
GRA	3	84	**8,16%**
BIN	4	82	**7,96%**
ABS	5	52	**5,05%**
LAN	6	38	**3,69%**
ORT	7	30	**2,91%**
DIV	8	1	**0,10%**
	Total	1030	**100,00%**
Mots sans modification		2393	69,91%
Mots étiquetés		1030	30,09%
	Total	3423	100,00%

Il en ressort que les phénomènes de *typographie* sont les plus saillants, suivis par les *modifications* (substitutions, réductions, ajouts, etc.). La balise qui concerne la *grammaire* arrive en troisième position, suivie, dans l'ordre, par les *binettes*, l'*absence*, la *langue*, l'*orthographe* et la balise *divers*. Il est également intéressant de constater que 70 % de l'extrait des 100 SMS ne subit aucune modification. 113 commentaires figurent dans le fichier annoté, en partie pour décrire les entrées dictionnairiques apparaissant dans le *Petit Robert 2014 en ligne*.
Quatre groupes étudiants ont également effectué le même travail (*cf.* le fichier « 100_SMS_annotations_etudiants.pdf »). Ensuite, une étudiante a effectué une harmonisation des résultats des quatre groupes. La synthèse de ces résultats est présentée dans le tableau 3.16.
En comparant les deux figures contenant les résultats chiffrés, on constate des différences importantes d'attribution de balises. De même, des attributions divergentes existent entre groupes d'étudiants (*cf.* les tableaux récapitulatifs en fin de fichier « 100_SMS_annotations_etudiants.pdf »). Cela peut être dû aux raisons suivantes, mais pas seulement : certains encodeurs étudiants ont utilisé la balise ⟨ABS⟩ pour une absence de ponctuation, alors que cela doit être codé avec la balise ⟨TYP⟩, ⟨ABS⟩ étant réservé pour les ellipses de pronoms, de négation, etc. De même pour la balise ⟨ORT⟩ qui doit être uniquement réservée pour l'orthographe

Tableau 3.16 – Synthèse de l'utilisation des balises, pour l'extrait des 100 SMS annotés, par des étudiants de Master en Sciences du Langage.

Synthèse étudiants			
MOD	1	271	**30,83%**
TYP	2	263	**29,92%**
ORT	3	113	**12,86%**
BIN	4	83	**9,44%**
GRA	5	72	**8,19%**
ABS	6	39	**4,44%**
LAN	7	30	**3,41%**
DIV	8	8	**0,91%**
	Total	879	**100,00%**

(lexicale) et non pas grammaticale. Dans ce dernier cas, on utilisera ⟨GRA⟩. Un exemple comme « quil » doit être codé en ⟨TYP⟩ et non en ⟨MOD⟩. Ces « erreurs » d'étiquetage posent tout de même la question de différenciations importantes au niveau des choix des annotateurs et peuvent poser un réel problème dans l'étiquetage d'un grand corpus.
Les écarts entre les différents traitements montrent à quel point il est extrêmement difficile de proposer une annotation standardisée. Lors du projet *sud4science LR*, les chercheurs ont invité les acteurs des collectes précédentes, dans le cadre de *SMS4science*, lors de deux journées d'étude à Montpellier (« Harmonisation/standardisation des méthodes de traitement de corpus écrits de type SMS. Anonymisation, transcodage, annotation. », 14-15 novembre 2011, MSH-M), à présenter leurs balises pour l'annotation de leurs corpus de SMS. Une harmonisation générale a ensuite permis aux chercheurs *sud4science* de réduire le nombre de balises précédemment utilisées, afin d'envisager le balisage éventuel du corpus *88milSMS*. Par la suite, ils ont décidé de fournir ici un échantillon d'annotation de 100 SMS, mais de renoncer à l'annotation de l'ensemble du corpus 88milSMS, précisément car [...] les chercheurs ne seraient pas nécessairement en accord avec le choix des balises. Cet échantillon permet de fournir des pistes de recherche, mais il est important que chacun ait accès au corpus anonymisé, sans que des initiatives de balisage supplémentaire leur soient imposées.

(Panckhurst et al. 2014a), Extrait du fichier : « 100_SMS_annotations_balises.pdf », http://88milsms.huma-num.fr.

Je crois qu'il est crucial d'insister sur le fait que toutes les étapes précédemment exposées ici (acquisition, anonymisation, transcodage, annotation) ont nécessité des choix théoriques précis, qui, de plus est, ont été effectués dans une démarche pluridisciplinaire. Le transcodage et l'annotation linguistique posent de sérieux problèmes, comme nous le précisons ci-dessous :

> Nous constatons qu'il est extrêmement difficile, voire impossible, de proposer un transcodage et une annotation linguistique standardisés consensuels. Mais ce n'est pas parce que cela prendrait un temps conséquent que nous avons décidé d'y renoncer. Pour nous, il s'agit plutôt d'une position théorique. Le transcodage et l'annotation suscitent des désaccords théoriques, que ce soit au sein d'une même discipline, ou de manière interpluridisciplinaire. Nous considérons qu'annoter n'est pas une opération descriptive neutre. Elle relève nécessairement d'un cadre interprétatif. On comprend alors pourquoi elle peut ne pas faire consensus, parce qu'il y a des cadres théoriques différents, des démarches pluridisciplinaires distinctes, des questionnements scientifiques variés, etc. Nous pensons qu'il est préférable que les chercheurs prennent en charge leurs transcodage et annotation en fonction de leur(s) propre(s) questionnement(s).
>
> (Panckhurst et al. 2016b).

(McEnery et Hardie 2012) pèsent le pour et le contre concernant l'annotation de corpus :

> Arguments against annotation are largely predicated upon the purity of the corpus texts themselves, with the analyses being viewed as a form of impurity. This is because they impose an analysis on the users of the data, but also because the annotations themselves may be inaccurate or inconsistent [...]. Such claims are interesting because, as has been noted, corpus annotation is the manifestation within the sphere of corpus linguistics of processes of analysis that are common in most areas of linguistics. To identify problems with accuracy and consistency, in corpus annotation is, in principle at least, to identify flaws with analytical procedures across the whole of linguistics. It is because of the issues of accuracy and consistency, in particular, that some linguists prefer to use unannotated corpora. But this does not mean to say that such linguists do not analyse the data they use; rather, it means that they leave no systematic record of either their analysis or their errors which can easily and readily be tied back to the corpus data itself. (McEnery and Hardie 2012, p. 14).

L'annotation peut s'avérer correspondre à des positions linguistiques et/ou à de l'encodage de données. Dans les deux cas, les chercheurs peuvent préférer mettre à disposition leurs corpus dans deux versions, l'une correspondant à des données « brutes », l'autre à des données « annotées ». L'important est que le choix reste ouvert.

> Another alternative is that researchers may of course prefer to provide both "raw" and tagged corpora: "Dissemination will take two different forms: one version of a corpus with the 'raw' text without any tokenization and annotation (v1), and a second version of the same corpus with the annotations (v2)." (Chanier et al. 2014, p. 2). For instance, (Riou and Sagot 2016) present morpho-syntactic tagging of a specific corpus within the French CoMeRe corpora repository (v2), following on from a previous version without it (v1). (Panckhurst 2016a).

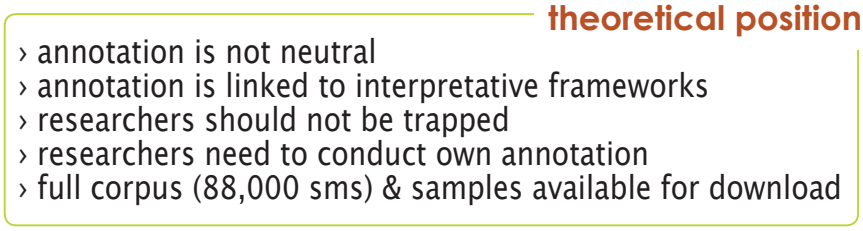


Figure 3.32 – Position théorique liée à l'annotation. (Lopez et al. 2016)

Dans (Panckhurst 2016a), je me suis également interrogée sur l'annotation linguistique et sa définition :

> Another issue is linguistic annotation of the corpus (*cf.* (Ide and Pustejovsky 2017), forthcoming). For example, the "raw" SMS "je met tout ça de coté et peux tout encaisser juste pour toi." (I'm leaving all of that aside and I can bear it all just for you.) could be transcoded into standardized French as follows: "Je mets tout ça de côté et je peux tout encaisser juste pour toi." It could then be linguistically annotated with information of interest to researchers, among other items: spelling, grammatical information, emoji insertion, code-switching, typography, missing accents, voluntary modification, etc. Therefore, *I define linguistic annotation of SMS data for the 88milSMS corpus, as "interpretative" linguistic information indicated*

> *via appropriate tags [...], related to the difference between a "raw" text message and its transcoded equivalent in standardized French. I do not include in this definition, lemmatization, or part-of-speech (POS) tagging [...], which do indeed also correspond to other methods of linguistic annotation (based mainly on providing lexico-morpho-syntactic information).* (Panckhurst 2016a), je souligne.

3.3.3.9 *88milSMS* : de 2014 à 2016

En juin 2014, nous avons donc mis à disposition du public, via un téléchargement direct (sur la grille de services d'Huma-Num : http://88milsms.huma-num.fr/, (Panckhurst et al. 2014a)), le corpus intégral, intitulé *88milSMS* entièrement anonymisé (format.ods), deux échantillons (SMS annotés, 1 000 SMS transcodés en français standardisé), le questionnaire sociolinguistique soumis aux donateurs, et leurs réponses.

En 2015, nous avons proposé une version du corpus encodé en XML dans le cadre d'une contribution Dariah [130]. Cela est déterminant pour un archivage à long terme au CINES [131]. En revanche, un certain nombre de problèmes se posaient pour la syntaxe du fichier XML, notamment, à cause de l'existence des quelque 30 000 binettes/émoticônes/smileys. J'avais pris la précaution d'encoder les emoji (graphiques) en Unicode, dans la version XML. En 2016, suite à une négociation avec notre collègue Thierry Chanier [132], j'ai pu proposer notre corpus pour mise sur Ortolang. Il a donc effectué un travail très important d'harmonisation et de standardisation pour transférer le corpus *88milSMS* dans une version XML syntaxiquement valide et notre licence (initialement plus restrictive) a évolué en une licence *Creative Commons Attribution 4.0 International (CC BY 4.0)*. La nouvelle version (Panckhurst et al. 2016a) est consultable et téléchargeable sur Ortolang : https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1

130. *Digital Research Infrastructure for the Arts and Humanities* : http://www.dariah.fr/
131. *Centre Informatique National de l'Enseignement Supérieur* : https://www.cines.fr/
132. Celle-ci s'est déroulée à Montpellier, le 27 mai 2016, entre Mathieu Roche, Thierry Chanier et moi-même.

Corpus « 88milSMS »

Plusieurs types de corpus/données sont disponibles :

1. « **88milSMS_88522.ods** » (le corpus de 88 522 SMS anonymisés, recueillis en 2011, à Montpellier dans le cadre du projet de recherche **sud4science LR**), par les six chercheurs suivants : (**Rachel Panckhurst**, **Catherine Détrie**, **Cédric Lopez**, **Claudine Moïse**, **Mathieu Roche**, **Bertrand Verine** (**Praxiling**, **Lirmm**, **Lidilem**, **Tetis**, **Viseo**)).
2. « **88milSMS.pdf** » (ce fichier explique comment lire le fichier de données « 88milSMS_88522.ods » et fournit un bref descriptif du projet) ;
3. « **100_SMS_annotations_chercheurs.ods** » (ce fichier contient un échantillon et un travail d'annotation de 100 SMS par les chercheurs du projet) ;
4. « **100_SMS_annotations_etudiants.ods** » (ce fichier contient un échantillon et un travail d'annotation de 100 SMS par les étudiants de Master ayant travaillé sur le projet) ;
5. « **100_SMS_annotations_balises.pdf** » (ce fichier contient les indications de lecture pour les fichiers : « 100_SMS_annotations_chercheurs.ods » et « 100_SMS_annotations_etudiants.ods ») ;
6. « **1000_SMS_transcodage.ods** » (ce fichier contient 1 000 SMS transcodés en français standardisé par les chercheurs et les étudiants de Master) ;
7. « **1000_SMS_transcodage.pdf** » (ce fichier contient les indications de lecture pour le fichier : « 1000_SMS_transcodage.ods ») ;
8. « **reponses_questionnaire.ods** » (ce fichier contient les réponses au questionnaire sociolinguistique, lié au corpus 88milSMS) ;
9. « **questions_explications_questionnaire.pdf** » (ce fichier contient les questions ainsi que les explications du questionnaire sociolinguistique) ;
10. « **88milSMS_88522_emoji-utf8.xml** » (ce fichier contient le corpus entier en format xml (encodage utf-8) ; les emojis graphiques sont également encodés en Unicode).

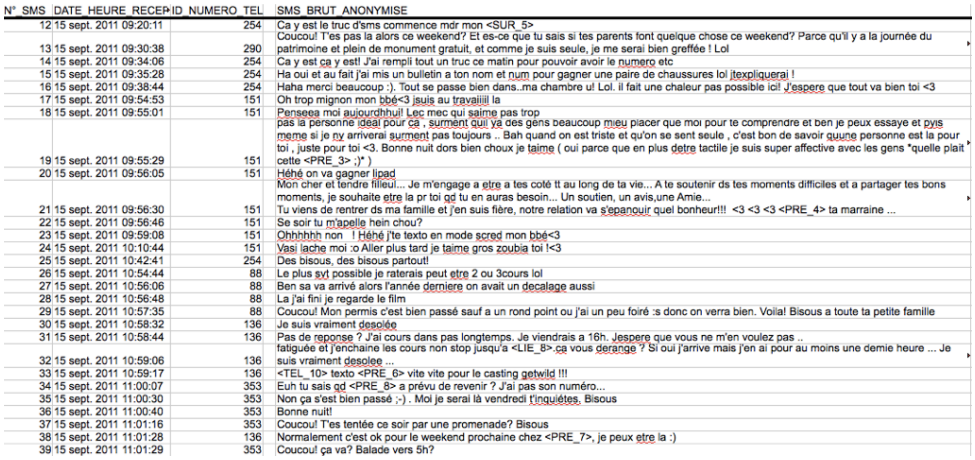
Remarque

Tous les SMS du corpus « *88milSMS* » ont été lus par un annotateur humain. Ils ont été anonymisés. Des SMS ont également été supprimés du corpus, soit parce qu'ils correspondaient à des SMS en 'chaîne', publicitaires, en provenance d'opérateurs téléphoniques, des doublons, etc., ou soit parce qu'ils contenaient des propos contraires à la loi ou aux règles de la collecte (ex : pas de propos sensibles concernant des tiers). Si vous consultez un SMS qui aurait échappé à la vigilance de notre équipe de relecteurs, vous êtes prié(e) de nous le signaler au plus vite en écrivant à l'adresse de courrier électronique suivante :
88milsms@univ-montp3.fr

Figure 3.33 – Une capture d'écran du site Web permettant la consultation et le téléchargement du corpus *88milSMS* et les données associées. http://88milsms.huma-num.fr/

3.3.3.10 Analyses des données

Tout au long du travail mené pendant le déroulement des projets (financement de 2011 à 2014), et depuis la mise à disposition de *88milSMS* sur la grille des services Huma-Num en 2014, nous avons continué à nous passionner pour l'analyse des données SMS, à partir d'approches pluridisciplinaires, en sciences du langage et en TAL/Informatique :



N° SMS	DATE_HEURE_RECEP	ID_NUMERO_TEL	SMS_BRUT_ANONYMISE
12	15 sept. 2011 09:20:11	254	Ca y est le truc d'sms commence mdr mon <SUR_5>
13	15 sept. 2011 09:30:38	290	Coucou! T'es pas la alors ce weekend? Et es-ce que tu sais si tes parents font quelque chose ce weekend? Parce qu'il y a la journée du patrimoine et plein de monument gratuit, et comme je suis seule, je me serai bien greffée ! Lol
14	15 sept. 2011 09:34:06	254	Ca y est ca y est! J'ai rempli tout un truc ce matin pour pouvoir avoir le numero etc
15	15 sept. 2011 09:35:28	254	Ha oui et au fait j'ai mis un bulletin a ton nom et num pour gagner une paire de chaussures lol jtexpliquerai !
16	15 sept. 2011 09:38:44	254	Haha merci beaucoup :). Tout se passe bien dans..ma chambre ul Lol. il fait une chaleur pas possible ici! J'espere que tout va bien toi <3
17	15 sept. 2011 09:54:53	151	Oh trop mignon mon bbé<3 jsuis au travaiiiil la
18	15 sept. 2011 09:55:01	151	Perseeea moi aujourdhhui! Lec mec qui saime pas trop
19	15 sept. 2011 09:55:29	151	pas la personne idéal pour ça , surment qui ya des gens beaucoup mieu placer que moi pour te comprendre et ben je peux essaye et pyis meme si je ny arriverai surment pas toujours .. Bah quand on est triste et qu'on se sent seule , c'est bon de savoir quune personne est la pour toi , juste pour toi <3. Bonne nuit dors bien choux je taime (oui parce que en plus detre tactile je suis super affective avec les gens *quelle plait cette <PRE_3> ;)*)
20	15 sept. 2011 09:56:05	151	Héhé on va gagner lipad
21	15 sept. 2011 09:56:30	151	Mon cher et tendre filleul... Je m'engage a etre a tes coté tt au long de ta vie... A te soutenir ds tes moments difficiles et a partager tes bons moments, je souhaite etre la pr toi qd tu en auras besoin... Un soutien, un avis,une Amie... Tu viens de rentrer ds ma famille et j'en suis fière, notre relation va s'epanouir quel bonheur!!! <3 <3 <3 <PRE_4> ta marraine ...
22	15 sept. 2011 09:56:46	151	Se soir tu m'appelle hein chou?
23	15 sept. 2011 09:59:08	151	Ohhhhhh non ! Héhé j'te texto en mode scred mon bbé<3
24	15 sept. 2011 10:10:44	151	Vasi lache moi :o Aller plus tard je taime gros zoubia toi !<3
25	15 sept. 2011 10:42:41	254	Des bisous, des bisous partout!
26	15 sept. 2011 10:54:44	88	Le plus svt possible je raterais peut etre 2 ou 3cours lol
27	15 sept. 2011 10:56:06	88	Ben sa va arrivé alors l'année derniere on avait un decalage aussi
28	15 sept. 2011 10:56:48	88	La j'ai fini je regarde le film
29	15 sept. 2011 10:57:35	88	Coucou! Mon permis c'est bien passé sauf a un rond point ou j'ai un peu foiré :s donc on verra bien. Voila! Bisous a toute ta petite famille
30	15 sept. 2011 10:58:32	136	Je suis vraiment desolée
31	15 sept. 2011 10:58:44	136	Pas de reponse ? J'ai cours dans pas longtemps. Je viendrais a 16h. Jespere que vous ne m'en voulez pas .. fatiguée et j'enchaine les cours non stop jusqu'a <LIE_8> ca vous derange ? Si oui j'arrive mais j'en ai pour au moins une demie heure ... Je suis vraiment desolee ...
32	15 sept. 2011 10:59:06	136	
33	15 sept. 2011 10:59:17	136	<TEL_10> texto <PRE_6> vite vite pour le casting getwild !!!
34	15 sept. 2011 11:00:07	353	Euh tu sais qd <PRE_8> a prévu de revenir ? J'ai pas son numéro...
35	15 sept. 2011 11:00:30	353	Non ça s'est bien passé ;-) . Moi je serai là vendredi t'inquiètes. Bisous
36	15 sept. 2011 11:00:40	353	Bonne nuit!
37	15 sept. 2011 11:01:16	353	Coucou! T'es tentée ce soir par une promenade? Bisous
38	15 sept. 2011 11:01:28	136	Normalement c'est ok pour le weekend prochaine chez <PRE_7>, je peux etre la :)
39	15 sept. 2011 11:01:29	353	Coucou! ça va? Balade vers 5h?

FIGURE 3.34 – Une capture d'écran du fichier « 88milSMS_88522.ods » contenant les SMS constituant le corpus *88milSMS*.

Les analyses en sciences du langage fusent depuis la collecte des SMS authentiques. Celles-ci entrent dans un cadre « guidé par corpus » (« corpus driven », *cf.* tableau 3.17) ou « fondé sur corpus » (« corpus-based » *cf.* tableau 3.18).

Tableau 3.17 – Guidé par corpus/corpus driven.

SMS brut anonymisé	**No du SMS**
Wesh ma vache : -) je lol	13 213
On va rater les bandes annonces espèce de nazgul en tongue lol	902
Wesh gros ! Et bien je sais pas si je pourrai parce que jai ptetre cours, enfin jte dirai ca ce soir ^^	38 593
Espèce de gloutonne des validations d'acquis^^	48 178
Lol, non j't-ai pas oublié !	59 947
Wesh trkl tkt ;) tu fou quoi ?	692

(DÉTRIE et VERINE 2015) ont découvert un phénomène qu'ils ont dénommé « insultes-mots doux » : des SMS incluant des « insultes » (« ma vache », « espèce de nazgul en tongue », « gros » « gloutonne »), radoucies par d'autres éléments textuels (« je lol »), ou des binettes/emojis (« ^^ », «: -) »), etc. Voyant l'utilisation de « wesh », « trkl », « tkt », « lol », etc. dans différents contextes, (MOÏSE 2013a,b) a décidé d'approfondir l'étude sociolinguistique des notions de « norme » et de

Figure 3.35 – Annonce de publication du corpus, juin 2014.

« faute » au sein de l'écriture SMS.

Panckhurst s'est interrogée sur le remplacement de l'accent aigu « é » par « er » ou vice versa : la fouille du corpus a révélé effectivement un grand nombre de formes infinitives en remplacement de formes participiales. Le tableau 3.18 montre les

Tableau 3.18 – Fondé sur corpus/corpus-based.

	SMS brut anonymisé	**No du SMS**
« er » au lieu de « é »	Moi non plus comment ça se fait que tu a changer de Num ?	86 936
	Ehh la j'en peux plus mdrr tu sais je sais plus si je te l'avais dit mais on a enfin acheter la machine a laver aujourd'hui et donc je dis : elle est belle ? Quel est blanche ? elle marche bien ? Et la mon pere me dit : elle est verte !! Nan mais VERTE quoi ! MDRRRR	20 475
« é » au lieu de « er »	J'arive ps a tlavoué mé jsuis tbr sr tn charme	8 530
	Oki :) jvais chercher ca, jte tiens au jus si j'ai réussi a réservé	63 150
mélange des deux	Je me suis levé, j'ai mangé, j'ai révisé mes cours de Jsp, j'ai jouer sur l'ordi, là, je cherche des applis a télécharger sur mon iPod, et toi ?	13 001
	Ma mere a retrouver la housse :) t as retrouve tes rido ?? Ta commencer a tt regroupe ?:)	41 203

> résultats de recherches « fondées sur corpus ». L'exemple suivant, trouvé au hasard d'une lecture, qui constituerait donc plutôt une recherche « guidée par le corpus », semble faire pencher la balance en direction d'une réponse ergonomique (l'accent étant le résultat d'un appui long sur la touche en question) : « Ok t a coter d ki ? » (n° 71 634).
> Les exemples des tableaux 3.17 et 3.18 montrent l'importance d'effectuer un va-et-vient constant entre les hypothèses et l'observation des données. Cela constitue le point essentiel de notre démarche et rejoint la position de (Tagg 2012) et de (Cougnon 2015). (Panckhurst et al. 2016b)

Outre les publications citées précédemment, d'autres travaux ont été menés en sciences du langage, sur des points spécifiques par notre équipe, à partir du corpus *88milSMS* : *apostrophes* (Détrie 2014), *termes d'adresse* (Détrie 2015),*néologie* (Détrie 2016b), *consensus et dissensus* (Détrie 2016a), *norme et faute* (Moïse 2013a,b), *abréviations sémantisées en écriture SMS néographique* (Roche et al. 2016), *genre* (discursif vs. textuel), (Verine 2013, 2015). De plus, j'évoquerai un travail succinct que j'ai effectué sur les binettes/emoji dans la section § 3.4 Réseaux, diffusion et valorisation. En TALNE, fouille d'opinion (*opinion mining*), les publications abondent également, entre autres : *analyse de sentiments* [133] : *SMS et TALNE*

133. (Pozzi et al. 2016) fournissent un état de l'art important, interdisciplinaire, sur l'analyse des sentiments au sein des réseaux sociaux, incluant des approches en TAL de type modèles d'apprentissage sémantiques, etc. Ce sera l'une de mes prochaines lectures.

écriture SMS, intégration des connaissances lexicales et sémantiques (Khiari et al. 2016a,b,c), *alignement* (Kogkitsidou et Antoniadis 2016 ; Lopez et al. 2014), *classification des items inconnus, identification automatique de la créativité scripturale* (Lopez et al. 2015), *néographie et extractions automatisées* (Roche et al. 2016).

Je reprends ci-dessous quelques points issus de notre travail, paru dans *Tranel* (Lopez et al. 2015) qui, je crois, illustre bien notre approche pluridisciplinaire, située quelque part *entre linguistique et informatique*.

> Dans cet article, nous présentons une première approche permettant de classer automatiquement des items inconnus, qui apparaissent dans *88milSMS*, en vue d'une aide à l'analyse de l'eSMS (Panckhurst et al. 2013) et plus précisément à l'identification de la *créativité scripturale* [134]. Nous définissons un item lexical comme l'unité autonome constituante du lexique de l'eSMS (au moins dans le cadre de notre corpus), compris entre deux espaces. Ainsi,« jtrouve » est considéré comme un seul item lexical, alors que « je trouve » est considéré comme une suite de deux items lexicaux.
>
> Une telle ressource contenant tous les items lexicaux SMS « inconnus » du français standard trouve son intérêt à la fois en linguistique et en informatique. D'un point de vue linguistique, cette ressource pourra faciliter l'étude à propos de la créativité scripturale, les mots du discours, l'agglutination, *etc*. D'un point de vue informatique, l'utilisation de la ressource sera utilisée dans la chaîne de traitement automatique des messages de blogs, fora, SMS, et réseaux sociaux. Elle constitue en effet un premier pas vers le transcodage automatique de l'écriture non standard vers l'écriture standardisée (Beaufort et al. 2010) qui permettra d'améliorer la qualité des applications fondées sur un traitement automatique de l'eSMS, par exemple dans un contexte médical (Stenner *et al.*, 2011 ; Vetulani et Marciniak,

134. Dans le cadre de cet article, notre terme créativité scripturale se veut générique et renvoie à différents phénomènes, qui questionnent encore et toujours en sciences du langage : la *néologie* (la créativité lexicale par suffixation (SMS, n° 52041 : « ça se passe bien la *voituration* ? »), mots-valises (« mdr j'avais une réponse bien cinglante, mais rien que de répondre, ça annule la *cinglicité* (?) de la chose… »)), la *néographie* (des variantes de graphies qui constituent des « écarts ludiques » (Anis 1998, p. 132), qui s'éloignent de la langue standardisée et qui sont très présentes et très instables dans l'écriture SMS : abréviations, troncations, notations sémio-phonologiques ou graphies phonétisées, etc.), l'écriture *non-intentionnée* (« fautes » de saisie, etc.). Nous ne prétendons pas répondre à ces questions, notamment concernant la frontière parfois ténue entre néologie et néographie, mais nous n'avons pas besoin d'une distinction fine ici.

> 2011), ou de reconnaissance vocale (Bove, 2005). (Lopez et al. 2015)

L'objectif de cet article était d'aider l'utilisateur à identifier les items non standard (INS) dans le corpus *88milSMS*, dans le sens où ces items n'existent dans aucun dictionnaire de langue française.

> Parmi les INS, nous proposons d'identifier les items non standard originaux (INSO). Un traitement manuel serait complexe principalement à cause de deux points :
>
> — la définition d'un « item non standard original »,
>
> — la taille du corpus, supérieure à un million d'items lexicaux.
>
> La question de l'originalité d'un item est largement discutable selon que l'on s'intéresse aux variations lexicales/scripturales, aux créations de termes, ou encore à l'alternance codique, par exemple. Afin de ne pas biaiser l'interprétation de la ressource produite et ne pas contraindre son utilisation à une application donnée, nous avons considéré que les INSO sont des items lexicaux que nous ne sommes pas en mesure de classer de façon triviale (par horaires, pseudonymes, termes du français, *etc.*). L'hypothèse sous-jacente est de considérer que les items n'ayant pu être classés dans les catégories prédéfinies sont potentiellement des INSO.
>
> (Lopez et al. 2015)

Après une classification par langue (*français, anglais, espagnol, allemand, italien*) (*cf.* (Lopez et al. 2015) pour de plus amples détails), une approche pour permettre d'identifier les items non standard (INS) a été adoptée. Elle fonctionne de la manière suivante :

> Notre approche d'identification d'INS consiste à fournir le corpus *88milSMS* en entrée du système et à obtenir en fin de traitement un ensemble de classes permettant d'aider à l'identification automatique de la créativité scripturale. Le système est développé en Java.
> Le corpus *88milSMS* est d'abord segmenté. Les segments habituellement considérés dans les approches de classification sont les mots ou bien les phrases. La segmentation par phrases n'est pas pertinente ici car notre objectif est d'identifier un ensemble d'items. Aussi, dans le contexte de la segmentation de SMS, il ne semble pas pertinent de prendre le mot comme segment puisque le lexique utilisé pour la rédaction de ces messages n'est pas formellement défini. De plus,

il est complexe d'identifier automatiquement les frontières des mots au sein d'une chaîne de caractères issue de données textuelles de type SMS (par exemple « a2min lami » = « à demain l'ami »). Notre objectif étant d'identifier des items lexicaux non standard, nous considérerons donc qu'un segment, ou item lexical, est une suite de caractères compris entre deux espaces (dans l'exemple précédent nous obtenons ainsi deux segments :« a2min » et « lami »). Notons qu'un prétraitement a consisté à ajouter une espace avant et après chaque élément de ponctuation lorsque ce dernier était absent. Au total, nous obtenons ainsi plus d'un million d'items lexicaux.

Notre approche consiste à déterminer, dans un premier temps, trois ensembles distincts, que nous nommerons « classe » : C1, C2 et C3.

La classe C1 recevra les items standard, c'est-à-dire les items présents dans le LEFFF, avec et sans accents. C2 recevra les INS reconnus grâce à des filtres que nous définissons ci-après, et C3 recevra tous les items qui n'ont été classés ni dans C1 ni C2 et qui peuvent donc correspondre à une forme de créativité scripturale non retenues par nos filtres. Les classes C1, C2, et C3 sont disjointes, *i.e.* un même item ne peut apparaître dans deux classes différentes.

L'objectif du travail étant d'identifier automatiquement les INS pour le français, nous cherchons, en premier lieu, à élaguer l'ensemble des items de *88milSMS* qui seraient également présents dans le LEFFF. Ainsi, la classe C1 contenant les items standard est constituée de deux sous-classes :

C1.1 : items standard présents dans le LEFFF.

Le filtre consiste ici à comparer un à un les items français de *88milSMS* avec les items du LEFFF. Les items présents dans le LEFFF sont attribués à la classe C1.1.

C1.2 : items standard présents dans le LEFFF sans accents.

Nous mettons en place un filtre permettant de comparer les items avec les mots du LEFFF auxquels nous avons supprimé les accents. La classe C1.2 accueille donc les items correctement orthographiés selon les normes du français standard mais dont l'accentuation est absente (par exemple : *qualites*, *degat*, *precisions*, *europeen*).

Cette première étape a permis de construire une approche automatique qui catégorise les items standard. Dans la suite, nous proposons une sous-catégorisation des items non standard (INS) de C2 (items non contenus dans C1) :

— ***C2.1 : items composés d'un caractère unique.*** Cette sous-classe contient les items constitués d'un seul caractère, incluant les caractères spéciaux, les chiffres, lettres, *etc.* Une telle classe est par exemple utile pour l'étude des abréviations sémantisées telles que *c* pour *c'est/ces/ce…* ou *t* pour *t'es/tu…*.

— ***C2.2 : items assimilables à des horaires.*** Cette sous-classe contient les items représentant une heure, ou plus généralement un rapport avec le temps. Le filtre correspondant est une succession de tests recherchant la présence d'une suite de caractères spécifiques telle qu'un chiffre suivi de la lettre « h » ou des lettres « min ». Par exemple, nous identifions *12h30*, *23 : 56*, *8heures*, *10minaperdre*, *6-7h, etc*.

— ***C2.3 : item avec allongement.*** Les termes de cette sous-classe ont subi une répétition de caractères, qui simule un allongement vocalique, et ce sur au moins un caractère (par exemple : *Jarriiiiiiiiive*, *Huuuummm*, *Meeerciiiiii*, *tkkkkt*). Le filtre mis en œuvre compare chaque caractère avec le caractère suivant. Si plus de deux caractères sont répétés, l'item est classé dans C2.3. Rappelons que les mots possédant deux mêmes caractères consécutifs issus de la langue standard (par exemple, passe, embrasse, apprendre) ont précédemment été classés dans C1. Si nous considérons qu'un allongement est un critère répondant à l'originalité des items, alors il faut considérer que C2.3 contient bon nombre d'INSO.

— ***C2.4 : item avec caractère spécial.*** Nous testons ici la présence d'un caractère spécial dans chaque item. Les caractères spéciaux considérés sont tous les caractères d'un clavier alphanumérique AZERTY classique, hors chiffres et lettres (30 caractères spéciaux au total). Les items de la classe C2.4 contiennent au moins un caractère spécial (par exemple : *Conn*rd*, *resto+cine*, *appeler/texto*, *dés˜annule*, *thèse/antithès/synthèse*, *fish&chips*). Cette sous-classe contient les binettes contenant un ou plusieurs caractères spéciaux (par exemple ^^ ou bien ;)).

— ***C2.5 : présence d'un chiffre.*** La sous-classe C2.5 contient tous les items incluant un chiffre (qui n'ont pas été précédemment repérés, par exemple, dans C1). Le filtre correspondant teste simplement la présence d'au moins un chiffre au sein de l'item. Nous obtenons par exemple, *numb3rs*, *mc2*, *106ounette*, *3615ma-vie*, *Ar5ggggggggh*. Ces items peuvent dès lors être considérés comme des INSO dont l'originalité réside dans la présence de chiffres.

— ***C2.6 : binettes*** [135]. Cette sous-classe contient les binettes d'après une liste construite en deux temps : 1) binettes acquises sur le Web [136], 2) binettes ajoutées manuellement d'après notre expertise sur le corpus. Les binettes inconnues de

135. « Binette » est le terme (québécois) que nous utilisons pour évoquer « smiley », « émoticône », « frimousse », par exemple :«;) »,« ^^ »,«;) »,«: D », etc. Dans un travail ultérieur, nous effectuerons le classement des « emoji » (les binettes graphiques) qui nécessitent un repérage Unicode.

136. https://support.skype.com/fr/faq/FA12330/qu-est-ce-que-la-liste-complete-d-emoticones, consulté le 11 janvier 2017.

> notre liste pourront être découvertes dans la sous-classe C2.4. Au total, nous disposons d'une liste de 54 variantes de binettes (par exemple «:-) » et « +.+ »). Cette liste ne contient aucun « emoji » (émoticône graphique).
> Enfin, la classe C3 contient les items qui ne sont ni dans C1 ni dans C2.
> — ***C3 : items non standard originaux (INSO).*** Cette catégorie contient les items qui n'ont pas été classés dans les catégories précédentes et ne nécessite donc pas la mise en place de filtre spécifique. Nous obtenons ainsi des items néologiques tels que *cinglicité*, *voituration* ou encore des items néographiques agglutinés tels que *tatende*, *tetrangle*. Les items présents dans C3 sont donc potentiellement des items non standard originaux (INSO).
> Il est important de noter que les sous-classes de C2.1 à C2.6 ne sont pas disjointes : plusieurs sous-classes peuvent contenir un même item. Par exemple, *Ar5ggggggggh* doit être classé dans C2.3 et C2.4.
> Nous avons défini 3 classes et 8 sous-classes qui représentent l'ensemble des items présents dans les SMS « français » identifiés à l'étape précédente (§ 2.1). D'autres classes et sous-classes peuvent être ajoutées dans le but de classer plus finement les items en fonction des objectifs visés. (Lopez et al. 2015)

Les classifications sont schématisées dans la figure 3.36.

Les résultats chiffrés correspondant à cette extraction sont présentés dans le tableau 3.19 (Roche et al. 2016) : pour chaque classe C, le nombre d'items différents, le nombre d'occurrences totales, ainsi que l'item ayant le nombre d'occurrences maximales.

Pratiques scripturales de *88milSMS*

Comme (Anis 1998), je désigne par *néographie* des variantes de graphie qui s'éloignent de la langue standardisée, souvent de manière délibérée, ludique, et qui sont très présentes et instables dans l'écriture SMS. Suite à (Lopez et al. 2015), nous avons voulu mieux comprendre quelles pratiques scripturales étaient les plus redondantes au sein du corpus *88milSMS*, et ce à partir de l'extraction informatique menée précédemment. La recherche que nous avons présentée au colloque *CINEO* à Salamanque en octobre 2015, portait également sur une partie de la classe C2.1 (tableau 3.19), afin d'analyser l'usage des lettres uniques

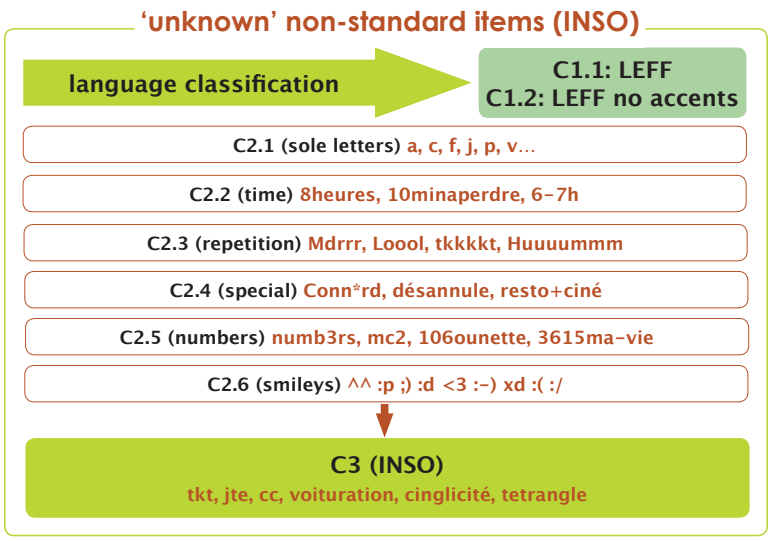


FIGURE 3.36 – Classification à partir de l'INS vers l'INSO.

qui renvoient à des mots, tels que 'c' pour c'est/ces/ce, ou 't' pour t'es/tu, etc. (*Cf.* (ROCHE et al. 2016) pour le détail de cette recherche).

Dans le tableau 3.20, on retrouvera les pratiques scripturales les plus récurrentes par classe. Parmi celles de la classe C3, et correspondant à ma typologie néographique (*cf.* la figure 3.10), on retiendra ce qui suit, par ordre décroissant :

1. Réduction phonétisée : acronyme : *lol*
2. Réduction graphique : agglutination : *jte, jsuis, jvais*
3. Suppression graphique : élision typographique/ponctuation : *cest, weekend, Jai*
4. Réduction graphique : squelette consonantique : *Dsl, avc, Cc*
5. Réduction phonétisée : troncation : *week*
6. Substitution graphique : signes diacritiques : *méme, étre*
7. Ajout graphique : signes diacritiques : *çà*
8. Ajout graphique : onomatopées : *Beh*
9. Substitution phonétisée avec variation : *Oue*

Tableau 3.19 – Formes et occurrences des classes C. (Roche et al. 2016)

Catégorie	Nombre d'items différents	Nombre d'occurrences totales	Item ayant le nombre d'occurrences maximales pour chaque classe
C1.1 : LEFFF	22 213	1 036 226	***de*** 24 907
C1.2 : LEFFF sans accents	2 526	39 791	***a*** 16 017
C2.1 : caractères uniques	125	390 072	**.** 74 922
C2.2 : horaires	500	4 837	***19h*** 172
C2.3 : allongement	2 537	6 309	***Mdrrr*** 741
C2.4 : caractères spéciaux	484	288 335	**.** 74 922
C2.5 : chiffres	1 971	16 802	**<3** 1 540
C2.6 : binettes	54	26 726	**:)** 6 704
C3 : items non standard originaux (INSO)	17 882	58 964	**lol** (3 341)

Les phénomènes complexes les plus récurrents sont les suivants, pour cette même classe C3 :

1. Réduction graphique : squelette consonantique/abréviation + Substitution phonétisée partielle : *tkt (= t'inquiète [pas]), pk (parce que / pourquoi)*
2. Suppression graphique : typographie + suppression graphique : fin de mot muette + substitution graphique avec variation : *cei*

Les autres classes indiquent une suppression graphique très importante de signes diacritiques (C1.2), une réduction en caractères uniques (C2.1), ainsi que des ajouts, correspondant à des répétitions de caractères afin de simuler des allongements vocaliques (C2.3), voire consonantiques (problématisés par (Ghliss et Verine 2016)) et des binettes (C2.6). (André 2014) a effectué une analyse comparative sur un échantillon de SMS extraits des corpus *88milSMS* (12 659 SMS) et *SMS4science* (1 165 SMS). Il a constaté « une large utilisation de phénomènes de variation d'ordre graphique modifiant la typographie et la ponctuation » (André 2014, p. 94), ce qui concorde avec les données du corpus entier. Entre 2004 (collecte belge *SMS4science*) et 2011 (*sud4science/88milSMS*), « la part d'utilisation des phénomènes d'ordre phonétique s'est réduite, mais la classe des moins de 18 ans regroupe toujours les scripteurs les plus enclins à y avoir recours, en vue

Tableau 3.20 – Items les plus fréquents par classe. (Roche et al. 2016)

C1.1 items standard présents dans le LEFFF :	***de*** **(24 907),** ***pas*** **(23 558),** ***je*** **(22 825),** ***est*** **(18 658),** ***la*** **(16 652),** ***tu*** **(15 956),** ***le*** **(15 892),** ***et*** **(15 510),** ***que*** **(15 492),** ***c*** **(12 862).**
C1.2 items standard présents dans le LEFFF sans accents :	***a*** **(16 017),** ***meme*** **(2 060),** ***etre*** **(1 762),** ***A*** **(1 628),** ***aprem*** **(641),** ***etait*** **(547),** ***apres*** **(426),** ***o*** **(385),** ***bientôt*** **(376),** ***tete*** **(330).**
C2.1 items composés d'un caractère unique :	**. (74 922), ' (68 107), ! (44 272), ? (34 477), , (23 812), a (16 017), " (15 076),** ***c*** **(12 862),** ***j*** **(12 429),** ***à*** **(12 343).**
C2.2 items assimilables à des horaires :	***19h*** **(172),** ***20h*** **(164),** ***18h*** **(163),** ***17h*** **(157),** ***16h*** **(142),** ***2h*** **(133),** ***1h*** **(127),** ***15h*** **(120),** ***8h*** **(113),** ***10h*** **(112).**
C2.3 item avec répétition :	***Mdrrr*** **(741),** ***mdrrr*** **(239),** ***bb*** **(153),** ***Mdrrrr*** **(122),** ***Biisoux*** **(67),** ***Loool*** **(66),** ***Ptdrrr*** **(7),** ***Ahh*** **(61),** ***Waii*** **(58),** ***Aaah*** **(57).**
C2.4 item avec caractère spécial :	**. (74922), ' (68107), ! (44272), ? (34477), , (23812), " (15076), :) (6704), ^^ (5624), - (4444),) (2538)**
C2.5 présence d'un chiffre :	***<3*** **(1540),** ***2*** **(1318),** ***3*** **(791),** ***1*** **(673),** ***5*** **(429),** ***4*** **(408),** ***10*** **(347),** ***15*** **(281),** ***30*** **(224),** ***20*** **(204).**
C2.6 binettes: spécial :	**:) (6704), ^^ (5624), ;) (2478), :D (1875), :-) (1735), :P (1382), :p (1336), :((1102),:/ (631), ;-) (541)**
C3 items non standard originaux (INSO) :	***lol*** **(3972),** ***tkt*** **(926),** ***jte*** **(858),** ***jsuis*** **(616),** ***week*** **(552),** **çà (459),** ***méme*****(306),** ***cest*** **(283),** ***Dsl*** **(264),** ***Beh*** **(263),** ***weekend*** **(253),** ***Pk*** **(248),** ***avc*** **(229),** ***Jai*** **(217),** ***cei*** **(204),** ***Cc*** **(196),** ***étre*** **(192),** ***facebook*** **(190),** ***Oue*** **(188),** ***jvais*** **(185).**

d'accentuer la portée ludique (voire cryptique) de leurs messages ».

(Roche et al. 2016)

3.3.3.11 Entre linguistique et informatique : applications

À l'avenir, avec mes collègues informaticiens, Mathieu Roche et Cédric Lopez, nous aimerions également approfondir nos recherches sur les applications informatiques envisageables :

> Outre les étapes d'anonymisation, de transcodage, d'annotation fondées, en partie, sur des techniques de TAL, d'autres applications informatiques sont envisageables : élaboration de lexiques transcodés français standardisé =>SMS ou vice versa, consultables en ligne ; mise en place de systèmes de vocalisation des SMS à l'usage de personnes aveugles ou de personnes momentanément empêchées de consulter leur écran de téléphone – en situation de conduite, etc.
>
> (Panckhurst et al. 2016b)

Nous pourrions également approfondir la question des techniques de normalisation automatique. Après une étude des phénomènes en linguistique liés à l'écriture SMS, dans (Lopez et al. 2014) il s'agissait d'étudier les phénomènes, puis le transcodage à partir de SMS « bruts » vers des SMS en français « standardisé » afin d'améliorer les résultats dictionnairiques. Pour les détails informatiques, on se reportera à Lopez et al (2014). Je fournis quelques exemples posant problème ci-dessous d'un point de vue linguistique et je les discute :

SMS « brut » (n° 32066, corpus *88milSMS*) :
*Vien a 17h10 pask **g pa** commencé encor*
SMS « transcodé » :
Viens à 17h10 parce que j'ai pas commencé encore

Dans cet exemple, on peut enrichir les items lexicaux apparaissant dans le dictionnaire : « parce que » =>« pask » ; « pas » =>« pa » ; « encore » =>« encor », etc. Le verbe *venir*, quant à lui, pose un problème d'ambiguïté, car *vien* ne correspondra pas systématiquement à la 1e ou la 2e personne, mais parfois à la 3e personne (ex, n° 29356 : *On vien d ariver*). De même, on ne peut effectuer un transcodage automatique entre l'usage d'un auxiliaire et d'une préposition (« a », « à »), seul le contexte phrastique permet de déterminer quel est le candidat approprié. Bien entendu, les SMS renferment des variantes graphiques très riches. Par exemple, dans notre corpus, « aujourd'hui » est graphié de 18 manières distinctes : *aujourd'hui, ajourd'hui, aujd, auji, aujii, aujiurd'hui, aujiurdhui, aujoirdhui, aujord'hui, aujordhui, aujordui, aujoud'hui, Aujoud'hui, aujourd'hui, aujourdghuo, aujourdhhui, aujourdhui, aujourfui*. Il ne s'agit pas de répertorier toutes ces variantes qui peuvent évidemment changer d'un corpus à l'autre, mais de mettre en place des priorités, via des résultats statistiques. Dans *88milSMS*, la graphie *aujourd'hui* est utilisée à 87 %, suivie de la forme agglutinée *aujourdhui* (9 %), puis de la forme *aujd* (1 %). Malgré tout, au-delà des analyses à des fins dictionnairiques, un approfondissement de ces recherches permettrait de mettre l'accent, dans une perspective sociolinguistique et à travers des variantes lexicales ciblées, sur le changement linguistique et les usages en cours dans les pratiques graphiées.

Les apocopes (SMS n° 33727, *les **appli** sont pas encore a jour*) et les aphérèses (SMS n° 47607, *bon allez espère que ta flemme s'est arrangée un peu.. Un**zou***) constituent

d'autres phénomènes intéressants à répertorier, car ils peuvent être utiles lors d'analyses automatiques ou de recherches dictionnairiques. Par exemple, le *Petit Robert 2016 en ligne* permet de retrouver « application » à partir de la forme en apocope « appli(s) », mais non pas « bisou » à partir de « zou » (l'entrée « zou » existe néanmoins, en tant qu'interjection-onomatopée).

Enfin, l'étude des formes élidées et les agglutinations permettra également de faire avancer les recherches dictionnairiques, ou moins pour en exclure certaines formes (SMS n° 47012, *Je c pa **jtapel** avant de sortir du gineco)*. Toutes les formes d'agglutination ne pourront bien sûr être étudiées, car la créativité lexicale est extrêmement riche au sein d'un corpus de textos :

SMS n° 38843, *bisoutoucalinourienkepourtoipuissance* ;

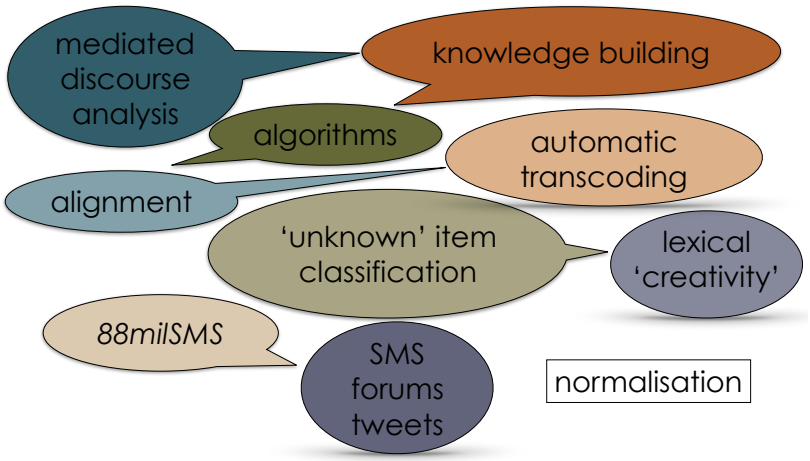
SMS n° 66143, *frontenormeetjouesdehamsterjovial)*.

Un approfondissement des recherches menées jusqu'à présent permettrait une meilleure compréhension des données textuelles en graphie « non-standardisée », constituant aujourd'hui un frein au moment d'analyser automatiquement le contenu du Web. Le transcodage utilisé devrait également conduire à un étiquetage catégoriel et une analyse syntaxique plus robustes.

Enfin, comme expliqué précédemment, nos publications les plus récentes (Panckhurst 2016a), (Panckhurst et al. 2016b), (Lopez et al. 2016) évoquent, entre autres, les choix théoriques concernant la décision de ne pas transcoder, ne pas annoter linguistiquement, l'ensemble du corpus *88milSMS*, mais seulement fournir des échantillons (1 000 SMS transcodés, 100 SMS annotés avec nos balises linguistiques), pour comprendre les enjeux et problèmes. Cela n'exclut pas nos recherches en techniques de TAL, permettant l'implémentation ultérieure d'applications réelles — au contraire :

> This theoretical position does not exclude exploring Natural Language Processing (NLP) investigation techniques, which could then be implemented in real-life applications. Examples of investigation techniques are indicated as follows : 1) Our corpus can be used to analyse current mediated electronic discourse, and help build knowledge on different SMS writing forms (Roche et al. 2016). 2) Algorithms may be used to learn from this : alignment methods for facilitating automatic transcoding have been explored (Aw et al. 2006 ; Beaufort et al. 2008 ;

Guimier de Neef et Fessard 2007 ; Kobus et al. 2008 ; Lopez et al. 2014). 3) We have devised a method for classifying « unknown » items within text messages, which may help to automatically identify lexical « creativity » within *88milSMS* and improve electronic dictionary approaches. (Lopez et al. 2015).



In order to refine automatic normalisation techniques for initially non-standard texts in French, the next logical step is to compare our resource with different types of instant media (i.e. SMS, forums, tweets). Firstly, a new typology of the detected « mistakes », based on existing typologies, will be elaborated. Secondly, automatic normalisation techniques — focussing on the most frequent errors — will be proposed. These will then be confronted with traditional automatic translation (Vilariño et al. 2012), speech recognition (Kobus et al. 2008) and spelling/grammatical checker principles (Beaufort et al. 2010). Finally, the approach should enable comparison between different types of instant media. (Lopez et al. 2016).

automatic normalisation techniques

› new typology of detected 'mistakes'
› normalisation based on most frequent errors
› confrontation with:
traditional automatic translation,
speech recognition,
spelling/grammatical checker principles
› comparison between different types of instant media (SMS, forums, tweets)

3.3.3.12 Conclusion

J'annonçais à la fin du volet 2 que le volet 3 serait l'occasion d'explorer de manière plus approfondie le lien entre *données authentiques*, *analyse du discours électronique/numérique médié*, et *mutation(s) des pratiques*.

En effet, depuis deux décennies, je m'intéresse à la communication « médiée » par ordinateur voire le « discours électronique médié », ou, suite à une actualisation terminologique, « discours numérique médié » (DNM). À mon sens, ce travail permet d'établir un lien direct entre mes recherches précédentes fondamentales en traitement automatique du langage et le domaine des TICE et de la FOAD. L'étude de l'évolution du langage m'a permis d'effectuer des analyses linguistiques et informatiques de courriers électroniques, de forums de discussion, de messageries instantanées (ou chats), et, plus récemment, d'écriture de type SMS (eSMS), qui est vraiment très riche, très innovante et également très créative.

Dans la figure 3.9 on pouvait constater l'évolution de l'utilisation de quatre catégories grammaticales (N, V, Adj, Adv) entre les corpus évoluant du courriel au chat, en passant par le forum de discussion (Panckhurst 2006 ascilite, 2009 sms). J'y ai ajouté une petite analyse effectuée à l'aide du logiciel *Cordial* (*cf.* la figure 3.38) pour les 30 000 SMS du corpus *SMS4science*, ainsi que pour le corpus des SMS conversationnels (p. 176) et, plus récemment, un échantillon de 1 000 SMS extraits du corpus *88milSMS* (Panckhurst et al. 2014a).

Outre le fait que j'ai évolué d'une recherche au sein de laquelle la donnée de type *exemple* était privilégiée (volet 1, § 3.3.1) à des *situations authentiques* dans lesquelles étaient investis les acteurs enseignants/chercheurs/étudiants (volet 2, § 3.3.2) puis à une recherche intégrant pleinement le recueil et l'analyse de *données authentiques* (volet 3, § 3.3.3), j'ai fait un cheminement dans ma façon

- **graphie très variable** : *aujourd'hui*, *ajourd'hui*, ***aujd***, *auji*, *aujii*, *aujiurd'hui*, *aujiurdhui*, *aujoirdhui*, *aujord'hui*, *aujordhui*, *aujordui*, *aujoud'hui*, *Aujoud'hui*, *aujourd'hui*, *aujourdghuo*, *aujourdhhui*, ***aujourdhui***, *aujourfui*.
- **squelettes consonantiques :** *slt*, *dsl*
- **apocopes** : *les **appli** sont pas encore a jour*
- **aphérèses** : *bon allez espère que ta flemme s'est arrangée un peu.. Un **zou****
- **élision/agglutination** : *Je c pa **jtapel** avant de sortir du gineco*
- **suppressions de fins de mots muettes** : *vou*
- **abréviations sémantisées** : *tu **f** koi ? (fais/feras/faisais/fous/foutais)*
- **substitutions phonétisées plus ou moins complexes :** *koi, boC, 2m1*
- **répétitions/ajouts de caractères** *: suuuuppppeeerrrr, les zamours, oki,*
- **binettes/emoji** ^^ :)
- **créativité lexicale riche** : *bisoutoucalinourienkepourtoipuissance frontenormeetjouesdehamsterjovial*

Figure 3.37 – Exemples d'écriture SMS (eSMS)

d'aborder le TAL également.

Au départ de ma carrière d'enseignant-chercheur, j'étais très impliquée dans la programmation, dans l'implémentation d'outils de TAL (volet 1, § 3.3.1). Puis, cela s'est assoupli vers les usages par autrui, l'évaluation de logiciels, la formation, etc. (volet 2, § 3.3.2). Maintenant, comme je l'ai déjà indiqué, je préfère laisser les aspects de programmation aux experts-informaticiens qui m'entourent (volet 3, § 3.3.3). Mais cela ne m'empêche pas de maintenir des dialogues nourris avec les informaticiens, car même si je ne programme plus sérieusement comme autrefois, j'ose espérer que j'ai encore les « réflexes » d'une linguiste-informaticienne.

Être résolument située entre linguistique et informatique. Hybride. Mais « une hybridité joyeuse », si j'ose dire, car ma situation entre-deux m'a toujours fourni un joker pour être *autre*, peut-être presque « électron libre ».

L'apogée de ma carrière a débuté en 2011, avec les recherches menées avec mes collègues Catherine, Cédric, Claudine, Mathieu, Bertrand (p. 186, § 3.4.2). Un bonheur sans cesse renouvelé. La recherche comme j'aime. Travailler sur un même objet, d'une manière pluridisciplinaire, où chacun apporte énormément aux autres, dans un respect total, avec les mêmes exigences scientifiques de rigueur. Rare. Très rare.

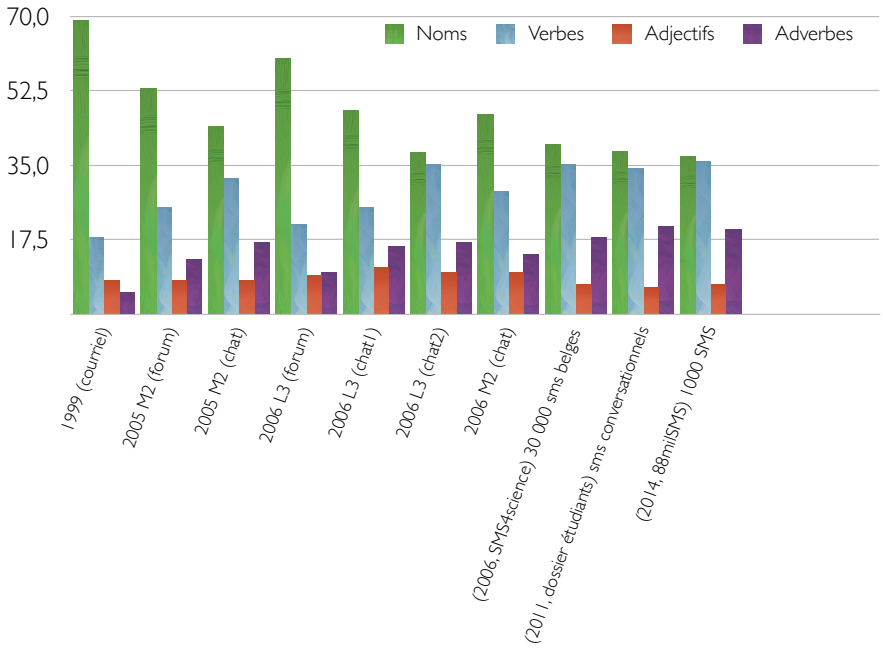


FIGURE 3.38 – Catégories syntaxiques (occurrences) utilisées (courriel, forums, chats, SMS) entre 1999 et 2014.

Coïncidence chronologique ? L'année 2011 marquait également mes 50 ans : j'estimais donc être suffisamment mûre pour décider de faire de la recherche pour mon unique plaisir. Je n'avais plus rien à prouver à quiconque. En effet, je dois dire que depuis le début de l'aventure en 2011, jusqu'à ce jour, les échanges scientifiques menés dans ce cadre, ont été les plus enrichissants de ma carrière. Tout me passionnait : communication nécessaire à l'extérieur de l'université (médias et commerces) ; travail pluridisciplinaire entre (enseignants-)chercheurs, étudiants, juristes ; recueil de données authentiques auprès du grand public (SMS et réponses à un questionnaire sociolinguistique) ; suivi, conseils et questionnements légaux ; anonymisation, transcodage, annotation, analyse de données.

Avec le recul, l'apogée de ma carrière — en tout cas, jusqu'à présent — avec l'aboutissement de ces projets de recherche autour des SMS, a sans doute pu être

effective, grâce à tout le chemin parcouru pendant de longues années en amont.

Avant 2011, je n'aurais pas imaginé qu'un travail réellement pluridisciplinaire, tel que je l'entends — travailler sur un objet commun, en respectant des approches disciplinaires diverses, dialoguer tout en acceptant de se remettre en question, faire parfois des compromis — puisse être aussi plaisant. Avant 2011, je n'aurais pas été prête pour aller convaincre des commerçants d'adhérer à notre projet de collecte, en faisant des dons (*cf.* (Panckhurst et al. 2014b,c)). Avant 2011, je n'aurais pas songé à l'idée de proposer à une entreprise informatique de s'associer à notre collecte via un prêt de smartphone, entre autres. Avant 2011, je n'aurais pas pris plaisir à apprendre différentes choses sur le plan juridique (*just the tip of the iceberg, mind you!*). Avant 2011, je n'aurais pas imaginé partager mon bureau avec des étudiants stagiaires pendant plusieurs mois. Avant 2011, je n'aurais jamais anticipé que nous arriverions à recueillir autant de SMS, ni même que le traitement d'anonymisation serait aussi long, malgré l'apport d'un logiciel de traitement semi-automatisé. Avant 2011, je n'aurais pas deviné qu'il y aurait un tel engouement médiatique à propos de notre projet, ou une telle sollicitation pour des entretiens, et cela continue aujourd'hui (*cf.* § 3.4 Réseaux, diffusion et valorisation). En somme, avant 2011, je n'aurais pas pu rêver mieux en recherche : avoir à sa disposition un corpus de données authentiques — qui nous tiendra en haleine encore longtemps ; pouvoir travailler et partager des moments de pur bonheur scientifique avec des collègues et des étudiants passionnants et passionnés, dans un contexte pluridisciplinaire détendu et fructueux ; remettre à la disposition du grand public et de la communauté scientifique les résultats de nos travaux de recherche et le corpus *88milSMS* ; assouplir notre licence en *Creative Commons* pour que le plus grand nombre puisse utiliser (encore plus) librement notre corpus ; prévoir des applications en TAL qui seraient — espérons-le, car ce serait effectivement mon rêve — directement utilisables par le grand public.

Ce n'est que maintenant, en effectuant ma rédaction pour l'habilitation à diriger des recherches, que je prends le temps nécessaire pour ce recul. Comme après un excellent dîner, je suis repue. J'ai l'impression que toutes mes recherches précédentes ont constitué des entrées en matière, et mon vrai plat principal — qui est le travail autour du corpus *88milSMS*, et, plus largement, l'écriture au quotidien, en évolution/mutation constante — est à peine entamé et sera

une source inépuisable pour la recherche pendant longtemps. Je crois que nous avons encore de longues années devant nous pour continuer à casser la croûte. Je proposerai d'autres menus (*cf.* § 4), mais avant cela, je me permets encore quelques tapas (*cf.* § 3.4 Réseaux, diffusion et valorisation).

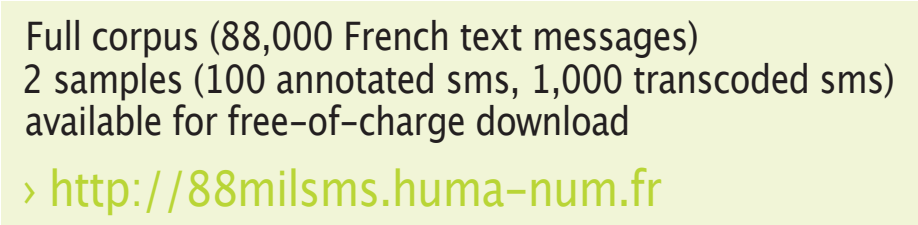


Figure 3.39 – 2014 : *88milSMS*

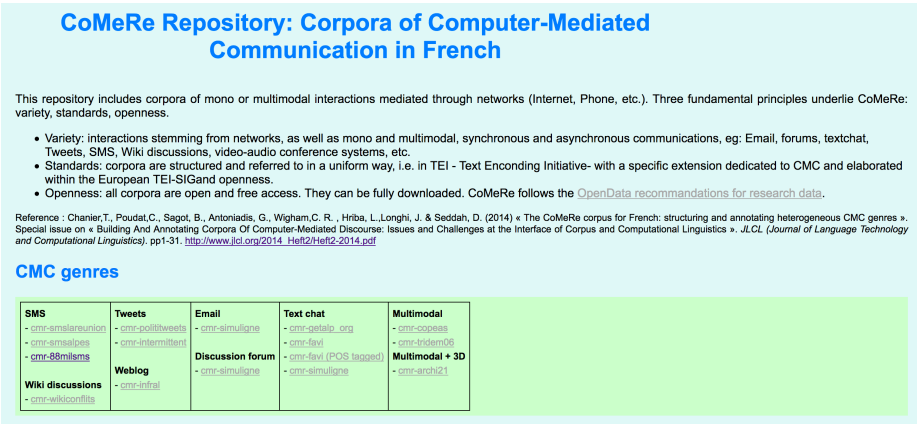


Figure 3.40 – 2016 : le corpus *88milSMS* v2 en XML/TEI est inclus dans *CoMeRe* et incorporé à la plateforme *Ortolang*.

3.3.3.13 Encadrement spécifique : volet 3

Mes implications d'encadrement d'étudiants ont continué. J'ai participé à un jury de doctorat (avril 2017) à la Sorbonne. J'ai (co-)dirigé les 17 mémoires de Master suivants dans le cadre du volet 3, entre 2006 et 2014, et j'ai co-encadré 8 stagiaires pour le projet *sud4science LR*, entre 2011 et 2013, tantôt avec mes collègues linguistes tantôt avec mes collègues informaticiens. J'ai continué de participer aux jurys de soutenance pour les Masters de M1 & de M2. Le détail de ces encadrements apparaît dans le tableau 3.21.

Tableau 3.21 – Encadrement de la recherche : volet 3*

Participation à jury de doctorat	
André Frédéric, « Pratiques scripturales et écriture SMS : analyse linguistique d'un corpus de langue française », Doctorat de Sciences du langage, université Paris-Sorbonne, Jury : C. Fairon, E. Stark, R. Panckhurst, S. Plane, G. Siouffi (directeur). Soutenance le 24 avril 2017.	2016-2017
Comités de suivi de thèses	
Michel Otell : « De quelques processus de production du sens dans les SMS conversationnels des sourds signants ».	2015-2016
Michel Otell : idem.	2014-2015
13 Directions de mémoires	
M2, Sciences du langage, « Comment comprendre l'acquisition et l'évolution du langage chez l'individu utilisant les nouveaux moyens de communication (sms, internet) », (C. Llorach).	2016-2017
M1, Sciences du langage, Discours médiatiques, institutionnels et politiques : « L'inscription de la complicité dans les SMS » (C. Luong).	2014-2015
1) M2, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « Écriture SMS et Néographie chez les jeunes de 11-30 ans » (G. Porto). 2) M1, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « Caractéristiques propres au scripteur et au Smartphone impactant l'écriture SMS » (J. Laboureau). M1, Gestion des connaissances, formations et médiations numériques : 3) « Camfranglais : Pratiques et usages plurilingues dans les SMS au Cameroun » (A. Moussa). 4) « La reconnaissance vocale sur mobile : lécritur d sms, 2 lécri a Loral » (A.Kaba).	2013-2014
M1 Gestion des connaissances, formations et médiations numériques : « Plurilinguisme et SMS » (N. Dos Santos).	2012-2013
M1 : « L'expression des émotions et des sentiments dans Twitter et les SMS : analyse comparée des usages, des formes et des objectifs » (E. Orlando).	2011-2012
« Conversations asynchrones en communication médiée : comparaison entre courriels, forums et SMS. » (D. Fontaine).	2008-2009
M1 : 1) « SMS et déficiences visuelles » (A. Gaussuron).	2007-2008
M1 : « Le langage SMS des jeunes. Approche lexicale et morpho-syntaxique » (M. Fayada).	2006-2007

M1 : 1) « Nouveaux usages de communication électronique en russe. » (N. Svarinska). 2) « Étude lingustico-communicationnelle des SMS en France et en Bulgarie. » (K. Filipov).	2005-2006
4 Co-directions de mémoires	
M2, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « L'interjection dans les SMS : usages et tendances scripturales », (J. Laboureau, co-directeur : L. Fauré).	2014-2015
M2, Sciences du Langage, Discours médiatiques, institutionnels et politiques : « Écriture SMS et Phénomènes Phonétiques. Évolution des pratiques scripturales entre 2004 et 2011 » (F. André, co-directeur : F. Hirsch).	2013-2014
M1 : « Conversations et SMS » (H. Catapano, co-directeur : L. Fauré.	2008-2009
M1 : « Discours électronique médié et SMS : étude phonologique de quelques morphèmes verbaux » (O. Caumont, co-directeur : L. Fauré.	2006-2007
4 Participations à jurys de soutenance	
M1 Recherche, « Interdiscours et créativité dans les slogans de manifestations » (Y. Ghliss, Département de Sciences du Langage, directeur : B. Verine).	2012-2013
M1 : Participation à jurys, journées pour l'ensemble de la promotion « Gestion des connaissances et formation ouverte et à distance », MSH, Montpellier.	2006-2010
1 participation à jury de soutenance, M2.	2004-2005
M1 Recherche : « Outils pour l'analyse automatique du discours » (S. Riou, Département de Sciences du Langage, directeur : P. Siblot). (Volets 1 et 3)	2003-2004
8 stagiaires ayant travaillé dans le cadre du projet de recherche SMS *sud4science/DGLFLF (volet 3)*	
Co-encadrement de cinq stagiaires de M1/M2, programme *sud4science* et *DGLFLF* : Camille Lagarde-Belleville, Michel Otell, Frédéric André, Yosra Ghliss, Reda Bestandji.	2012-2013
Co-encadrement de deux stagiaires de M2 (LIRMM, Namrata Patel et Pierre Accorsi), juin-juillet, 2012 Encadrement d'un stagiaire de M1 (A. Stifani), septembre-décembre 2011.	2011-2012

*Dans ce tableau ne sont inscrits que les encadrements en rapport avec mes sujets de recherche. J'ai également encadré d'autres mémoires « généralistes », qui ne sont pas indiqués ici. *Cf.* le tableau 3.1 pour l'ensemble de mes encadrements de recherche.

Récapitulatif des publications sélectionnées du volet 3

Le volet 3 montre une forme d'équilibre entre publications à auteur unique et en co-auteurs. Pour la recherche autour de la **communication médiée par ordinateur** et le **discours électronique/numérique médié**, j'ai sélectionné **neuf** publications : 1 fiche terminologique (Panckhurst 1997a), 3 chapitres dans ouvrage : 1 chapitre dans (Bres et al. 1999), (Panckhurst 1999b), 1 chapitre dans l'ouvrage de Jacques Anis, (Anis 1999), (Panckhurst 1999a), 1 chapitre dans l'ouvrage de la psychologue Annie Piolat, (Piolat 2005), (Panckhurst 2006a) ; 5 actes : *Colloque international sur le document électronique, CIDE*, Rabat, (Panckhurst 1998c), *Colloque GRESICO, Communication, société et internet*, Vannes, (Panckhurst 1998a), *Simposio Internacional de Comunicación Social*, Santiago de Cuba, (Panckhurst et Bouguerra 2003), *Online Educa*, Berlin (Panckhurst 2003a), *ascilite*, Sydney, (Panckhurst 2006b). Pour la deuxième section, concernant les **SMS**, j'ai choisi **douze** publications : 3 chapitres dans ouvrage : ma typologie sur l'écriture SMS, (Panckhurst 2009), 2 chapitres dans l'ouvrage collectif coordonnée par (Cougnon et Fairon 2014) : (Accorsi et al. 2014 ; Panckhurst et Moïse 2014) ; 4 articles : 1 article sur le déroulement général du projet dans le journal *Epistémè* (Panckhurst et al. 2013), 1 article à propos de l'identification automatique de l'écriture SMS, dans *Travaux neuchâtelois de linguistique, TRANEL*, (Lopez et al. 2015), 1 article sur notre approche pluridisciplinaire, dans *Histoire des théories linguistiques*, (Panckhurst et al. 2016b), 1 article sur l'exclusion du transcodage et de l'annotation de notre corpus *88milSMS*, dans *Digital Scholarship in the Humanities*, (Panckhurst 2016a) ;

5 actes : aspects plurilingues des SMS, *International Conference on Meaning and Interaction, i-mean*, Bristol, (Panckhurst 2010), description du corpus de SMS, depuis la collecte jusqu'à l'analyse, *Congreso Internacional de Linguïstíca de Corpus, CILC*, Alicante, (Panckhurst 2013), alignement de SMS, *Language Resources and Evaluation Conference, LREC*, Reykjavik, (Lopez et al. 2014), affiche, *Digital Humanities conference, DH*, Sydney, (Panckhurst 2015), néographie, *Congreso Internacional de Neología en las Lenguas Románicas, CINEO*, Salamanque (Roche et al. 2016). Le corpus *88milSMS* se décline en **deux** versions, sur la plateforme *Huma-Num*, (Panckhurst et al. 2014a) et sur *Ortolang*, (Panckhurst et al. 2016a).
Enfin, j'ai retenu **cinq** publications, dont 3 **articles explicatifs**, (Panckhurst et al. 2014b,c, 2015a), et 2 conférences invitées, dont les **vidéos** sont à consulter en ligne (Panckhurst 2016b,c).
La **page presse** contenant plus de 35 articles et reportages est à consulter également : http://sud4science.org?q=fr/node/5

3.3.3.14 Sélection des publications : volet 3

Remarque. — Dans le Volume II, je scinde les publications en trois sections, pour le volet 3 : 1) Communication médiée par ordinateur (CMO), Discours électronique médié (DEM), Discours numérique médié (DNM) ; 2) Service de messages succincts, SMS ; 3) Corpus *88milSMS* ; 4) Articles explicatifs, vidéos, presse. Le lecteur pourra s'y reporter pour ce détail.

Accorsi, Pierre, Namrata Patel, Cédric Lopez, Rachel Panckhurst et Mathieu Roche (2014). « Seek and Hide : Anonymising a French SMS corpus using natural language processing techniques ». In : *SMS Communication. A Linguistic Approach*. Sous la dir. de Louise-Amélie Cougnon et Cédrick Fairon. Amsterdam/Philadelphia : John Benjamins, p. 11–28.

Lopez, Cédric, Reda Bestandji, Mathieu Roche et Rachel Panckhurst (2014). « Towards Electronic SMS Dictionary Construction : An Alignment-based Approach ». In : *Actes du colloque LREC*. Reykjavik, p. 2833–2838. url : http://www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf.

Lopez, Cédric, Mathieu Roche et Rachel Panckhurst (2015). « Classification des items inconnus de 88milSMS : aide à l'identification automatique de la créativité scripturale ». In : *Tranel (Travaux neuchâtelois de linguistique)* 63, p. 71–86. URL : https://www.unine.ch/files/live/sites/islc/files/Tranel/63/71-86_lopez_al_corr.pdf.

Panckhurst, Rachel (1997a). « La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? » In : *Terminologies nouvelles* 17, p. 56–58.

– (1998a). « Analyse linguistique du courrier électronique ». In : *Communication, société et internet, Actes du colloque Les relations entre individus médiatisées par les réseaux informatiques*. Sous la dir. de Nicola Guéguen et Laurence Tobin. GRESICO. Paris : L'Harmattan, p. 47–60.

– (1998c). « Marques typiques et ratages en communication médiée par ordinateur ». In : *Actes du colloque CIDE 98*. INPT, Rabat : Paris : Europia Productions, p. 31–43.

– (1999a). « Analyse linguistique assistée par ordinateur du courriel ». In : *Internet, communication et langue française*. Sous la dir. de Jacques Anis. Paris : Hermès, p. 55–70.

– (1999b). « La Communication médiée par ordinateur : un discours autre ? » In : *L'autre en discours*. Sous la dir. de Jacques Bres, Régine Delamotte-Legrand, Françoise Madray et Paul Siblot. Dyalang-Praxiling, Service des publications de l'Université Paul-Valéry Montpellier 3, p. 307–331.

– (2003a). « Computer-mediated communication and linguistic issues in French University online courses ». In : *Actes du colloque Online Educa*. Berlin, p. 454–457.

– (2006a). « Le discours électronique médié : bilan et perspectives ». In : *Lire, écrire, communiquer et apprendre avec Internet*. Sous la dir. d'Annie Piolat. Marseille : Éditions Solal, p. 345–366.

– (2006b). « Mediated electronic discourse and computational linguistic analysis : improving learning through choice of effective communication methods ». In : *Actes du colloque ascilite*. Sydney, p. 633–637. URL : http://www.ascilite.org/conferences/sydney06/proceeding/pdf_papers/p16.pdf.

Panckhurst, Rachel (2009). « Short Message Service (SMS) : typologie et problématiques futures ». In : *Polyphonies, pour Michelle Lanvin*. Sous la dir. de Teddy Arnavielle. Université Paul-Valéry Montpellier 3, p. 33–52.

– (2010). « Txtng in three European languages : does the linguistic typology differ ? » In : *Actes du colloque i-Mean*. University of the West of England, Bristol, p. 122–137. URL : http://www2.uwe.ac.uk/faculties/CAHE/Documents/Conferences/imean/IMEAN-Conference-Proceedings-2009.pdf.

– (2013). « A large SMS corpus in French : from design and collation to anonymisation, transcoding and analysis ». In : *Proceedings du colloque CILC*. Sous la dir. de Social Procedia et Elsevier Behavioural Sciences. Alicante. URL : http://www.sciencedirect.com/science/article/pii/S1877042813041475.

– (2016a). « A digital corpus resource of authentic anonymized French text messages : 88milSMS—What about transcoding and linguistic annotation ? » In : *Digital Scholarship in the Humanities*. DOI : 10.1093/llc/fqw049.

– (2016b). « De Sud4science à 88milSMS (un grand corpus de SMS authentiques) : entre linguistique et informatique ». In : *Conférence invitée*. ENS, Lyon. URL : http://cle.ens-lyon.fr/conf-aperos/de-sud4science-a-88milsms-un-grand-corpus-de-sms-authentiques-entre-linguistique-et-informatique--303624.kjsp.

– (2016c). « Les SMS en langue française, Conférence invitée ». In : Médiathèque André-Malraux, Béziers. URL : https://www.youtube.com/watch?v=jMH2a2v8Qdo.

Panckhurst, Rachel et Tayeb Bouguerra (2003). « Communicational and methodological/linguistic strategies using electronic mail in a French University ». In : *Actes du colloque 8th International Symposium on Social Communication*. Santiago de Cuba, p. 548–554.

Panckhurst, Rachel, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche et Bertrand Verine (2013). « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS ». In : *Épistème, revue internationale de sciences sociales appliquées* Des usages numériques aux pratiques scripturales électroniques.9, p. 107–138.

– (2014a). *88milSMS. A corpus of authentic text messages in French, produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère*

de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo, ISLRN : 024-713-187-947-8. URL : http://88milsms.huma-num.fr/.

– (2014b). « Un grand corpus de SMS en français : 88milSMS ». In : *La lettre de l'InSHS, la Tribune d'HumaNum*, p. 22–25. URL : http://www.cnrs.fr/inshs/Lettres-information-INSHS/lettre_infoinshs_31hd.pdf.
– (2014c). « Une grande collecte de SMS authentiques en français : démarche, remarques et conseils ». In : *Le français à l'université* 19.03. URL : http://www.bulletin.auf.org/index.php?id=1875.
– (2015a). « Dites-le dans le français que vous voulez ! » In : *Mediapart*. URL : https://blogs.mediapart.fr/edition/les-invites-de-mediapart/article/020415/dites-le-dans-le-francais-que-vous-voulez.
– (2016a). *88milSMS. A corpus of authentic text messages in French. Version nouvelle du corpus ISLRN 024-713-187-947-8, In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang : Nancy. cmr-88milsms-tei-v1*. URL : https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1.

Panckhurst, Rachel et Claudine Moïse (2014). « French text messages. From SMS data collection to preliminary analysis ». In : *SMS Communication. A Linguistic Approach*. Sous la dir. de Louise-Amélie Cougnon et Cédrick Fairon. Amsterdam/Philadelphia : John Benjamins, p. 141–168.

Panckhurst, Rachel, Mathieu Roche et Cédric Lopez (2015b). « Données authentiques : un grand corpus de SMS en français ». In : *Colloque SHESL, Corpus et constitution des savoirs linguistiques*, p. 33–35. URL : shesl-htl2015.sciencesconf.org/conference/shesl-htl2015/pages/Livret_resumes_SHESL_HTL2015.pdf.

Panckhurst, Rachel, Mathieu Roche, Cédric Lopez, Bertrand Verine, Catherine Détrie et Claudine Moïse (2016b). « De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire ». In : *Histoire des théories linguistiques (HEL)* 38.2, p. 73–86.

Roche, Mathieu, Bertrand Verine, Cédric Lopez et Rachel Panckhurst (2016). « La néographie dans un grand corpus de SMS français : 88milSMS ». In : *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones, Actes du colloque CINEO 2015, 22-24 octobre, Salamanque*. Sous la dir. de Joaquín García Palacios, Goedele De Sterck, Daniel Linder, Nava Maroto,

Miguel Sánchez Ibáñez et Jesús Torres del Rey. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt, Peter Lang, p. 279–302. URL : DOI:http://dx.doi.org/10.3726/978-3-631-69859-4.

3.4 Réseaux, diffusion et valorisation

Je prendrai un seul exemple ici : notre projet de collecte et d'analyse de SMS. Comme indiqué précédemment (§ 3.3.3.4), le recueil de SMS s'est déroulé en 2011 et le corpus final *88milSMS* a été mis à disposition pour téléchargement auprès du grand public en juin 2014 (http://88milsms.huma-num.fr/). Fin 2016, une deuxième version du corpus *88milSMS* — structurée en XML/TEI, et dotée d'une licence CC BY 4.0 plus ouverte que la précédente — a intégré la plateforme *Ortolang* (Panckhurst et al. 2016a), cmr-88milsms-tei-v1 ; https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1 (*cf.* p. 219).

3.4.1 Séminaires, conférences invitées et journées d'étude

Grâce à un financement MSH-M pour le projet *sud4science LR* (années 2011 et 2012), essentiellement pour de la recherche entrante (chercheurs se déplaçant à Montpellier), et ce dans la continuité du projet international *SMS4science*, j'ai pu consolider un réseau de chercheurs, accueillir des chercheurs à Montpellier et tisser de nouveaux liens scientifiques. J'ai organisé un ensemble de manifestations scientifiques, menés en présence et/ou par visio-conférence, de janvier 2011 à novembre 2012, ainsi que deux journées d'étude, en novembre 2011 (http://www.msh-m.fr/programmes-2011/sud4science-lr/).

Dans le tableau 3.22, je récapitule ces manifestations scientifiques (séminaires, conférences invitées et journées d'étude) pour les années 2011-2012.

Tableau 3.22 – Manifestations scientifiques : 2011-2012,
http://www.msh-m.fr/programmes-2011/sud4science-lr/

Manifestation scientifique n° 1, 8 fevrier 2011. *SMS4science, Alpes4science, sud4science LR* (Belgique et France).
Louise-Amelie Cougnon, CENTAL, Universite catholique de Louvain, (Belgique) : « sms4science : un projet international de collecte de sms pour la recherche scientifique ». Georges Antoniadis, Lidilem, Universite Stendhal, Grenoble 3 : « Alpes4science : Buts et preliminaires de la collecte ». Virginie Zampa, Lidilem, Universite Stendhal, Grenoble 3 & Gaelle Chabert, (étudiante), Lidilem, Universite Stendhal, Grenoble 3 : « Alpes4science : Collecte et traitement des SMS ». Rachel Panckhurst, Praxiling UMR 5267 CNRS — Universite Paul-Valery Montpellier 3 : « sud4science Languedoc-Roussillon : choix techniques et mise en place de la collecte ».
Manifestation scientifique n° 2, 15 mars 2011. SMS & psychologie (France métropole et La Réunion)

Antonine Goumi [137], Josie Bernicot, (Psychologie, Universite de Poitiers-CNRS UMR 6234) : « Un corpus de SMS produits par de jeunes adolescents : methode de recueil et premieres donnees » (Goumi et Bernicot 2011).
Gudrun Ledegen [138], Laboratoire LCF-UMR 8143 du CNRS, Universite de La Reunion : « SMS4science a La Reunion : langues en contact ». Cette intervention sera en visio-conference.

Manifestation scientifique n° 3, 21 avril 2011. Contacts de langues, surdite, evolution des pratiques scripturales des SMS (France)

Marion Blondel, SFL, CNRS-Paris 8 ; Jeanne Gonac'h, Ingenieur d'etudes contractuel (LCPE, Paris 6) et Chargee de cours a l'Universite de Rouen (LIDIFRA EA4503) : « Existe-t-il des sms "pi-sourd" ? ».
Fabien Lienard, Departement Information-Communication - IUT Le Havre, (LiDiFra - Universite de Rouen) : « Analyse diachronique de la pratique scripturale SMS : enjeux et evolutions ».

Manifestation scientifique n° 4, 16 mai 2011. SMS4science.ch (Suisse)

Elisabeth Stark, (Linguistique romane), Romanisches Seminar, Universite de Zurich ; Christa Durscheid, (Linguistique allemande), Deutsches Seminar, Universite de Zurich : « Le nouveau corpus SMS de suisse : traitement des donnees mulitlingues et premiers resultats d'analyse ».
Simone Ueberwasser, coordinatrice du projet *sms4science.ch* a l'universite de Zurich : « Annotation du corpus sms4science.ch : problemes et procedures envisagees. ».

Manifestation scientifique n° 5, 29 septembre 2011. SMS et TALN (France et Belgique)

Mathieu Roche [139], (informatique), Universite de Montpellier, LIRMM, « Traitement automatique des messages courts par des approches de Fouille de Textes »
Thomas Francois, (TAL) Universite catholique de Louvain : « Une approche statistique des corpus de SMS : outils et defis ».

Manifestation scientifique n° 6, 20 octobre 2011. SMS : genre, orthographe, plurilinguisme (France, Norvège)

Olga Volckaert-Legrier, (Psychologie), Laboratoire PDPS, E.A. 1687, Universite de Toulouse II - Le Mirail : « Usage des SMS chez les adolescents : une etude comparative de filles et de garcons de 13 a 18 ans ».
Celine Combes [140], Doctorante (Psychologie), Universite de Toulouse II - Le Mirail : « La production orthographique dans les SMS de collegiens : etude de l'effet d'une double tache ».
Kristin Vold Lexander, Institut des etudes en culture et langues orientales, Universite d'Oslo : « SMS plurilingues des etudiants de Dakar – une approche integree » [141].

Manifestation scientifique n° 7 : journées d'étude internationales : 14 et 15 novembre, 2011. Harmonisation/standardisation des méthodes de traitement de corpus écrits de type SMS. Anonymisation, transcodage, annotation. [142].

Manifestation scientifique n° 8 : 29 mai 2012. Art, culture, sciences.

137. Antonine Goumi est désormais MCF à l'université Paris-Ouest Nanterre.
138. Gudrun Ledegen est désormais PU à l'université Rennes 2.
139. Mathieu Roche est désormais chercheur HDR au Cirad, Montpellier.
140. Céline Combes est désormais MCF à l'université d'Angers.
141. On consultera également sa thèse de doctorat (Vold Lexander 2010), et le mémoire de Master (Moussa 2014) sous ma direction pour le contexte camerounais.
142. http://www.msh-m.fr/la-recherche/programmes-2012/sud4science-lr/manifestations-scientifiques-460/article/journees-des-14-et-15-novembre

Franck Bauchard [143], Directeur de La Panacee, Centre d'art & culture contemporaine, Montpellier ; Eli Commins, Auteur et metteur en scene, Paris. « Nouvelles machines d'ecriture », et « Le Textopoly et la place de l'ecrit dans la manifestation en preparation sur l'art et le telephone ». Table ronde : « Art, culture et sciences ». Intervenants : Catherine Detrie, Rachel Panckhurst, Bertrand Verine, Claudine Moise, Cedric Lopez.
Manifestation scientifique n° 9, 4 septembre 2012. Anonymisation, informatisation, droit.
« Seek & Hide. L'anonymisation d'un corpus de SMS » Namrata Patel [144] et Pierre Accorsi, etudiants en Master 1 Informatique, specialite : Donnees, Connaissances et Langage Naturel (DECOL) a l'Universite de Montpellier ; stagiaires au LIRMM dans le cadre de *sud4science LR*. Table ronde « Anonymisation, informatisation, droit ». Intervenants : Catherine Detrie, Rachel Panckhurst, Bertrand Verine, Claudine Moise, Mathieu Roche, Cedric Lopez, Stephanie Delaunay Nicolas Hvoinsky, (Direction des affaires juridiques et institutionnelles, Universite Paul-Valery Montpellier 3).
Manifestation scientifique n°10, 26 novembre 2012. Anonymisation, linguistique-informatique.
Louise-Amelie Cougnon, CENTAL, UCL, Louvain, (Belgique) : « Variation individuelle de l'ecrit sms. Lexique, syntaxe et orthographe ». Michel Otell et Camille Lagarde-Bellville [145], etudiants en Master 2 Sciences du Langage a l'Universite Paul-Valéry Montpellier 3 ; stagiaires dans le cadre de sud4science LR : « Anonymisation et linguistique ». Table ronde « Anonymisation, linguistique-informatique ». Intervenants : Catherine Detrie, Bertrand Verine (Praxiling UMR 5267 CNRS — Universite Paul-Valery Montpellier 3), Claudine Moise (Lidilem, Universite Stendhal Grenoble 3), Mathieu Roche (Lirmm, université de Montpellier), Cedric Lopez (Viseo, Grenoble).

Certains séminaires ont été filmés et les vidéos ont été mises à disposition (MSH-M.TV : http://msh-m.tv/spip.php?rubrique138 ; itunes : https://itunes.apple.com/fr/itunes-u/sud4science-languedoc-roussillon/id594173977?mt=10).

3.4.2 Réseaux de chercheurs

Notre réseau « interne » était déjà constitué entre linguistes et informaticiens : Catherine Détrie, Bertrand Verine et moi-même (Praxiling, université Paul-Valéry Montpellier 3), Cédric Lopez (Viseo, Grenoble), Claudine Moïse (Lidilem, université Grenoble-Alpes), Mathieu Roche (Tétis, Cirad, CNRS, LIRMM). Il s'agissait de l'élargir sur le triple plan national, européen, international. Pendant tout le déroulement des projets *sud4science LR* et DGLFLF, j'ai maintenu des contacts avec le réseau des chercheurs mis en œuvre lors des séminaires

143. Franck Bauchard est désormais Director, Arts Management Program, University at Buffalo, New York State.

144. Namrata Patel est désormais docteure en informatique et ingénieure à Viseo, Grenoble.

145. Michel Otell et Camille Lagarde-Belleville sont actuellement doctorants à l'Universite Paul-Valéry Montpellier 3.

scientifiques qui se sont tenus dans le cadre de nos projets. Le réseau « externe » était constitué des chercheurs suivants : Cédrick Fairon, Louise-Amélie Cougnon, Hubert Naets, Thomas François, Richard Beaufort (CENTAL, université catholique de Louvain, Belgique), Elisabeth Stark, Christa Dürscheid, Simone Ueberwasser (université de Zurich, Suisse), Georges Antoniadis, Virginie Zampa, Gaëlle Chabbert (Lidilem, université Grenoble-Alpes), Josie Bernicot, (CLIF, université de Poitiers), Olga Volckaert-Legrier (CLLE, université de Toulouse Jean-Jaurès) Gudrun Ledegen (PREFics, université de Rennes), Fabien Liénard (CIRTAI-IDEES, I.U.T du Havre), Marion Blondel (LCA, CNRS-Paris 8), Tita Kyriacopoulou (LIGM-MOA, université Paris-Est Marne-la-Vallée), Kristin Vold Lexander (Oslo), Patrick Drouin (université de Montréal), Christian Guilbault (Simon Fraser University, Vancouver), Franck Bauchard et Eli Commins (La Panacée, Montpellier, Paris, University at Buffalo, États-Unis), Alena Podhorná-Polická (Masaryk University, Brno, Tchéquie). Ces chercheurs représentent les disciplines suivantes : Sciences du Langage, Informatique, Linguistique-Informatique, Information-Communication, Psychologie, Arts et culture.

Chercheur invité

La MSH-M a accueilli Cédrick Fairon (Directeur CENTAL, UCL, Belgique) du 12 au 19 novembre 2011 en tant que chercheur invité. Pendant son séjour, il est intervenu (entre autres) dans un séminaire du projet *sud4science LR*, intitulé : « Correction/normalisation de SMS : les apports du corpus SMS4science » que j'ai organisé le 19 novembre 2011.

3.4.3 Diffusion

3.4.3.1 Mise à disposition du corpus *88milSMS*

Tout en saluant notre initiative, maints collègues ont été surpris d'apprendre que nous souhaitions mettre à disposition du grand public, notre corpus *88milSMS*. En 2014, cette démarche était encore assez rare et l'est encore en 2017 [146].

146. *Cf.* également (Chanier et al. 2014) pour CoMeRe, ou, par exemple, pour des mises à disposition d'un grand ensemble de données liées aux *lexiques-grammaires*, http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html

Étant donné le suivi juridique minutieux par notre juriste-CIL, Nicolas Hvoinsky, depuis le début du projet *sud4science LR*, nous n'avions pas d'inquiétude quant à la mise à disposition du contenu :

> L'aspect juridique est primordial. Une collecte de données qui passe outre la règlementation en vigueur risque de produire des données inexploitables pour des raisons légales. Rappelons que les SMS constituent des données personnelles, sensibles. La vie privée doit donc être protégée. Nous avons choisi d'associer, dès le départ de notre projet, [la direction] des affaires juridiques et institutionnelles de l'université Paul-Valéry Montpellier 3 [DAJI] par l'entremise de sa directrice, la juriste Stéphanie Delaunay, et de son juriste correspondant informatique et libertés, CIL, Nicolas Hvoinsky. **Parfois, des universitaires recueillent des données, des années durant, pour se rendre compte trop tard que celles-ci sont inutilisables, car la collecte n'a pas respecté les normes juridiques.**
>
> (Panckhurst et al. 2014c) Je souligne.

Dès l'annonce de la diffusion du corpus, en juin 2014, l'enthousiasme était au rendez-vous. Entre juin 2014 et décembre 2016, le corpus *88milSMS* a été téléchargé à 447 reprises, ce que j'estime plus qu'honorable. Les cinq continents sont représentés pour les téléchargements sur la carte mondiale (*cf.* figure 3.41).

Quarante et un pays figurent dans la liste des téléchargements, avec le pourcentage le plus important en France (60 %), et plus généralement dans 15 pays parmi les 28 états membres de l'Europe (77 % incluant la France) (*cf.* figure 3.42). À partir de septembre 2016, une nouvelle version du corpus a également été mise à disposition sur la plateforme *Ortolang*. J'ai donc pensé que les statistiques seraient moins révélatrices quelques mois après cette date, car la nouvelle licence CC BY 4.0 est plus ouverte et n'exige pas qu'un formulaire soit rempli par la personne effectuant le téléchargement.

3.4.3.2 Communication et médias

Dès notre premier communiqué de presse annonçant le projet de récolte de SMS le 11 septembre 2011 (*cf.* figure 3.43) et la visibilité sur les réseaux sociaux (https://www.facebook.com/pages/sud4science/160054617409199), un véritable engouement médiatique s'est emparé du projet.

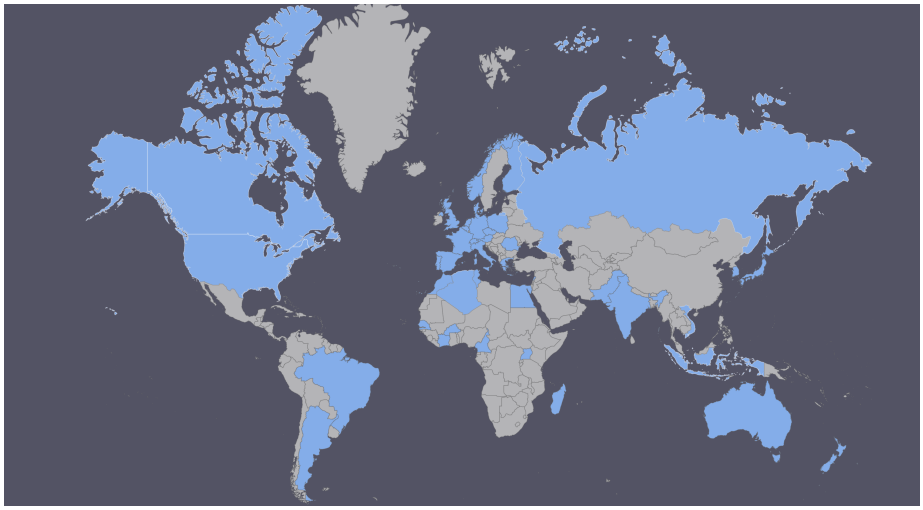


Figure 3.41 – Visualisation (en bleu) des 41 pays à partir desquels des téléchargements du corpus *88milSMS* ont été effectués au niveau mondial entre juin 2014 et décembre 2016, https://www.amcharts.com

.

La consultation de notre page presse récapitule ces événements : http://sud4science.org/?q=fr/node/5).

En effet, la diffusion de l'événement concernant la collecte s'est mutée en véritable valorisation du projet.

3.4.4 Valorisation

Nous sommes intervenus dans les médias locaux, nationaux et internationaux : presse écrite/en ligne, radio et télévision.

3.4.4.1 Presse écrite, en ligne

En tant que responsable de projet, j'ai été interviewée par des journalistes français, belges, suisses, québécois, argentine (journal espagnol, *ABC*), pour 33 articles de presse écrite/en ligne, incluant les journaux et revues/magazines suivants, entre septembre 2011 et août 2016. Je les indique ci-dessous par ordre chro-

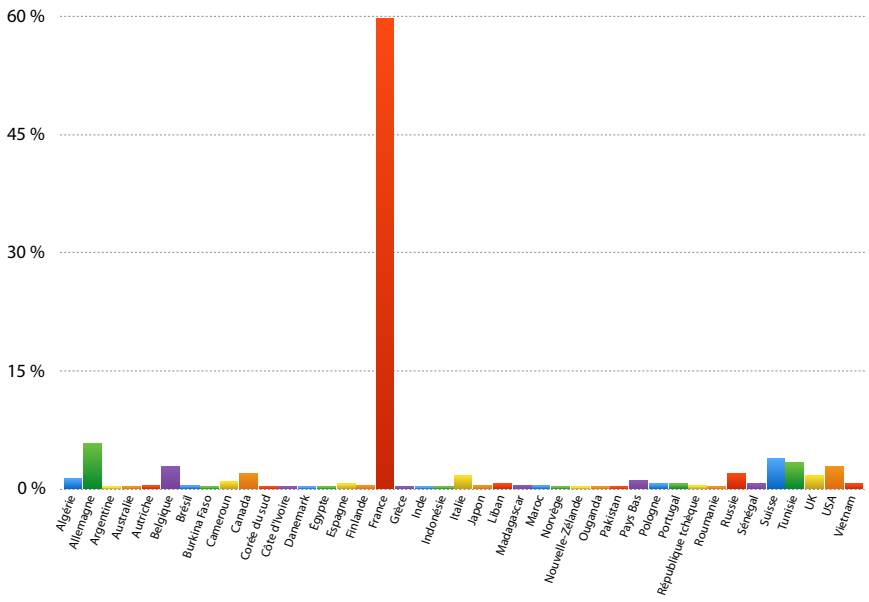


FIGURE 3.42 – Pourcentage de téléchargement du corpus *88milSMS* par pays, pour un total de 40 pays distincts, entre juin 2014 et décembre 2016.

nologique inversé : *Télérama* (août 2016), *Brain Magazine* (juin 2016), *Femina Suisse* (mars 2016), *Apprentis d'Auteuil* (octobre-novembre 2015), *La Libre*, Belgique (août 2015), *Le Figaro* (août 2015), *Science et Vie* (juillet 2015), *Le Parisien Magazine* (juin 2015), *Le Nouvel Observateur* (mai 2015), *Le Soir,* Belgique (mai 2015), *Madame Figaro* (avril 2015), *ABC*, Espagne (couverture et article portrait du supplément culturel, avril 2015), *20 Minutes* (2 articles, février 2015), *Le Nouvel Observateur* (décembre 2014), *Glamour* (décembre 2014), *Stylist* [147] (octobre 2014), *La Presse+*, Canada (mai 2014), *Journal du CNRS*, en ligne (mars 2014), *Journal du CNRS* [148] (Hiver 2014), *M le magazine du Monde* (novembre 2013), *Version Femina* (octobre 2013), *Le Monde* (décembre 2012), *La Libre*, Belgique (juillet 2012), *Slate* (juillet 2012), *Midi Libre* (décembre 2011), *La Gazette*, Montpellier (octobre 2011), *Languedoc Pages*, en anglais (octobre 2011), *Hérault du*

147. Claudine Moïse a également été interviewée pour cet article.
148. Mathieu Roche a également été interviewé pour les deux articles du CNRS.

COMMUNIQUÉ DE PRESSE

2011
2012

FAITES DON DE VOS SMS À LA SCIENCE !
- DU 15 SEPTEMBRE AU 15 DECEMBRE 2011 -
http://www.sud4science.org

Des chercheurs des universités de Montpellier, en partenariat avec la Maison des Sciences de l'Homme de Montpellier (http://www.msh-m.fr/), le CNRS et des laboratoires de recherche, lancent un vaste projet autour des textos (SMS).

La responsable de ce projet, Rachel Panckhurst est enseignant-chercheur en Sciences du Langage (linguistique-informatique) et membre de l'équipe de recherche Praxiling (http://praxiling.univ-montp3.fr) UMR 5267 CNRS Université Paul-Valéry Montpellier 3 (http://www.univ-montp3.fr). Elle est responsable du programme scientifique « sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée » de la MSH-M.

Les SMS sont aujourd'hui d'une utilisation relativement commune et généralisée. Messages utilitaires, amicaux ou amoureux, ils facilitent et agrémentent notre quotidien ; signes de relations sociales, ils sont ainsi un lien qui, le plus souvent, nous « bipe » et nous attache avec bienveillance aux autres.
Ces écrits, de longueur variable mais au nombre de caractères limités, sont rédigés dans une orthographe plus ou moins standard ou cryptée, mais souvent avec créativité. Ils peuvent jouer aussi de diverses langues ou variétés de langues. Avec les téléphones nouvelle génération, ils se présentent sous forme de conversations continues ou interrompues. Entre usage informatif ou ludique, les SMS sont donc d'un grand intérêt pour les linguistes, les informaticiens, les psychologues, les spécialistes de la communication, etc.

Face à ce phénomène social est né un projet de recherche international, *sms4science,* (www.sms4science.org) initié en 2004 par le laboratoire CENTAL de l'université Catholique de Louvain (UCL) en Belgique et auquel ont déjà participé les universités de la Réunion, de Montréal, de Zürich et de Grenoble. La collecte à grande échelle de plusieurs dizaines de milliers de SMS permet d'analyser des SMS authentiques et de comprendre leurs différentes écritures, selon les téléphones utilisés, selon les contextes, monolingues ou plurilingues, selon les générations…

À l'automne 2011 (**du 15 septembre au 15 décembre),** Rachel Panckhurst et ses collègues (des chercheurs des universités de Montpellier en partenariat avec la MSH-M, le CNRS et des laboratoires de recherche) organisent une collecte de SMS auprès du grand public. L'objectif est de recueillir **50 000 SMS** authentiques en langue française émis par la population vivant en Languedoc-Roussillon. Pour réaliser cette collecte, les participants sont invités à s'inscrire sur le site internet www.sud4science.org et à 1) nous mettre en copie des SMS qu'ils envoient (et dont ils sont l'auteur) ou 2) nous faire suivre les SMS qu'ils ont déjà envoyés et qui figurent dans la mémoire de leur téléphone. **Des cadeaux sont tirés au sort toutes les semaines et un ipad 2 est à gagner pour celle ou celui qui aura envoyé le plus grand nombre de SMS !**

Les informations recueillies seront anonymisées et rassemblées dans une base de données. Le recueil ainsi constitué fera l'objet d'une diffusion dans le monde scientifique, à l'usage des chercheurs et des étudiants, voire du grand public (sous forme de CD/DVD et/ou de livre).

>>>>> Lancement de la collecte sud4science **le jeudi 15 septembre à 18h à iTribu Mauguio-Fréjorgues.**

> Contact Presse - Université Paul-Valéry / Montpellier III : communication@univ-montp3.fr
- Tél. : 04 67 14 22 67 / 22 74
> Responsable projet sud4science LR : Rachel Panckhurst, MCF en linguistique-informatique, Praxiling, UMR 5267 CNRS Université Montpellier 3, sud4science@msh-m.org

> **Toute l'actualité de l'Université sur http://www.univ-montp3.fr**

Figure 3.43 – Communiqué de presse : récolte de SMS

jour (septembre 2011), *Direct Montpellier* (septembre 2011), *Objectif LR* (septembre 2011), *Tout Montpellier* (septembre 2011), *Midi Libre* (septembre 2011). (http://sud4science.org/?q=fr/node/5)

Curieusement, la sollicitation pour des entretiens ne s'est pas interrompue après la fin de la collecte des SMS (décembre 2011). Après 8 articles de septembre à décembre 2011, 3 en 2012, 2 en 2013, cela est reparti à la hausse en 2014 (6 articles) et en 2015 (11 articles), avant de redescendre à 3 articles en 2016. Je reviendrai sur ce pic d'articles en 2015 plus bas (§ 3.4.5.2).

Par ailleurs, mes collègues et moi avons décidé d'accepter toutes les demandes d'entretien [149] : du magazine grand public, à la revue scientifique spécialisée, en passant par le journal local, national, voire international. Au début, je demandais systématiquement à relire (avant publication) l'article écrit par le/la journaliste afin d'en vérifier le contenu et éventuellement de demander des rectificatifs. En tant qu'universitaire, relectures et corrections font évidemment partie de notre quotidien de publiant(e). Cela étant, une fois, suite à la réception de mes nombreuses corrections, le jeune journaliste qui m'avait interviewée s'est vexé et a refusé de nous publier. Par la suite, je n'ai pas toujours demandé la relecture, estimant qu'il était peut-être plus important qu'on évoque notre projet, quitte à accepter (à contre-cœur) que des raccourcis ou des erreurs se glissent dans le contenu. Au bout de 5 ans d'entretiens à raison de quelques articles à une dizaine par an, j'ai commencé à prendre l'habitude, arrivant assez rapidement à discerner, au téléphone, le journaliste quasi-débutant, ou stagiaire, du journaliste chevronné.

Enfin, cela est parfois frustrant d'être interviewée pendant plus d'une heure, pour qu'au final, seulement une ou deux phrases soient retenues — avec, parfois, l'insertion de guillemets, me citant, alors que cela est loin de correspondre à la réalité de ce que j'avais prononcé. Mais, à la décharge des journalistes, ceux-ci sont parfois eux-mêmes frustrés de ne pas pouvoir imposer leurs articles tels quels à la rédaction de leur journal, qui effectue également à son tour des coupes éditoriales.

149. Nous en avons néanmoins manqué quelques-unes, car les demandes d'entretiens journalistiques s'accordent en général le jour même voire le lendemain. Il faut être extrêmement réactif.

3.4.4.2 Télévision

Au niveau télévisuel, notre premier reportage a été diffusé sur *TF1*. Catherine Détrie, Bertrand Verine, un stagiaire étudiant en Master, Anthony Stifani, et moi-même avons été interviewés à propos de notre collecte de SMS et le reportage a été diffusé lors du JT de 20h présenté par Laurence Ferrari en octobre 2011 (figure 3.44).

Figure 3.44 – Reportage JT, 20h, *TF1*, 4/10/2011

En fait, la télévision régionale (*France 3*) nous avait interviewés avant cette date, mais la diffusion du reportage (présentée par Anne-Sophie Mandrou au JT 19/20 Édition Languedoc-Roussillon), a été reportée à une date postérieure en octobre 2011 (figure 3.45).

Deux ans plus tard, en novembre 2013, j'ai été interviewée pour le *Campus Mag LR*. Cela a relevé d'un véritable défi pour moi : gérer le stress devant une caméra *live* [150], expliquer l'essentiel du projet via les questions impromptues de l'animateur, en sachant que nous ne disposions que de 4 minutes 30 maximum (*cf.* l'entretien avec Damien Conaré, *Campus Mag LR*, 19/11/2013, http:

150. Même si la diffusion du reportage n'a pas été immédiate, il n'y avait qu'une séance de tournage, sans possibilité de la refaire — les animateurs voulaient prendre sur le vif pour que cela soit le plus naturel possible. La seule édition postérieure effectuée concernait l'incrustation d'images. Aucune autre coupe n'a été effectuée.

FIGURE 3.45 – Reportage JT, 19/20h, *France 3*, 19/10/2011

//www.youtube.com/watch?v=Ig_1C1RVAg8). Le dernier reportage en date sur une chaîne de télévision publique a été effectué en juin 2015. Le tournage a duré plus de deux heures, pour un résultat extrait de 25 secondes, incrusté dans un reportage présenté par Laurent Delahousse au JT-Weekend de 20h, mi-juin 2015 (figure 3.46).

FIGURE 3.46 – Reportage JT, 20h, *France 2*, 15/06/2015

Cela semble habituel à la télévision, mais la frustration est de mise lorsqu'on a l'impression qu'une ou deux phrases plutôt insignifiantes sont retenues. Néanmoins, d'un point de vue de la valorisation, il est indéniable que des passages au JT de 20h touchent le grand public et ont un impact positif sur nos projets de recherche, quoique sans répercussions sur les subventions accordées ![151].

3.4.4.3 Radio

La radio, comme la télévision et la presse écrite/en ligne, a été présente dès le départ. *Radio France Bleu Hérault* et *Sud Radio* m'ont sollicitée pour des entretiens menés soit par téléphone, soit de visu, le 20 septembre 2011. En 2014, Daniel Fiévet m'a interviewée pour Info Sciences, France Info : « Les SMS sous la loupe des scientifiques » (http://www.francetvinfo.fr/replay-radio/info-sciences/les-sms-sous-la-loupe-des-scientifiques_1755977.html).

Les apparitions médiatiques plus « stressantes », mais également passionnantes, furent lors de deux invitations par des journalistes de *France Inter*.

En janvier 2014, Claudine Moïse, Bertrand Verine et moi-même avons participé en direct à l'émission « Le Grand Bain », animé par Sonia Devillers (« Le SMS réinvente-t-il le français ? », https://www.franceinter.fr/emissions/le-grand-bain/le-grand-bain-25-janvier-2014).

Puis, en mars 2016, Mathieu Roche et moi-même sommes intervenus en direct avec un troisième invité, le chercheur Bertrand Bergier, dans l'émission « La Tête au Carré », animé par Mathieu Vidard (« La science du texto », https://www.franceinter.fr/emissions/la-tete-au-carre/la-tete-au-carre-29-mars-2016 [152]).

3.4.4.4 Articles explicatifs

Outre les articles de presse/en ligne écrits par des journalistes, nous avons nous-mêmes souhaité écrire des articles pour le grand public ou pour des chercheurs émanant d'autres disciplines. Je n'aime pas le terme « vulgarisation », préférant privilégier celui d'« explication ». La liste de ces articles, écrits en co-auteurs avec mes collègues du projet, figure ci-dessous, par ordre chronologique inverse :

151. Actualité oblige, les reportages ne sont plus consultables sur les sites des chaînes, mais je les ai archivés.

152. Ces trois URL ont été consultés le 12 janvier 2017.

FIGURE 3.47 – Mathieu Roche, Rachel Panckhurst, Mathieu Vidard, Bertrand Bergier — *La Tête au Carré, France Inter*, 29 mars 2016.

1. Panckhurst R., Détrie C., Lopez C., Moïse C. Roche M., Verine B., (2015), « Dites-le dans le français que vous voulez ! », *Les invités de Mediapart*, 2 avril 2015, https://blogs.mediapart.fr/edition/les-invites-de-mediapart/article/020415/dites-le-dans-le-francais-que-vous-voulez
2. Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014), « Une grande collecte de SMS authentiques en français : démarche, remarques et conseils. », *Le français à l'université*, 19-03 | 2014, mise en ligne le 25 septembre 2014 : http://www.bulletin.auf.org/index.php?id=1875
3. Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014), « Un grand corpus de SMS en français : 88milSMS », *La lettre de l'InSHS*, pages 22-25, la Tribune d'Huma-Num, septembre 2014, http://www.cnrs.

fr/inshs/Lettres-information-INSHS/lettre_infoinshs_31hd.pdf.

3.4.4.5 Conférences et débats invités, participation à film

Par ailleurs, j'attache une grande importance aux rencontres avec le public. Les conférences et débats invités l'attestent, ainsi que ma participation à un film autour d'une exposition sur le téléphone mobile à *La Panacée* (Montpellier, www.lapanacee.org) :

1. Panckhurst R. (2016, mars), « Les SMS en langue française », Conférence invitée, Médiathèque André-Malraux (MAM), Béziers, 26 mars 2016, https://www.youtube.com/watch?v=jMH2a2v8Qdo
2. Panckhurst, R. (2013, novembre). Rencontre et débat avec le public après le visionnage à la Panacée du film « The World », de Jia Zhang Ke. Animation par Caroline Barraud, association *Brand à part*, le 28 novembre 2013.
3. Panckhurst R. (2013, juillet). Participation sur invitation au projet « La Cabine », [séquence filmée de 3 minutes à la Panacée], Montpellier, par Christine Bouteiller (Réalisation), David Olivari (conception multimédia), Production : la Panacée / les films du Tamarin.
4. Panckhurst R. (2012, avril), « sud4science Languedoc-Roussillon : collecte et analyse de 90 000 SMS authentiques », conférence invitée, La Fabrique-Mons (Belgique), 27 avril 2012.
5. Panckhurst, R. (2011, décembre). Participation sur invitation à la soirée *Convergences*. Rencontre publique avec Isabelle Massu, organisée à l'espace « Kawenga — Territoires numériques » le 08 décembre 2011, Montpellier, France.
6. Panckhurst, R. (2011, septembre). Participation sur invitation à la 4e édition de *la Nuit des Chercheurs*, organisée le 23 septembre 2011 au Domaine d'Ô par l'association *ConnaiSciences*, Montpellier, France.

3.4.5 Apports mutuels

J'aimerais croire que nos interventions, nos apparitions médiatiques, etc. ne soient pas une simple opération de diffusion ou de valorisation de nos projets

— bien que ces aspects soient également importants. Je suis profondément convaincue que la recherche publique doit être « rendue », « présentée », « expliquée », au public. Nous ne devons pas garder nos corpus enfermés sur nos serveurs, cachés derrière des pare-feux. Ils doivent être mis à disposition sur des plateformes ouvertes, en téléchargement libre pour tous. Nos discours doivent être plus facilement compréhensibles par le grand public. Puis, notre recherche appliquée doit servir dans les situations de tous les jours. Vaste programme, sans cesse renouvelé.

3.4.5.1 Vers autrui

J'ai l'espoir (l'illusion ?) que nous, chercheurs, pouvons avoir une influence sur autrui. Parfois, en discutant avec des journalistes, par exemple, je me suis rendue compte que mon, nos point(s) de vue a/ont pu modifier leurs positionnements. J'évoquerai deux exemples :

1. Dans un article apparaissant dans *Slate* en 2012, le journaliste a bien voulu inclure une diapositive que j'avais préparée pour aller à l'encontre de quelques idées reçues (figure 3.48) — avec lesquelles il n'était pas nécessairement en accord au début de notre entretien. Il a même titré l'article : « Les mobiles modernes ne vont pas tuer le langage SMS ».
2. Pour préparer notre intervention sur *France Inter,* début 2014, dans l'émission *Le Grand Bain*, Sonia Devillers — dans un grand souci de professionnalisme, que j'ai vraiment apprécié, car j'ai pu constater que tous les journalistes ne pratiquent pas systématiquement cela — a téléphoné à chaque intervenant (Claudine Moïse, Bertrand Verine et moi-même), dans la semaine qui a précédé notre direct. Avant nos entretiens téléphoniques — préparatoires au direct — l'émission devait s'intituler : « Le SMS a-t-il tué le français ? ». Après ceux-ci, le titre a été modifié en « Le SMS réinvente-t-il le français ? ».

En tant que chercheurs, ces types de « petites victoires » [153] nous apportent de la

153. J'avais apprécié également le titre de l'article paru dans *La Gazette* (13/10/11) « Non, les SMS ne nuisent pas à l'orthographe », suite à l'entretien que j'ai accordé de visu dans les locaux du journal à Montpellier, au journaliste Olivier Rioux.

Figure 3.48 – Contre-carrer les idées reçues sur les scripteurs de SMS.

satisfaction, nous confortent dans notre travail et nous convainquent que nos efforts servent à autrui.

À l'inverse, il arrive que nous nous énervions. Parfois, nous avons l'impression que le message ne passe pas du tout !

L'article paru dans Mediapart en 2015 [154] constitue une réaction de notre part à une campagne médiatique, produite à l'occasion de la première journée de la langue française dans les médias audiovisuels, « Dites-le en français », organisée en partenariat avec la *Délégation générale à la langue française et aux langues de France* (DGLFLF) et *Radio France*. Le conseil supérieur de l'audiovisuel (CSA) avait réalisé une campagne de trois vidéos. De notre point de vue, celles-ci dévalorisaient l'écriture SMS et nous avons protesté fortement contre leur médiatisation :

> Nous, chercheurs du projet *sud4science/88milSMS*, linguistes et informaticiens (Rachel Panckhurst, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu

154. (Panckhurst et al. 2015a) « Dites-le dans le français que vous voulez ! », *Les invités de Mediapart*, 2 avril 2015, https://blogs.mediapart.fr/edition/les-invites-de-mediapart/article/020415/dites-le-dans-le-francais-que-vous-voulez (consulté le 12 janvier 2017).

Roche, Bertrand Verine), appartenant à des structures de recherche publiques [...] et privées [...], nous insurgeons contre le contenu de ces vidéos. De notre point de vue, elles dévalorisent l'une des créativités majeures de la langue française écrite du 21e siècle : l'écriture SMS.
Dans l'une de ces vidéos, intitulée « Stop au langage SMS - Campagne CSA "Dites-le en français" », on voit un adolescent frigorifié entrer dans une voiture : il avait envoyé un SMS à sa mère pour lui demander de venir le chercher car son cours était annulé. On voit la mère lui rétorquer : « J'ai rien compris », puis, une fois que l'adolescent lui a expliqué le contenu du SMS qu'elle n'avait pas compris, elle poursuit :« La prochaine fois que t'as besoin de moi, tu m'achètes un dictionnaire pour SMS, oh, non, mieux, t'apprends à écrire ». Puis on entend, en arrière-fond, « Dites-le en français. Notre langue est belle. Utilisons-la. »
En tant qu'enseignants-chercheurs et chercheurs en Sciences du Langage et en Traitement automatique des langues (TAL), notre travail consiste, entre autres, à étudier l'évolution de la langue française. Les linguistes ne jugent pas ; ils observent, constatent, analysent et rendent les résultats de leurs travaux de recherche disponibles pour consultation par le grand public. [...]
Nous avons bénéficié de deux subventions publiques pour mener à bien ce travail : MSH-M (Maison des Sciences de l'Homme de Montpellier) et DGLFLF.
Dans le rapport de 50 pages rendu à la DGLFLF à l'issue du projet intitulé « Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS », nous avons expliqué pourquoi nous nous intéressons aux SMS : en tant que linguistes, nous observons, sans jugement, sans nous référer à une norme quelconque, les changements linguistiques. Nous nous intéressons à l'étude du langage, des langues et des pratiques langagières, donc, lorsqu'il y a des mutations éventuelles, nous saisissons l'occasion pour étudier de nouveaux phénomènes. La langue est dynamique, en mouvance constante et surtout vivante (ce que semble méconnaître la vidéo consacrée aux SMS !). Puis, en étudiant les SMS, nous pouvons envisager des applications industrielles en TAL : des systèmes de transcodage SMS vers le français standardisé ; des logiciels de reconnaissance des SMS, ou de vocalisation, pour des personnes aveugles, ou pour des conducteurs, etc. Ces applications pourront aussi aider à traiter de grandes masses de données textuelles très présentes aujourd'hui dans les réseaux sociaux (tweets, Facebook).
Grâce à des subventions publiques, nous avons pu observer et analyser une écriture SMS d'une très grande créativité dans ses formes, et qui renferme une

> dimension ludo-affective très forte.
> Nous déplorons que l'écriture SMS soit trop souvent considérée comme une facette de la langue à bannir. Acceptons enfin (!) que cette écriture fasse partie intégrante de la langue française dans sa dimension évolutive. Nous savons qu'à l'oral plusieurs registres, plusieurs genres sont possibles, et même nécessaires, selon les contextes. Refusons, qu'à l'écrit, seule la langue française normée soit acceptable. Toutes les écritures, toutes les créativités doivent être à l'honneur. Les scripteurs de SMS, eux-mêmes, font la différence entre une écriture utilisée entre amis, proches, dans un contexte où l'affect et la volonté de faire partager cet affect priment, et une autre situation exigeant d'autres pratiques. Nos études scientifiques le montrent. Alors, oui au slogan « Dites-le en français », qui, pour nous, ne doit pas équivaloir à : « Dites-le en français normé » mais plutôt à « Dites-le dans le français que vous voulez ! » (Panckhurst et al. 2015a), je souligne.

J'étais ravie que *Mediapart* nous publie, même si notre protestation écrite n'ait peut-être pas atteint les sphères que nous ciblions. Nous continuerons la lutte.

Orthographe

Les travaux sur les SMS (et également sur les discours numériques apparaissant au sein des réseaux sociaux) foisonnent dans divers domaines et disciplines : sciences du langage, informatique, information-communication, psychologie, etc. en France et ailleurs. Ce n'est pas le lieu ici pour recenser les nombreux travaux qui existent, mais je m'arrête un instant sur un point récurrent pour les journalistes : l'orthographe. Si « le niveau d'orthographe des jeunes » est sans cesse critiqué, les recherches, entre autres, de psychologues et de linguistes(-informaticiens) (Bernicot et Bert-Erboul 2014 ; Bernicot et al. 2014a,b ; Combes et al. 2014 ; Cougnon 2015 ; Cougnon et al. 2016 ; Fairon et al. 2006a ; Lénaïs Maskens et Fairon 2015 ; Volckaert-Legrier et al. 2009 ; Wood et al. 2013) révèlent un aspect positif : une « pluricompétence » (« multi-skills » (Cougnon et al. 2016)) plutôt qu'une « incompétence » des scripteurs.

> Chez les jeunes adolescents le registre SMS présente deux caractéristiques fortes qui s'acquièrent dans les interactions : des formes orthographiques différentes de celle de l'écrit traditionnel, et une structure dialogique différente de celles des interactions traditionnelles orales ou écrites. De plus, la pratique des SMS

> n'a pas d'influence sur l'orthographe des collégiens, c'est leur niveau en orthographe qui détermine le type de fautes dans les SMS. Des études récentes pour les langues anglaise et finlandaise (Wood et al. 2013) vont dans le même sens. L'écrit traditionnel et l'écrit SMS constituent deux registres indépendants régis par les mêmes capacités cognitives symboliques. Les SMS, avec les autres pratiques numériques, ne sont pas une menace pour l'orthographe, mais une occasion nouvelle et supplémentaire de pratiquer l'écrit, qui auparavant, pour des enfants de 11-12 ans, était restreint au milieu scolaire et à quelques cartes postales. Dans les SMS, si en moyenne 58 % de mots contiennent des textismes, 48 % sont écrits selon les règles traditionnelles. LES SMS pourraient être utilisés comme un allié pour les apprentissages scolaires en se basant sur quatre faits : les élèves pratiquent cette forme de communication (en écriture comme en lecture) avec facilité et enthousiasme, aucune étude n'a démontré de lien négatif de cette pratique avec la maîtrise de l'écrit traditionnel, un pourcentage important d'élèves possède un téléphone mobile qui constitue l'une des nouvelles technologies les moins onéreuses. Pour toutes ces raisons, le téléphone mobile et les SMS pourraient être utilisés comme support d'apprentissages scolaires (et ne plus être réservé uniquement aux échanges entre proches). L'UNESCO a publié, dès 2010, un document appelant au développement de ce type de projet.
>
> (Bernicot et Bert-Erboul 2014, p. 146-147).

Les travaux de (Bernicot et al. 2014a) ont fait un « buzz » médiatique en 2014 suite à la publication de leur communiqué de presse « Les SMS, une menace pour l'orthographe des adolescents ? », le 18 mars 2014, http://www2.cnrs.fr/presse/communique/3475.htm. Cela a été relayé à maints endroits, dès le lendemain : télévision/Web, radio, journaux écrits/en ligne [155].

Quel impact ces reportages et articles auront-ils eu vraiment ? Les normes sont tellement difficiles à bousculer. Il serait vraiment intéressant que nous, (enseignants-)chercheurs, arrivions à contribuer à des projets innovants tels

155. JT, *France 2* : https://www.youtube.com/watch?v=km2D3GVncVs, JT Régional *France 3* : https://www.youtube.com/watch?v=cyc-ihfe-UU), *France Info* : https://www.youtube.com/watch?v=-s3DFzjWf3Q), « Écrire "SMS" ne nuit pas à l'orthographe », Blog, *Le Monde*, 19 mars 2014, http://lemonde-educ.blog.lemonde.fr/2014/03/19/ecrire-sms-ne-nuit-pas-a-lorthographe/, « Les élèves font la différence entre SMS et langage courant », *L'Express*, 19 mars 2014, http://www.lexpress.fr/education/orthographe-les-eleves-font-la-difference-entre-sms-et-langage-courant_1501468.html

que ceux préconisés par l'UNESCO [156], en 2014, sur l'utilisation des TIC en tant que véritable support d'apprentissage (alphabétisation, *numératie*, etc.).

3.4.5.2 Depuis autrui

Près de 15 articles de presse, reportages télévisuels, contiennent l'un des mots suivants dans leur titrage : « smiley », « émoticône », « emoji ». Pour l'unique année 2015, 10 articles portent sur ce phénomène. Actualité oblige : l'engouement médiatique pour les emojis a eu une conséquence directe sur les demandes d'entretien à mon encontre.

J'ai commencé à m'intéresser aux *binettes* (appellation québécoise, introduite en 1995 et remise à jour en 2016, http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26534418), aux *frimousses* (appellation officielle, publiée dans le *Journal officiel* du 16/03/1999 *FranceTerme*, http://www.culture.fr/franceterme — que personne, à ma connaissance, n'utilise), aux *émoticônes*, aux *emoji* (l'*Office québécoise de la langue française* a incorporé ce terme en 2016, http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26532301 [157]), grâce aux journalistes qui me sollicitaient sans cesse sur le sujet.

L'année 2011 a marqué un tournant : les emoji, introduits dans les iPhones japonais dès 2007, ont été mis à disposition pour le public européen en 2011. Comme notre collecte s'est déroulée fin 2011, nous avons quelques centaines d'exemples d'emoji (précisément 378, *vs*. 30 000 *binettes ponctuatives*, terme que j'ai adopté pour les émoticônes incluant des caractères de type ponctuatif, directement saisissables sur un clavier). (*cf*. figure 3.49 pour le *top ten* du corpus *88milSMS*). [158]

156. « Exploiter le potentiel des TIC dans l'enseignement et l'apprentissage des compétences de base. Programmes efficaces d'alphabétisation et de numératie utilisant la radio, la télévision, le téléphone mobile, les tablettes et les ordinateurs. », http://unesdoc.unesco.org/images/0023/002317/231726f.pdf ; Sélection d'études de cas à http://www.unesco.org/uil/litbase.

157. URL consultés le 12 janvier 2017.

158. Si on compare ces chiffres au corpus *SMS4science*, on constate une évolution de l'utilisation des binettes ponctuatives, entre 2004 et 2011. Les 10 émoticônes les plus fréquentes en 2004 sont, par ordre décroissant : 1. «;-) » 2. «:-) » 3. «:p » ou «:P » 4. «:) » 5. «;) » 6. «:-(» 7. « @+ » 8. «:d » ou «:D » 9. «:(» 10. « <3 ». La binette japonaise (double accent circonflexe) est la plus changeante : elle passe de 5 occurrences (moins d'1 % en 2004) à plus de 6 500 occurrences (22 % du total des binettes ponctuatives) en 2011). Le cœur, très utilisé en France en forme d'emoji également,

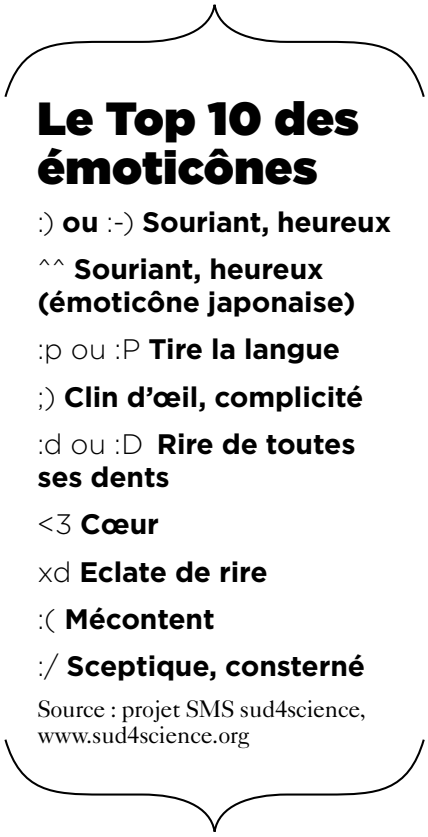


FIGURE 3.49 – Binettes ponctuatives du corpus *88milSMS*. *Cf.* M. *Le magazine du Monde*, 2/11/2013.

Dans ce paragraphe, je n'entre pas dans un débat scientifique sur le sujet, mais je fournis les types d'informations écrites que j'envoie aux journalistes, suite à des entretiens (téléphoniques, de visu ou par courriel), dans un souci de clarté pour une lecture par le grand public.

Par exemple, voici un extrait du courriel que j'ai envoyé à Frédéric Strauss (Journaliste à *Télérama*, en juin 2016) à propos des emoji, suite à un entretien téléphonique d'environ une heure.

monte de la 10e à la 7e position.

J'ai observé trois situations distinctes dans lesquelles les emoji sont utilisés dans notre corpus *88milSMS* (http://88milsms.huma-num.fr/) :

1. Un emoji peut correspondre à de l'information redondante par rapport au contenu textuel, mais il aide à renforcer, à appuyer ce que l'on écrit.

Hehehe, faut séduire le ventre avant tout 🍦🍨🍩🍪🎂🍫🍮🍯

Code dans la poche🚗🚘 !!!

2. Un emoji/emoticône peut orienter vers une interprétation du message (si le contenu textuel est ambigu, ou s'il y a des malentendus potentiels, par exemple). Dans ce cas, il peut être un véritable adoucisseur.

Exemple extrait de *88milSMS* : Mais tu fais chier aussi là :D

3. Un emoji peut apporter une information absente du contenu verbal (c'est donc une fonction lexicale)

Ok. Courage et à demain. Nos ados sont des 😇.Non, plutôt des😈😈

Pour information, dans notre corpus *88milSMS*, l'emoji du cœur est le plus fréquent, suivi de l'emoji du visage souriant aux yeux rieurs, puis, en troisième position, du visage clignant d'un œil.

De mon point de vue l'utilisation des emojis constitue un enrichissement de l'écrit, si ceux-ci sont utilisés en même temps que les données textuelles.

Quand nous avons une conversation en face-à-face avec autrui, nous échangeons des paroles, mais également des regards, des gestes, des expressions faciales, etc. Autrement dit, nous communiquons de manière verbale quand nous parlons, mais également de manière non-verbale avec notre corps. L'intonation de notre voix correspond à de la communication para-verbale.

Dans les textes courts écrits échangés en ligne ou par téléphone (courriels, chats, forums, tweets, SMS, etc.) il peut être important d'ajouter de l'expressivité, de réinjecter des émotions, car la « communication électronique médiée » (néologisme de Panckhurst) peut paraître parfois un peu sèche, brutale, déshumanisée.

Les emoji (« e » = image ; « moji » = caractère/lettre) permettent d'ajouter une dimension non verbale à l'écriture présente dans les SMS (que j'appelle « écriture SMS » ou eSMS, (Panckhurst 2009)). Les emoji peuvent alors être ressentis comme des adoucisseurs, mais aussi comme une façon ludique de réintroduire les émotions.

Les emoji fonctionnent bien pour maintenir un lien, pour faire un clin d'œil à l'autre, en tant qu'adoucisseurs, ou pour apporter une dimension ludique, etc. Il s'agit d'une dimension hautement afffective. Cela correspond à un ajout non-

> verbal, visuel.
> On ne parle pas emoji. À mon avis, les emoji ne deviendront pas une sorte d'esperanto visuel.
> Si un emoji remplace un mot, il le fait de manière très ponctuelle.
> Dans le discours d'Obama (janvier 2015), on remarquera qu'il n'est pas composé exclusivement d'emoji : http://www.theguardian.com/us-news/ng-interactive/2015/jan/20/-sp-state-of-the-union-2015-address-obama-emoji.
> Les verbes, les adjectifs, les prépositions, les conjonctions, etc. sont écrits normalement. Dans certains emoji l'écrit est utilisé, par exemple, l'emoji pour « SOS », 🆘. [159]
>
> Il est peut-être facile d'utiliser les emoji, mais gare aux interprétations erronées, notamment dans des situations interculturelles [160] Puis, je ne suis pas certaine que cela soit si facile de s'exprimer exclusivement en emoji, ou tout au moins d'exprimer certaines nuances dans nos sentiments.
> Par exemple, je ne vais pas utiliser l'emoji avec un bisou sur la joue 😘, pour dire « bises » à un copain masculin car cela pourrait être ressenti comme étant ambigu. Par contre, je peux très bien écrire « bises », plus neutre.
> Puis, les emoji peuvent changer en fonction des types de téléphone utilisés. Par exemple, si vous regardez l'unicode des emoji, ici : http://unicode.org/emoji/charts/full-emoji-list.html vous verrez que ceux-ci changent assez d'un type de téléphone à un autre. Est-ce que cela va dans le sens de l'universalité ? Ce n'est pas si évident. Cette recherche le démontre : http://grouplens.org/ [161]
> Par ailleurs, certaines applications ont vu le jour ces dernières années, par exemple *Emojli*, http://emoj.li/. Selon moi, cela correspond à un phénomène de mode. L'application a d'ailleurs été arrêtée en juillet 2015.

Même si peu de phrases ont été retenues à partir de mes indications écrites et de notre entretien téléphonique, celles-ci sont suffisantes pour faire connaître notre projet :

> Des universitaires se sont penches sur nos SMS, leurs formules courtes et leurs emoticones. Bonne nouvelle : ces messages desinhibent leurs auteurs. Et jouent avec la langue en sourdine… sans la menacer. […] À Montpellier, ou s'est etendue la

159. Ou encore l'expression maorie, *Kia Kaha* (« Soyez fort(e) ») 🆘.
160. Par exemple, il n'est pas certain que les *emotiki* (*cf.* figure 3.50) soient compris de tous.
161. URL consultés le 12 janvier 2017.

> recherche *sms4science*, Rachel Panckhurst, maitre de conferences en linguistique-informatique, evoque un test comparatif etonnant : « J'ai eu l'occasion de presenter des textes rediges dans des SMS et d'y meler des textes tires de cartes postales de poilus : il arrive qu'on ne puisse pas voir la difference entre l'ecriture de 2014 et celle de 1914. Le SMS entre dans l'histoire des pratiques scripturales quotidiennes. Ce n'est pas une autre langue. » Sa specificite reside surtout dans son immediatete : sitot envoyes, sitot distribues, ces messages s'echangent en direct et vont directement a l'essentiel. Ils permettent de passer sous silence les formules de politesse. Et meme de se passer de mots… […]
>
> Au lieu de taper sur une lettre, on tape alors sur une image appelee emoji ou emoticone : un pouce dresse pour dire « parfait » ou « tout va bien », une cible au centre de laquelle est plantee une fleche pour dire « c'est exactement ca »… L'an dernier, un discours d'Obama a ete traduit en emoticones par le quotidien britannique *The Guardian*. Certains mots n'ont pas trouve d'equivalent et ont ete maintenus, mais l'exercice se voulait surtout symbolique, suggerant que de la pratique du SMS pourrait peut-etre un jour surgir un nouveau langage, qui se lirait mais ne se parlerait plus… À travers les emoticones, la vraie nature du texto s'eclaire : il est moins charge de sens que de sensations, d'emotions. « L'image du cœur est la plus utilisee dans les SMS ; viennent ensuite celle du visage qui sourit et celle du visage qui fait un clin d'œil, cela montre qu'on est dans le relationnel plus que dans l'information, note Rachel Panckhurst. Les SMS peuvent etre percus comme un mode de communication deshumanise, on a besoin d'y reinjecter de l'expressivite, de l'affectif. »
>
> (Extraits de l'article « C + s1mple en mo2 sil1ce », Frédéric Strauss, *Télérama*, 03/08/2016).

En effet, les apports, les influences, semblent être mutuels. Nos recherches peuvent avoir un effet sur autrui, tout comme les questionnements d'autrui peuvent nous inciter, nous aiguiller, nous pousser à effectuer des études ciblées sur des points d'actualité, entre autres phénomènes.

FIGURE 3.50 – 200 emojis, ou *emotiki*, reflétant la culture maorie en Nouvelle-Zélande/Aotearoa, introduits en 2016 (*http ://www.emotiki.com/*).

4

Future horizons

A meal isn't complete without dessert.

As at the onset, this end, or this new beginning — this meeting point, shall we say? — is in English. My first language has a habit of popping out at specific moments.

This journey, this looking back over the past decades, stepping back to observe from the interior, has been a very stimulating exercise for me. *Je ne regrette rien*…or maybe just one thing: having imposed a long reading session on the jury members. Sorry.

My very late *habilitation* indeed, can be explained — I do hope you agree — from the perspective of having required the time (ok, granted, a long time), to wander, to gather, to reflect, to explore, to teach, to learn, to exchange, to stumble, to rise, to share, *de parcourir différents étapes*. As in any career, *le chemin a été parfois semé d'embûches*.

I have learnt an enormous amount during this past year. For one, I really enjoyed taking some sustained time just for myself, to ponder on many issues. For two, I realise there are still many things that I would like to accomplish over the next ten years.

4.1 Looking back

My three-fold research *triptyque* started out in a conventional way, applying what I had learnt to do in a Natural Language Processing (NLP) perspective (volet 1, § 3.3.1). As my interaction with students progressed, my research networks grew, and my administrative responsibilities emerged, my interests shifted, and branched out to focus on two aspects: on the one hand, training, evaluation, eLearning pedagogical networks (volet 2, § 3.3.2) and, on the other hand, computer-mediated communication (CMC), *communication médiée par ordinateur (CMO)*, mediated electronic discourse (MED), *discours électronique médié (DEM)*, mediated digital discours (MDD), *discours numérique médié (DNM)* (volet 3, § 3.3.3).

Computational linguistics tools, my initial programming, did indeed pave the way to my next step: using tools and helping others to use them efficiently, all the while being involved in software evaluation, and interacting with a vast number of researchers from other disciplines. Also, my 2 ½ years directing our *METICE* service (*Multimédia, Enseignement, Technologies de l'Information et de la Communication Éducatives*), including the distance-education service of the University, in order to initiate the transition from paper to online access (among other factors), gave me important insight into how to help colleagues and students move to the digital era. This period (1999-2001) and the teaching-training sessions I initiated and taught (1996-2003), had a strong influence on how I perceived applied research.

Setting up the SMS project [1] — and working with Cathérine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, Bertrand Verine — has indeed been the highlight, the *apogée*, of my career. The *sud4science LR* project, followed on by the DGLFLF one, both of which culminated in the *88milSMS* corpus, were for me, pure examples of applied research. This meant I was able to discover a number of new *facettes* which are not necessarily systematically investigated by academics, and meant I needed to sometimes go beyond the University boundaries out into the *real world*:

1. Thanks to the initial incentive by Cédrick Fairon, Jean René Klein, Sébastien Paumier, Louise-Amélie Cougnon, through the *SMS4science* project *pour ne citer qu'eux*.

1. Legal advisors (*Correspondant Informatique et Libertés, CIL*);
2. Communications department (*communiqués de presse*, etc.);
3. Local firms: persuading them to contribute prizes (during the text-message collection);
4. Media (Local, national, international, written & online press, radio, television);
5. Pluridisciplinary research;
6. Student internships, leading to student authored publications.

However, I would not like the SMS-related projects to take all of the limelight. Each moment, each path, has contributed a crucial stepping stone to this *ensemble*. I am reminded of my co-author, here: Gilles Pérez, for our first book, published in 2000 (Panckhurst and Pérez 2000). The two following books were co-edited with Laurence Vincent-Durroux (Vincent-Durroux and Panckhurst 2002), then Sophie David and Lisa Whistlecroft (Panckhurst et al. 2004a). Debra Marsh and I also co-authored numerous articles between 2007 and 2011, before the text message research took over and became my full-time research project.

4.2 What next?

I intend to strive in order to achieve the following 5 (unordered) key-points — which I consider to be fundamental for research — during the next ten years. Maybe I can be accused of being utopian:

1. Deliver crucial research information to the general public.
2. Request cross-disciplinary PhDs.
3. Demand that research results be factored into Ministerial reforms.
4. Help provide scientific expertise for devising real-life applications/software.
5. Continue applied research — including PhD supervision — and help link up academic institutions with other organisations.

I am quite proud that in our recent SMS project we have had such attention from the media, which in turn helps get information out — even if the interviews are, at times, extremely time-consuming. But we need to continue. I feel we have a strong debt to the general public. Our projects are funded with public money. I would therefore like to write a book — *en termes limpides*, that, I hope, would be handy for all — about evolving scriptural practices in French in the 21st century. [2] I need to explain to the general public why we study these phenomena and convince them that it is positive to do so. These sorts of books exist in English, but we need to provide more of them for French [3] I hope a workshop/conference on how the *88milSMS* corpus downloads have been used in different domains could also be of interest.

Many people have very conventional views on text messages. In our DGLFLF report, we addressed this, and wrote the following paragraph, related to two ever-present questions: "Why do you want to study text-messages which are written in bad French?", "What do you gain for the field of linguistics by doing so?" I would like to pursue.

> Lors de la préparation de notre collecte, nous avons discuté à maintes reprises avec des personnes du grand public qui posaient systématiquement les questions suivantes aux linguistes de l'équipe : "Pourquoi voulez-vous étudier les SMS qui sont écrits dans un mauvais français ?" "Qu'est-ce que cela apporte à votre discipline, les sciences du langage ?" Nous répondions de la manière suivante : en tant que linguistes, nous observons, sans jugement, sans nous référer à une norme quelconque. Nous nous intéressons à l'étude du langage, des langues et des pratiques langagières, donc, lorsqu'il y a des mutations éventuelles, nous saisissons l'occasion pour étudier de nouveaux phénomènes. La langue est dynamique, en mouvance constante. Cette collecte a permis de recueillir en grand nombre des utilisations spontanées de la langue française. À partir de là, nous pouvons les comparer avec ce que nous connaissons d'autres types d'utilisations : par exemple,

2. *Cf.* (Cougnon 2015) for her excellent book, emanating from her PhD, *Langage et SMS*, an international study on current writing practices; (André 2017) for his PhD, *Pratiques scripturales et écriture SMS : analyse linguistique d'un corpus de langue française*.

3. Maybe I will come back to *mes premières amours* : question forms in French. Negation (Stark 2011) is also of interest, as are other morphosyntactic issues (Stark 2014) or indeed semantic/pragmatic trends of *unités verbales polylexicales*, for instance — or something entirely different — in relation to current digital discourse trends.

> les écrits plus normés sur papier, ou encore les courriers électroniques, mais aussi les échanges oraux. Nous pourrons en tirer des conclusions sur la graphie, sur l'écriture, sur les choix de vocabulaire, sur la construction des phrases, sur la façon de s'adresser à l'autre, etc. Tous ces éléments varient selon les personnes, selon les âges et selon les situations. Par exemple, nous allons pouvoir étudier de quels mots se servent les gens pour entrer en contact (*salut, hello, coucou*) ou s'ils entrent immédiatement dans le vif du sujet ; ou encore, de quels mots se servent-ils pour solliciter leur destinataire. Vont-ils utiliser leur prénom, un mot doux, un surnom, etc. ? Puis, en étudiant les SMS, nous pouvons envisager des applications industrielles en TALN : des systèmes de transcodage SMS vers le français standardisé, pour des personnes qui ne comprendraient pas une écriture SMS très codée ; des logiciels de reconnaissance des SMS, ou de vocalisation, pour des personnes aveugles, ou pour des conducteurs, etc. Ces applications pourront aussi aider à traiter de grandes masses de données textuelles très présentes aujourd'hui dans les réseaux sociaux (tweets, Facebook). (DGLFLF report, 2013).

Currently, PhD candidates can have two co-directors. However, they are always (to my knowledge) enrolled in solely one discipline, from the point of view of the CNU (*Conseil National des universités*) in France. Certain students truly cover two disciplines, between linguistics (CNU 7) and computer science (CNU 27), for instance. Why not open up the access to a double PhD enrolment? Candidates for Senior lecturer/Associate Professor (*maîtres de conférences*) positions — once they are qualified for a particular CNU discipline — are already also allowed to apply for a job within any other CNU discipline. Why not include this legislation at an earlier level, if it is indeed pertinent? For instance, the current *Data Scientist* diploma trend requires competencies, aptitudes, on several levels.

The *Mediapart* example I gave earlier (§ 3.4.5.1) shows to what extent the research results are not sufficiently communicated to ministerial spheres. This aspect needs to be greatly improved, in order to avoid biases, *préjugés*, stereotypes, popular beliefs, *idées reçues*.

I am also very interested in real-life applications/software for improving people's daily lives. We have come a long way in *voice recognition* and *speech synthesis* over the decades. To indicate one example, I hope that our current SMS research will provide some insight into how we can modify electronic lexica in order

to improve vocal tools used by the blind and/or those who are momentarily impeded from writing on their mobile devices, as suggested above.

Finally, I feel that academics need to spend more time off-campus, mingling with people from other walks of life, in order to understand how their own research can become truly applied and useful for all. Links between Universities and other institutions/private enterprise are also crucial [4]. All of these aspects are important for effective PhD supervision.

After having initially created some of my own prototypes, having then used and evaluated tools by others, on my next path, I would like to continue to collaborate in a relaxed pluridisciplinary environment, with trustworthy colleagues emanating from both public and private sectors. For instance, the new Belgian project (organised by the CENTAL (http://www.vospouces.org/) is very stimulating. Analysing the collected data (from *WhatsApp* and other applications) will provide openings for further computational linguistical "conversational" analyses, of the type I have mentioned earlier (p. 176) and elsewhere (Panckhurst and Moïse 2014). In turn, these avenues will provide interesting opportunities for future doctoral candidates.

Like the feeling of contentment after *une dégustation*, *un menu gastronomique* in a top-notch restaurant, I feel satiated, *repue*. I have been very privileged to have been able to share the starters, the tapas and the *plat principal* with many generous family members, friends and colleagues. I am now ready for the dessert, *mon pêché mignon*. For the ten remaining years of my academic career, *l'avenir, que me réserve-t-il ?* I would like to continue tasting and indeed discover yet other wonders of the research world: *cheesecakes, crèmes brûlées, pastéis de Belém, mince pies, pavlovas, tiramisus, makroutes, mousses au chocolat, gelati, baklavas, brigadeiros, waffles, lamingtons, zitos, churros*, to name a few delights.

4. I am very grateful that I have be able to continue (beyond our *sud4science/88milSMS* project) working with Cédric Lopez, at *Viseo*, Grenoble. His R&D director, Frédérique Segond, fortunately has a very open view about research, and encourages these sorts of collaborations. For instance, currently, an M2 student from Grenoble-Alpes University, is conducting an internship at *Viseo* on normalization — she is co-supervised by Cédric Lopez and Georges Antoniadis — a good example of essential public/private collaboration.

Desserts, and food in general, are like research. In the conclusion of my recipe book (*Fait Maison. Recipes from a Kiwi in France*, 2014), I wrote the following:

> On 16 November 2010, France became the first country in the world to have its gastronomy classified as a Unesco "intangible cultural heritage of humanity". But this cultural heritage is not about fine restaurants and luxury cuisine; it concerns the everyday normal relationship with anything culinary, the social ties, the discourse about what one eats. Gastronomy in France is a popular form of culture, shared by all French people, not just the elite. (Panckhurst, 2014, *Fait Maison. Recipes from a Kiwi in France*, https://leanpub.com/faitmaison).

My dream is that this culinary quote could also be applied to research, by changing a few words, here and there:

> In 2017, France became the first country in the world to have its research classified as a Unesco "intangible cultural heritage of humanity". But this cultural heritage is not about fine research teams and luxury laboratories; it concerns the everyday normal relationship with anything to do with applied research, the social ties, the discourse about what one studies. Research in France is a popular form of culture, shared by all French people, not just the elite.

5

Bibliographie générale

5.1 Sélection de mes publications

Remarque. — Ne sont incluses que les publications parues ou sous presse.

Accorsi, Pierre, Namrata Patel, Cédric Lopez, Rachel Panckhurst et Mathieu Roche (2014). « Seek and Hide : Anonymising a French SMS corpus using natural language processing techniques ». In : *SMS Communication. A Linguistic Approach*. Sous la dir. de Louise-Amélie Cougnon et Cédrick Fairon. Amsterdam/Philadelphia : John Benjamins, p. 11–28.

Amar, Muriel, Sophie David, Rachel Panckhurst et Lisa Whistlecroft (2008). « Classification procedures for software evaluation ». In : *Actes du colloque LREC*. Marrakech, p. 623–630. url : www.lrec-conf.org/lrec2008/.

Charnet, Chantal et Rachel Panckhurst (1998). « Le correcteur grammatical : un auxiliaire efficace pour l'enseignant ? Quelques éléments de réflexion ». In : *ALSIC* 1.2, p. 103–114. url : https://alsic.revues.org/1494.

David, Sophie et Rachel Panckhurst (2004). « Questionnaire results : from the competitors' point of view ». In : *Evaluation in e-learning : the European Academic Software Award*. Sous la dir. de Rachel Panckhurst, Sophie David et Lisa Whistlecroft. MédiaTic 3. Service des publications, Université Paul-Valéry Montpellier 3, p. 83–88.

David, Sophie, Rachel Panckhurst et Lisa Whistlecroft (2005). « Revising the evaluation procedure of the European Academic Software Award, Eunis ».

In : *Actes du colloque Eunis*. Manchester. URL : http://web.archive.org/web/20061009022604/http://www.mc.manchester.ac.uk/eunis2005/medialibrary/papers/paper_111.pdf.

Lopez, Cédric, Reda Bestandji, Mathieu Roche et Rachel Panckhurst (2014). « Towards Electronic SMS Dictionary Construction : An Alignment-based Approach ». In : *Actes du colloque LREC*. Reykjavik, p. 2833–2838. URL : http://www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf.

Lopez, Cédric, Mathieu Roche et Rachel Panckhurst (2015). « Classification des items inconnus de 88milSMS : aide à l'identification automatique de la créativité scripturale ». In : *Tranel (Travaux neuchâtelois de linguistique)* 63, p. 71–86. URL : https://www.unine.ch/files/live/sites/islc/files/Tranel/63/71-86_lopez_al_corr.pdf.

Panckhurst, Rachel (2001, 2017). « Discours électronique médié ; Linguistique(s) de corpus ; Néographie ; Traitement automatique des langues ». In : *Termes et concepts pour l'analyse du discours. Une approche praxématique*. Sous la dir. de Catherine Détrie, Paul Siblot, Bertrand Verine et Agnès Steuckardt. Nouvelle édition augmentée. Paris : Honoré Champion, p. 103–105, 205–207, 239–240, 406–408.

– (1992a). « Comment allier les besoins du linguiste et l'utilisation intelligente de bases de données ? » In : *Actes du XVe Congrès international des Linguistes (CIL)*. Québec : Les Presses de l'Université Laval, Sainte-Foy, p. 301–304.
– (1992b). « Description linguistique et implémentation en FX des structures interrogatives (directes) du français. Résumé de thèse. » In : *Traitement automatique des langues (T.A.L.), Spécial trentenaire*. 33.1-2, p. 250–252.
– (1993a). « Analyseurs et bases de données pour des besoins spécifiques ». In : *Actes du colloque Informatique et langue naturelle, ILN '93*. Nantes, p. 207–222.
– (1993b). « Scatlex : une aide informatisée pour la construction d'entrées lexicales verbales ». In : *Revue de la liaison de la recherche en informatique cognitive des organisations (ICO)* 5.3, p. 61–67.
– (1994a). « A Database for linguists : intelligent querying and increase of data. » In : *Computers and the Humanities* 28.1, p. 39–52.
– éd. (1994b). *Cahiers de Praxématique, PULM* 22. URL : http://praxematique.revues.org/1887.

– (1994c). « Constitution d'une base lexicale verbale. » In : *Actes du colloque Consensus Ex Machina, Association for Computers and the Humanities/Association for Literary and Linguistic Computing, ACH/ALLC*. Paris, p. 187–188.
– (1994d). « Présentation ». In : *Cahiers de Praxématique* 22, p. 5–6. URL : http://praxematique.revues.org/1890.
– (1994e). « Une structure de classification verbale basée sur des contrastes ». In : *Cahiers de Praxématique* 22, p. 105–134. URL : http://praxematique.revues.org/2275.
– (1995a). « Behind the scenes : Building a tool for Verb Classification in French. » In : *Actes du colloque Association for Computers and the Humanities/Association for Literary and Linguistic Computing, ACH/ALLC*. Santa Barbara, p. 89–92.
– (1995b). « Décrire le système verbal indépendamment d'un cadre grammatical. » In : *Actes du colloque Le traitement automatique du langage naturel, TALN*. Marseille, p. 172–180.
– (1995c). « Poly…quelque chose et classification verbale. » In : *Actes du colloque Lexiques-Grammaires comparés et traitements automatiques, Université du Québec à Montréal (UQAM)*. Montréal, p. 199–206.
– (1996a). « Formation en linguistique-informatique : une expérience montpelliéraine ». In : *Traitement automatique des langues (T.A.L.), Enseignement du TAL*. 37.1, p. 51–64.
– (1996b). « Linguistique-informatique : la crise (Addendum) ». In : *Traitement automatique des langues (T.A.L.), Grammaire et théorie de la preuve* 37.2, p. 176–177.
– (1996c). « Quelques problèmes posés pour l'analyse automatique des unités verbales en français. » In : *Informatique et langue naturelle (ILN)*. Nantes, p. 465–476.
– (1997a). « La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? » In : *Terminologies nouvelles* 17, p. 56–58.
– (1997b). « Sens et informatique ». In : *Hommages à Xavier Mignot*. Sous la dir. de Paul SIBLOT. Université Paul-Valéry Montpellier 3, p. 115–130.
– (1998a). « Analyse linguistique du courrier électronique ». In : *Communication, société et internet, Actes du colloque Les relations entre individus médiatisées par les réseaux informatiques*. Sous la dir. de Nicola GUÉGUEN et Laurence TOBIN. GRESICO. Paris : L'Harmattan, p. 47–60.

Panckhurst, Rachel (1998b). « Des unités verbales polylexicales ». In : *De l'actualisation*. Sous la dir. de Jeanne-Marie Barbéris, Jacques Bres et Paul Siblot. Paris : CNRS-Éditions, p. 161–178.
– (1998c). « Marques typiques et ratages en communication médiée par ordinateur ». In : *Actes du colloque CIDE 98*. INPT, Rabat : Paris : Europia Productions, p. 31–43.
– (1999a). « Analyse linguistique assistée par ordinateur du courriel ». In : *Internet, communication et langue française*. Sous la dir. de Jacques Anis. Paris : Hermès, p. 55–70.
– (1999b). « La Communication médiée par ordinateur : un discours autre ? » In : *L'autre en discours*. Sous la dir. de Jacques Bres, Régine Delamotte-Legrand, Françoise Madray et Paul Siblot. Dyalang-Praxiling, Service des publications de l'Université Paul-Valéry Montpellier 3, p. 307–331.
– (2001a). « Distance, open and virtual lifelong learning : shaping the transition within a French University ». In : *Proceedings, 20th World conference on open learning and distance education*. Sous la dir. de Norvège ICDE – Oslo et Allemagne FernUniversität Hagen. Düsseldorf. isbn : ISBN-NR.3-934093-01-9.
– (2001b). « Les unités verbales polylexicales : problèmes de repérage en traitement automatique ». In : *La locution et la périphrase du lexique à la grammaire*. Sous la dir. de Francis Tollis. Paris : L'Harmattan, p. 55–63.
– (2003a). « Computer-mediated communication and linguistic issues in French University online courses ». In : *Actes du colloque Online Educa*. Berlin, p. 454–457.
– (2003b). « La glose, le document électronique et l'extraction automatisée ». In : *Le mot et sa glose*. Sous la dir. d'Agnès Steuckardt et Aïno Niklas-Salminen. T. Langues et langage. 9. Publications de l'université de Provence., p. 271–292.
– (2006a). « Le discours électronique médié : bilan et perspectives ». In : *Lire, écrire, communiquer et apprendre avec Internet*. Sous la dir. d'Annie Piolat. Marseille : Éditions Solal, p. 345–366.
– (2006b). « Mediated electronic discourse and computational linguistic analysis : improving learning through choice of effective communication methods ». In : *Actes du colloque ascilite*. Sydney, p. 633–637. url : http://www.ascilite.org/conferences/sydney06/proceeding/pdf_papers/p16.pdf.

– (2009). « Short Message Service (SMS) : typologie et problématiques futures ». In : *Polyphonies, pour Michelle Lanvin*. Sous la dir. de Teddy ARNAVIELLE. Université Paul-Valéry Montpellier 3, p. 33–52.
– (2010). « Txtng in three European languages : does the linguistic typology differ ? » In : *Actes du colloque i-Mean*. University of the West of England, Bristol, p. 122–137. URL : http://www2.uwe.ac.uk/faculties/CAHE/Documents/Conferences/imean/IMEAN-Conference-Proceedings-2009.pdf.
– (2011). « Réseaux sociaux et pédagogie dans un contexte d'enseignement supérieur français ». In : *2e journée TICE à l'UAPV*. Avignon, Voir la vidéo de l'intervention (47 minutes). URL : https://pod.univ-avignon.fr/video/0108-r-panckhurst-journee-tice-2011/.
– (2012). « The digital tutor : accepting to lose control and make mistakes ». In : *Actes du colloque ascilite "Future challenges/Sustainable challenges"*. Wellington, p. 735–739. URL : http://www.ascilite.org/conferences/Wellington12/2012/images/custom/panckhurst,_rachel_-_the_digital.pdf.
– (2013). « A large SMS corpus in French : from design and collation to anonymisation, transcoding and analysis ». In : *Proceedings du colloque CILC*. Sous la dir. de Social PROCEDIA et Elsevier BEHAVIOURAL SCIENCES. Alicante. URL : http://www.sciencedirect.com/science/article/pii/S1877042813041475.
– (2016a). « A digital corpus resource of authentic anonymized French text messages : 88milSMS—What about transcoding and linguistic annotation ? » In : *Digital Scholarship in the Humanities*. DOI : 10.1093/llc/fqw049.
– (2016b). « De Sud4science à 88milSMS (un grand corpus de SMS authentiques) : entre linguistique et informatique ». In : *Conférence invitée*. ENS, Lyon. URL : http://cle.ens-lyon.fr/conf-aperos/de-sud4science-a-88milsms-un-grand-corpus-de-sms-authentiques-entre-linguistique-et-informatique--303624.kjsp.
– (2016c). « Les SMS en langue française, Conférence invitée ». In : Médiathèque André-Malraux, Béziers. URL : https://www.youtube.com/watch?v=jMH2a2v8Qdo.
PANCKHURST, Rachel et Tayeb BOUGUERRA (2003). « Communicational and methodological/linguistic strategies using electronic mail in a French University ». In : *Actes du colloque 8th International Symposium on Social Communication*. Santiago de Cuba, p. 548–554.

Panckhurst, Rachel et Bas Cordewener (2004). « A review of the European Academic Software Award : Year 2000 ». In : *Evaluation in e-learning : the European Academic Software Award*. Sous la dir. de Rachel Panckhurst, Sophie David et Lisa Whistlecroft. MédiaTic 3. PULM, Université Paul-Valéry Montpellier 3, p. 23–41.

Panckhurst, Rachel, Sophie David et Lisa Whistlecroft, éds. (2004a). *Evaluation in e-learning : the European Academic Software Award*. MédiaTic 3. Service des publications, Université Paul-Valéry Montpellier 3.

– (2004b). « Overview, Sommaire ». In : *Evaluation in e-learning : the European Academic Software Award*. Sous la dir. de Rachel Panckhurst, Sophie David et Lisa Whistlecroft. MédiaTic 3. Service des publications, Université Paul-Valéry Montpellier 3, xi–xiv and xv–xviii.

Panckhurst, Rachel, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche et Bertrand Verine (2013). « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS ». In : *Épistème, revue internationale de sciences sociales appliquées* Des usages numériques aux pratiques scripturales électroniques.9, p. 107–138.

– (2014a). *88milSMS. A corpus of authentic text messages in French, produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo, ISLRN : 024-713-187-947-8*. URL : http://88milsms.huma-num.fr/.

– (2014b). « Un grand corpus de SMS en français : 88milSMS ». In : *La lettre de l'InSHS, la Tribune d'HumaNum*, p. 22–25. URL : http://www.cnrs.fr/inshs/Lettres-information-INSHS/lettre_infoinshs_31hd.pdf.

– (2014c). « Une grande collecte de SMS authentiques en français : démarche, remarques et conseils ». In : *Le français à l'université* 19.03. URL : http://www.bulletin.auf.org/index.php?id=1875.

– (2015a). « Dites-le dans le français que vous voulez ! » In : *Mediapart*. URL : https://blogs.mediapart.fr/edition/les-invites-de-mediapart/article/020415/dites-le-dans-le-francais-que-vous-voulez.

– (2016a). *88milSMS. A corpus of authentic text messages in French. Version nouvelle du corpus ISLRN 024-713-187-947-8, In Chanier T. (ed) Banque de corpus CoMeRe*.

Ortolang : Nancy. cmr-88milsms-tei-v1. URL : https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1.

PANCKHURST, Rachel et Debra MARSH (2006). « A French Master's degree in eLearning : are the students' needs met ? » In : *Actes du colloque ascilite*. Sydney, p. 985–986. URL : http://www.ascilite.org/conferences/sydney06/proceeding/pdf_papers/p25.pdf.

– (2007). « eLEN — eLearning Exchange Networks : reaching out to effective bilingual and multicultural University collaboration ». In : *Actes du colloque EADTU*. Lisbon.

– (2008a). « Communities of Practice. Moving from Institutional Platforms to the open Web as a platform ». In : *Actes du colloque iLearn Forum*. Paris. URL : http://www.eife-l.org/publications/proceedings/ilf08/contributions/designing-estrategies-for-learningorganisations/panckhurst/_marsh.pdf/view.

– (2008b). « REEL : réseaux d'échanges pédagogiques en eLearning. Améliorer la qualité de l'apprentissage en favorisant l'autonomie des apprenants ». In : *25e Congrès de l'AIPU, Le défi de la qualité dans l'enseignement supérieur : vers un changement de paradigme*. Sous la dir. de Chantal CHARNET, Claire GHERSI et Jean-Louis MONINO. Montpellier, Actes consultables en ligne. URL : www.aipu2008-montpellier.fr.

– (2009). « eLEN2 — 2nd generation eLearning Exchange Networks ». In : *Actes du colloque Online Educa*. Berlin, p. 245–248. URL : http://www.online-educa.com.

– (2011a). « Les frontières pédagogiques sont-elles remises en question par l'utilisation des réseaux sociaux ? L'implémentation d'objets d'apprentissage sociaux dans un espace de communication électronique médiée ». In : *La communication électronique : enjeux de langues*. Sous la dir. de Fabien LIÉNARD et Sami ZLITNI. Lambert-Lucas, Limoges, p. 293–301.

– (2011b). « Using Social Networks for Pedagogical Practice in French Higher Education : Educator and Learner Perspectives ». In : *RUSC, Revista de Universidad y Sociedad del Conocimiento, « Globalisation and Internationalisation of Higher Education »* 8.1, Monographie en ligne. ISSN : 1698-580X. URL : http://www.raco.cat/index.php/RUSC/article/view/225632/306988.

Panckhurst, Rachel et Claudine Moïse (2014). « French text messages. From SMS data collection to preliminary analysis ». In : *SMS Communication. A Linguistic Approach*. Sous la dir. de Louise-Amélie Cougnon et Cédrick Fairon. Amsterdam/Philadelphia : John Benjamins, p. 141–168.

Panckhurst, Rachel et Gilles Pérez (2000). *Introduction aux technologies de l'information et de la communication. Problèmes et méthodes : MacOS et Windows*. MédiaTic 1. Service des publications, Université Paul-Valéry Montpellier 3.

Panckhurst, Rachel, Mathieu Roche et Cédric Lopez (2015b). « Données authentiques : un grand corpus de SMS en français ». In : *Colloque SHESL, Corpus et constitution des savoirs linguistiques*, p. 33–35. url : shesl-htl2015.sciencesconf.org/conference/shesl-htl2015/pages/Livret_resumes_SHESL_HTL2015.pdf.

Panckhurst, Rachel, Mathieu Roche, Cédric Lopez, Bertrand Verine, Catherine Détrie et Claudine Moïse (2016b). « De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire ». In : *Histoire des théories linguistiques (HEL)* 38.2, p. 73–86.

Roche, Mathieu, Bertrand Verine, Cédric Lopez et Rachel Panckhurst (2016). « La néographie dans un grand corpus de SMS français : 88milSMS ». In : *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones, Actes du colloque CINEO 2015, 22-24 octobre, Salamanque*. Sous la dir. de Joaquín García Palacios, Goedele De Sterck, Daniel Linder, Nava Maroto, Miguel Sánchez Ibáñez et Jesús Torres del Rey. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt, Peter Lang, p. 279–302. url : DOI:http://dx.doi.org/10.3726/978-3-631-69859-4.

Vincent-Durroux, Laurence et Rachel Panckhurst, éds. (2002). *Autoformation et autoévaluation : une pédagogie renouvelée ?* MédiaTic 2. Service des publications, Université Paul-Valéry Montpellier 3.

5.2 Bibliographie

Remarque. — Cette bibliographie ne contient qu'une sélection de références. Pour un ensemble plus conséquent, le lecteur pourra se reporter aux références bibliographiques à la fin de chacune de mes publications du volume II.

Abeillé, Anne (1991). « Une grammaire lexicalisée d'arbres adjoints pour le français. » Thèse de doct. Université Paris VII.

– (1993). *Les nouvelles syntaxes*. Paris : Armand Colin.

Alcouffe, Philippe et Bruno Revellin Falcoz (1994). « Notion de position du modèle GENELEX et structuration d'une base de données syntaxiques issue des Tables du LADL. » In : *Cahiers de Praxématique* 22, p. 81–104.

André, Frédéric (2014). « Écriture SMS et Phénomènes Phonétiques. Évolution des pratiques scripturales entre 2004 et 2011 », mémoire de Master 2, Sciences du Langage, Discours médiatiques, institutionnels et politiques, Université Paul-Valéry Montpellier 3, co-direction : Rachel Panckhurst et Fabrice Hirsch.

– (2017). « Pratiques scripturales et écriture SMS : analyse linguistique d'un corpus de langue française ». Thèse de doct. Université Paris-Sorbonne. Jury : Cédrick Fairon, Elisabeth Stark, Rachel Panckhurst, Sylvie Plane, Gilles Siouffi (directeur).

Anis, Jacques (1998). *Texte et ordinateur. L'écriture réinventée ?* Méthodes en sciences humaines. Bruxelles, De Boeck Université.

– éd. (1999). *Internet, communication et langue française*. Paris, Hermès.

Anis, Jacques, Michel de Fornel et Béatrice Fraenkel (2004). « La communication électronique : Approches linguistiques et anthropologiques ». In : *Colloque international, EHESS*. Paris.

Antoniadis, Georges (2008). Du TAL et son apport aux systèmes d'apprentissage des langues : Contributions, Habilitation à diriger des recherches, Université Stendhal.

– (2014). *Corpus de SMS réels dans les Alpes, smsalpes. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy*. URL : https://hdl.handle.net/11403/comere/cmr-smsalpes.

Antoniadis, Georges, Gaëlle Chabert et Virginie Zampa (2011). « Alpes4science : Constitution d'un corpus de SMS réels en France métropolitaine ». In : Sherbrooke : 79th Acfas colloquium.

Arnavielle, Teddy, éd. (2009). *Polyphonies, pour Michelle Lanvin*. Université Paul-Valéry Montpellier 3.

Arrivé, Michel, Françoise Gadet et Michel Galmiche (1986). *La grammaire d'aujourd'hui*. Paris, Flammarion.

Aw, AiTi, Min Zhang, Juan Xiao et Jian Su (2006). « A phrase-based statistical model for SMS text normalization ». In : *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics. Sydney, p. 33–40. url : http://anthology.aclweb.org/P/P06/P06-2.pdf#page=43..

Baschung, Karine (1991). *Grammaire d'unification à traits et contrôle des infinitives en français*. Clermont-Ferrand : Adosa.

Baschung, Karine, Gabriel Bès, Rachel Panckhurst et Henk Zeevat (1986). *Contextual phenomena in dialogue*. Rapp. tech. Université Blaise-Pascal Clermont 2, ACORD ESPRIT Project 393, Task 2.3.

Baylon, Christian et Xavier Mignot (1995). *Sémantique du langage. Initiation*. Fac. linguistique. Paris, Nathan.

Beaufort, Richard, Sophie Roekhaut, Louise-Amélie Cougnon et Cédrick Fairon (2010). « A hybrid rule/model-based finite-state framework for normalizing SMS messages. » In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Sous la dir. de J. Hajič et al. Association for Computational Linguistics. Uppsala : Sweden, p. 770–779.

Beaufort, Richard, Sophie Roekhaut et Cédrick Fairon (2008). « Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition ». In : *Actes du colloque JADT*, p. 155–166.

Bennett, Shirley, Debra Marsh et Clare Killen (2007). *Handbook of Online Education*. London/New York, Continuum.

Benveniste, Émile (1974). *Problèmes de linguistique générale II*. Paris, Gallimard.

Bernicot, Josie et Alain Bert-Erboul (2014). *L'acquisition du langage par l'enfant, 2e édition actualisée*. Paris, Éditions In Press.

Bernicot, Josie, Antonine Goumi, Alain Bert-Erboul et Olga Volckaert-Legrier (2014a). « How do skilled and less-skilled spellers write text messages ? A longitudinal study ». In : *Journal of Computer Assisted Learning*.

Bernicot, Josie, Olga Volckaert-Legrier, Antonine Goumi et Alain Bert-Erboul (2014b). « SMS experience and textisms in young adolescents : Presentation

of a longitudinally collected corpus. » In : *SMS Communication. A linguistic Approach*. Sous la dir. de Louise-Amélie COUGNON et Cédrick FAIRON. Amsterdam : John Benjamins, p. 29–46.

BÈS, Gabriel G. (1994). « Les tables de Méthodes en syntaxe : introduction à un mode d'emploi. » In : *Cahiers de Praxématique* 22, p. 57–80. URL : https://praxematique.revues.org/1888.

– (2002). « La linguistique entre science et ingéniérie. » In : *TAL* 43.3, p. 57–81.

BÈS, Gabriel G. et Karine BASCHUNG (1985). *Feasibility of a GPSG French grammar*. Rapp. tech. Université Blaise-Pascal Clermont 2, ACORD ESPRIT Project 393.

BÈS, Gabriel G. et Pierre-François JURIE (1989). *UCG Grammars ; the control of their descriptive adequacy*. Rapp. tech. ACORD ESPRIT Project 393, Laboratoires de Marcoussis.

BEVILACQUA, Sabrina (2012). « La communication médiée par téléphone ». In : *Synergies* 1, p. 117–126.

BILGER, Mireille, éd. (2000). *Linguistique sur Corpus, Études et Réflexions*. 31. Cahiers de l'Université de Perpignan.

BLACHE, Philippe (1994). « Le verbe en HPSG : nucleus d'un système lexicalisé ». In : *Cahiers de Praxématique* 22, p. 37–56.

– (2001). *Les grammaires de propriétés. Des contraintes pour le traitement automatique des langues naturelles*. Paris : Hermès Sciences.

BOONS, Jean-Paul, Alain GUILLET et Christian LECLÈRE (1976). *La structure des phrases simples en français : constructions intransitives*. Genève : Droz.

BOUDRIQUE, Dorian, Hélène CATAPANO et Sonia POLLET (2008). « SMS espagnols et typologie des SMS français », Dossier de Master, Université Paul-Valéry Montpellier 3, direction : Rachel Panckhurst.

BOURIGAULT, Didier (1993). « Analyse syntaxique locale pour le repérage de termes complexes dans un texte. » In : *TAL, Traitements automatiques de la composition nominale*. 34.2, p. 105–118.

BOURIGAULT, Didier et Cécile FABRE (2000). « Approche linguistique pour l'analyse syntaxique de corpus ». In : *Cahiers de Grammaire* 25, p. 131–151.

BRES, Jacques, Régine DELAMOTTE-LEGRAND, Françoise MADRAY et Paul SIBLOT, éds. (1999). *L'autre en discours*. Dyalang-Praxiling, Service des publications de l'Université Paul-Valéry Montpellier 3.

Buvet, Pierre-André, éd. (2015). *« Linguistique et informatique », Études de linguistique appliquée*. 180. Klincksieck.

Calder, Jo, Ewan Klein, Marc Moens et Henk Zeevat (1986). *A Unification Categorial Grammar Interpreter.* Rapp. tech. ESPRIT PROJECT 393 ACORD ; Deliverable 2.6.

Chambreuil, Michel (1990). *Grammaire de Montague. Langage, traduction, interprétation.* Clermont-Ferrand : Adosa.

Chanier, Thierry, éd. (2016). *Banque de corpus CoMeRe, Ortolang : Nancy.* url : https://hdl.handle.net/11403/comere.

Chanier, Thierry, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi et Djamé Seddah (2014). « The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres. Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse : Issues and Challenges at the Interface of Corpus and Computational Linguistics ». In : *JLCL (Journal of Language Technology and Computational Linguistics)* 29.2, p. 1–31. url : http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf.

Chaudiron, Stéphane, éd. (2004). *Évaluation des systèmes de traitement de 'information*. Paris, Hermès.

Combes, Céline, Olga Volckaert-Legrier et Pierre Largy (2014). « The Effet of a Dual Task on SMS Writing in Novice and Expert Adolescents ». In : *SMS Communication. A Linguistic Approach*. Sous la dir. de Louise-Amélie Cougnon et Cédrick Fairon. Amsterdam/Philadelphia : John Benjamins, p. 47–66.

Constant, Matthieu et Elsa Tolone (2010). « A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. » In : *Actes du 27e Colloque international sur le lexique et la grammaire.* Sous la dir. de Michele De Gioia. L'Aquila, p. 79–93. url : https://hal.archives-ouvertes.fr/hal-00483662/document.

Cori, Marcel, Sophie David et Jacqueline Léon (2002a). « Pour un travail épistémologique sur le TAL ». In : *TAL* 43.3, p. 7–22.

– éds. (2002b). *« Problèmes épistémologiques », TAL*. T. 43. 3.

– éds. (2008a). *« Construction des faits en linguistique : la place des corpus », Langages.* T. 3. 171. Armand Colin.

– (2008b). « Présentation : éléments de réflexion sur la place des corpus en linguistique ». In : *Langages* 3.171, p. 5–11.

Cori, Marcel et Jean-Marie Marandin (1993). « Grammaires d'arbres polychromes ». In : *TAL* 34.1, p. 105–135.
– (2001). « La linguistique au contact de l'informatique : de la construction des grammaires à la grammaire de construction ». In : *Histoire, épistémologie, langage (H.E.L.)* 23.1, p. 49–79.
Cougnon, Louise-Amélie (2015). *Langage et SMS. Une étude internationale des pratiques actuelles*. Cahiers du Cental 8. Louvain-la-Neuve : Presses universitaires de Louvain.
Cougnon, Louise-Amélie et Cédrick Fairon, éds. (2014). *SMS Communication. A linguistic Approach*. Amsterdam/Philadelphia : John Benjamins.
Cougnon, Louise-Amélie et Thomas François (2011). « Étudier l'écrit SMS. Un objectif du projet sms4science. » In : *La communication par SMS en Suisse. Usages et variétés linguistiques. Linguistik Online, Adrian Stähli and Christa Dürscheid and Marie-José Béguelin (éds)* 48.4, p. 19–34.
Cougnon, Louise-Amélie et Gudrun Ledegen (2010). « C'est écrire comme je parle. Une étude comparatiste de variétés de français dans l'écrit sms. » In : *Les voix des Français. Modern French Identities* 2.94, p. 39–57.
Cougnon, Louise-Amélie, Lénaïs Maskens, Sophie Roekhaut et Cédrick Fairon (2016). « Social media, spontaneous writing and dictation. Spelling variation. » In : *Journal of French language studies*.
Courtois, Blandine, Mylène Garrigues, Gaston Gross, Maurice Gross, René Jung, Michel Mathieu-Colas, Anne Monceaux, Anne Poncet-Montange, Max Silberztein et Robert Vivès (1997). *Dictionnaire électronique DELAC : les mots composés binaires*. Rapport technique 56. LADL, Université Paris-7.
Crystal, David (2001). *Language and the Internet*. Cambridge : Cambridge University Press.
– (2009). *Txtng : the Gr8 Db8*. Oxford, Oxford University Press.
– (2011). *Internet Linguistics : A Student Guide*. London/New York : Routledge.
Daille, Béatrice (1994). « Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques ». Thèse de doct. Université Paris-7.
Dalle, Laurine, Joséphine Faisant, Marie Jaffal et Véronique de Martino (2013). « Transcodage de 1 001 SMS ». Dossier étudiant, Master 1, Sciences du Langage, Université Paul-Valéry Montpellier 3. Direction : Rachel Panckhurst.

Danlos, Laurence et Benoît Sagot (2007). « Comparaison du Lexique-Grammaire des verbes pleins et de DICO-VALENCE : vers une intégration dans le Lefff ». In : *Actes du colloque TALN*. Toulouse.

David, Sophie (1993a). « Les unités nominales polylexicales. Éléments de description et reconnaissance automatique. » Thèse de doct. Université Denis-Diderot Paris-7.

– (1993b). « Remarques à propos du mode de construction des unités de forme NN. » In : *TAL, Traitements automatiques de la composition nominale*. 34.2.

David, Sophie et Pierre Plante (1990). « De la nécéssité d'une approche morpho syntaxique dans l'analyse de textes. » In : *Intelligence Artificielle et Sciences Cognitives au Quebec* 3, p. 140–154.

Desclés, Jean-Pierre et Frédérique Segond (1990). « Topicalization : categorial analysis and Applicative Grammar. » In : *Actes du Colloque Ordre des mots dans le cadre des grammaires catégorielles*. Sous la dir. d'Alain Lecomte. Clermont-Ferrand, France : Clermont-Ferrand : Adosa., p. 13–37.

Détrie, Catherine (2014). « (Comment) faire public dans la sphère privée ? Ou le rôle des apostrophes dans les SMS ». In : *Actes du 5e Congrès Mondial de Linguistique Française*. Université de Lorraine, Metz.

– (2015). « Gentlemanminette d amour, ma chou, colocounette et autres formes nominales d'adresse dans les SMS : de quelques spécificités liées au genre. » In : *Actes du colloque « Interpréter selon les genres »*. Université Cadi Ayyad. Marrakech, Maroc.

– (2016a). « Être contre et ou tout contre en textotant : l'expression du consensus et du dissensus dans les SMS, entre rupture et continuum. » In : *Actes du 5e Congrès Mondial de Linguistique Française*. Sous la dir. de Franck Neveu, Gabriel Bergounioux, Marie-Hélène Côté, Jean-Michel Fournier, Linda Hriba et Sophie Prévost. url : DOI:http://dx.doi.org/10.1051/shsconf/20162702004.

– (2016b). « Produire du sens en textotant : de quelques innovations lexicales, morphosyntaxiques et sémantiques dans les SMS ». In : *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones, Actes du colloque CINEO 2015, 22-24 octobre, Salamanque*. Sous la dir. de Miguel Sánchez Ibáñez, Goedele De Sterck, Daniel Linder, Nava Maroto, Jesús Torres del Rey et

Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Joaquín García Palacios. Frankfurt, Peter Lang.

Détrie, Catherine, Paul Siblot, Bertrand Verine et Agnès Steuckardt, éds. (2017). *Termes et concepts pour l'analyse du discours. Une approche praxématique (nouvelle édition augmentée)*. Paris, Honoré Champion.

Détrie, Catherine et Bertrand Verine (2015). « Quand l'insulte se fait mot doux : la violence verbale dans les SMS ». In : *Dialogic Language use 3. Dimensions du dialogisme 3. Dialogischer Sprachgebrauch 3, Helsinki : Société néophilique*, p. 195–207.

Develotte, Christine, Richard Kern et Marie-Noëlle Lamy, éds. (2011). *Décrire la conversation en ligne. Le face à face distanciel*. Lyon, ENS Éditions.

Dik, Simon C. (1978). *Functional unification grammar*. Amsterdam : North-Holland.

Doehler, Simona Pekarek (2011). « Hallo ! Voulez vous luncher avec moi hüt ? Le "code switching" dans la communication par SMS ». In : *La communication par SMS en Suisse. Usages et variétés linguistiques. Linguistik Online, Adrian Stähli and Christa Dürscheid and Marie-José Béguelin (éds)* 48.4, p. 49–70.

Dos Santos, Nicolas (2013). « Plurilinguisme et SMS. » mémoire de Master 1, Gestion des connaissances, formations et médiations numériques, Université Paul-Valéry Montpellier 3, direction : Rachel Panckhurst.

Downes, Stephen (2008). « "Web 2.0, e-Learning 2.0 and the New Learning" ». In : *Speech, Learning Technologies Conference [online document]*. London. URL : http://www.downes.ca/cgi-bin/page.cgi?presentation=173.

Drouin, Patrick et Christian Guilbault (2016). « De Viens watcher la partie avec moi à Come regarder the game with me. » In : *Abstracts, PLIN 2016*. Louvain-la-Neuve, Belgium. URL : http://www.plindayucl.com.

Dubois, Jean et Françoise Dubois-Charlier (2013 (1997)). *Les verbes français*. version en ligne. URL : http://rali.iro.umontreal.ca/rali/?q=fr/lvf.

Dürscheid, Christa et Elisabeth Stark (2011). « SMS4science : an international corpus-based texting project and the specific challenges for multilingual Switzerland ». In : *Digital Discourse, Language in the New Media*. Sous la dir. de Crispin Thurlow et Kristine Mroczek. Oxford : Oxford University Press, p. 299–320.

Fabre, Ludivine et Manon Ravel (2011). « L'étude des conversations sms français / français et l'utilisation des Smartphones. » Dossier de M1, Université Paul-Valéry Montpellier 3.

Fairon, Cédrick, Jean René Klein et Sébastien Paumier (2006a). « Le langage SMS : révélateur d'1compétence ». In : *Le français m'a tuer. Actes du colloque" L'orthographe française à l'épreuve du supérieur"*. Sous la dir. de Presses universitaires de Louvain Louvain-la-Neuve, p. 33–42.

– (2007). « Un corpus transcrit de 30 000 SMS français ». In : *Actes du colloque CMT, La langue du cyberespace : de la diversité aux normes*. Sous la dir. de Jeannine Gerbault. L'Harmattan, p. 173–182.

Fairon, Cédrick, Jean René Klein et Sébastien Paumier Paumier (2006b). *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation (Manuel+CD-Rom)*. url : http://www.smspourlascience.be/.

Fairon, Cédrick, Jean René Klein et Sébastier Paumier (2006c). *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »*. Cahiers du Cental 1. Louvain-la-Neuve : Presses universitaires de Louvain. url : http://www.smspourlascience.be/.

Fairon, Cédrick et Sébastien Paumier (2006). « A translated corpus of 30,000 French SMS ». In : *Actes du colloque LREC*.

Franckel, Jean-Jacques (1994). « Effets sur le sens des verbes et la structuration de la relation prédicative de l'alternance sujet-prédicatif/sujet non-prédicatif ». In : *Cahiers de Praxématique* 22, p. 135–156.

Gadet, Françoise (1996). « Une distinction bien fragile : oral/écrit ». In : *Tranel (Travaux neuchâtelois de linguistique)* 25, p. 13–27.

Ganascia, Jean-Gabriel (1996). *Les sciences cognitives*. Paris : Flammarion.

Gardent, Claire, Bruno Guillaume, Guy Perrier et Ingrid Falk. (2006). « Extraction d'information de sous-catégorisation à partir des tables du LADL. » In : *Actes du colloque TALN*. Leuven, Belgique. url : https://hal.inria.fr/inria-00103163.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum et Ivan Andrew Sag (1985). *Generalized Phrase Structure Grammar*. Cambridge, MA : Harvard University Press.

Gazdar, Gerald et Chris Mellish (1989). *Natural Language Processing in Prolog. An Introduction ti Computational Linguistics*. Wokingham, Addison-Wesley.

Gerbault, Jeannine, éd. (2007). *La langue du cyberespace : de la diversité aux normes*. Paris : L'Harmattan.

Ghliss, Yosra (2014). « Quelques formes de l'inscription des émotions dans les SMS », Master 2, Sciences du Langage, Discours médiatiques, institutionnels et politiques, Université Paul-Valéry Montpellier 3, directeur : Bertrand Verine.

Ghliss, Yosra et Frédéric André (2017). « Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88milSMS. » In : *Corpus de communication médiée par les réseaux. Construction, structuration, analyse*. Sous la dir. de Ciara R. Wigham et Gudrun Ledegen. Humanités numériques. Paris : L'Harmattan, p. 71–84.

Ghliss, Yosra et Bertrand Verine (2016). « Je t'aime fortttttttttt : la répétition graphémique, marqueur d'émotion dans le genre du discours SMS ? » In : *Les Émotions et les valeurs dans la communication*. Sous la dir. de Katarzyna Wolowska et Anna Krzyzanowska. Francfort-sur-le-Main, Peter Lang.

Goumi, Antonine et Josie Bernicot (2011). « Un corpus de SMS produits par de jeunes adolescents : méthode de recueil et premières données ». In : *Conférence invitée, projet sud4science LR*. MSH-M, Montpellier.

Gréciano, Gertrud (1982). « Signification et dénotation en allemand. La sémantique des expressions idiomatiques. » Thèse de doct. Thèse d'Etat. Paris-Sorbonne. Klincksieck.

Gross, Maurice (1975). *Méthodes en syntaxe : régime des constructions complétives*. Paris : Hermann.

Guilbault, Christian et Patrick Drouin (2016). « Pratiques liées aux alternances de code dans un corpus anglais et français au Canada ». In : *Conférence invitée, Cercle linguistique Belge*. Louvain-la-Neuve, Belgique.

Guimier de Neef, Émilie et Sébastien Fessard (2007). « Évaluation d'un système de transcription de SMS ». In : *Proceedings of the 26th International Conference on Lexis and Grammar*. Bonifacio, France.

Guimier de Neef, Émilie et Jean Véronis (2004). « Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.) » In : *Journée d'Etude. Association pour le Traitement Automatique des Langues (ATALA)*. ENST, Paris.

Habert, Benoît, éd. (2004). « *Linguistique et informatique : nouveau défis », Revue française de linguistique appliquée*. T. IX. 1. http ://www.rfla-journal.org/. url :

https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2004-1.htm.

Habert, Benoît et Christian Jacquemin (1993). « Noms composés, termes, dénominations complexes : problématiques et traitements automatiques. » In : *TAL, Traitements automatiques de la composition nominale.* 34.2, p. 5–42.

Habert, Benoît, Adeline Nazarenko et André Salem (1997). *Les linguistiques de corpus*. Paris : Armand Colin, Masson.

Hak, Tony et Niels Helsloot, éds. (1995). *Michel Pêcheux. Automatic Discourse Analysis*. Utrecht Studies in Language et Communication : Amsterdam-Atlanta, Éditions Rodopi.

Halliday, Michael A. K. (1989). *Spoken and written language*. Oxford, Oxford University Press.

Herring, Susan C., éd. (1996). *Computer-mediated communication*. Pragmatics and Beyond. Amsterdam/Philadelphia, John Benjamins.

– (1997). « Computer-mediated discourse analysis : Introduction. » In : *Electronic Journal of Communication* 6.3. url : http://www.cios.org/www/ejc/v6n396.htm.

– (2001). *Computer-mediated Discourse*. Sous la dir. de Deborah Schiffrin, Deborah Tannen et Heidi E. Hamilton. Oxford, Blackwell.

Herring, Susan C., Dieter Stein et Tuija VirtanenStein, éds. (2013). *Handbook of Pragmatics of Computer-Mediated Communication*. Handbooks of Pragmatics. Berlin, De Gruyter Mouton.

Hulme, Keri (1983). *The Bone People*. London, Picador.

Huston, Nancy (1999). *Nord Perdu*. Littérature. Actes Sud.

Ide, Nancy et James Pustejovsky, éds. (2017). *Handbook of Linguistic Annotation*. Berlin : Springer.

Joshi, Aravind K., Leon S. Levy et Masako Takahashi (1975). « Tree Adjunct Grammars ». In : *Journal of Computer and Systems Sciences* 10.1, p. 55–75.

Kaba, Aminata (2014). « La reconnaissance vocale sur mobile : lécritur d sms, 2 lécri a Loral », mémoire de Master 1, Gestion des connaissances, formations et médiations numériques, Université Paul-Valéry Montpellier 3, directrice : Rachel Panckhurst.

Kaplan, Ronald M. et Joan Bresnan (1982). « Lexical-functional grammar : A formal system for grammatical representation. » In : *The Mental Representation*

of Grammatical Relations. Sous la dir. de Joan Bresnan. Cambridge : MIT Press, p. 173–281.

Kay, Martin (1979). « Functional Grammar. » In : *Proceedings of the 5th meeting of the Berkeley Linguistics Society*.

Khiari, Wejdene, Asma Bouhafs et Mathieu Roche (2016a). « Comment prendre en compte les spécificités de 'l'écriture SMS' pour l'analyse de sentiments ? » In : *Proceedings of Journées internationales d'analyse statistique des données textuelles (JADT)*. Nice. url : https://jadt2016.sciencesconf.org/108031/document.

– (2016b). « Integration of Lexical and Semantic Knowledge for Sentiment Analysis in SMS ». In : *Proceedings of Language Resources and Evaluation (LREC)*. Portoroz, Slovenia, p. 1185–1189. url : www.lrec-conf.org/proceedings/lrec2016/pdf/610_Paper.pdf.

Khiari, Wejdene, Mathieu Roche et Asma Bouhafs (2016c). « Intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS ». In : *Actes de la conférence Extraction et Gestion des Connaissances (EGC), Affiche*. Reims, p. 553–554. url : http://editions-rnti.fr/?inprocid=1002227.

Kobus, Catherine, François Yvon et Géraldine Damnati (2008). « Transcrire les SMS comme on reconnaît la parole ». In : *Proceedings, TALN 2008*. Avignon, p. 128–138. url : https://perso.limsi.fr/yvon/publications/sources/Kobus08transcrire.pdf.

Kogkitsidou, Eleni et Georges Antoniadis (2016). « L'architecture d'un modèle hybride pour la normalisation de SMS. » In : *Actes du 23e colloque Traitement automatique des langues naturelles*. Inalco, Paris, p. 355–363. url : https://jep-taln2016.limsi.fr/actes/index.php?lang=fr.

Krstev, Cvetana, Biljana Lazić, Ranka Stanković, Giovanni Schiuma et Miladin Kotorčević (2015). « Development of Open Educational Resources (OER) for Natural Language Processing ». In : *Proceedings of the Sixth International Conference on eLearning (eLearning 2015)*. Belgrade, Serbia.

Krstev, Cvetana, Anđelka Zečević, Duško Vitas et Tita Kyriacopoulou (2016). « NERosetta for the Named Entity Multi-lingual Space ». In : Springer International Publishing, p. 327–340. url : DOI:10.1007/978-3-319-43808-5_25.

Kyriacopoulou, Panayota (1989). « Les dictionnaires electroniques et la flexion verbale en grec moderne ». Thèse de doct. Université Paris 8.

Kyriacopoulou, Panayota (2010). « Lexique-Grammaire des verbes en grec moderne : bilan et perspectives ». In : *Les tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*. Sous la dir. de Takuya Nakamura. Cahiers du Cental 6. Presses Universitaires de Louvain, p. 181–191.

Kyriacopoulou, Panayota et Claude Martineau (2015). « Extraction de "segments complexes" : enrichissement des dictionnaires ». In : *Études de linguistique appliquée, « Linguistique et informatique », Pierre-André Buvet (coor.)* P. 407–416.

Laboureau, Jérémy (2014). « Caractéristiques propres au scripteur et au Smartphone impactant l'écriture SMS », mémoire de Master 1, Sciences du Langage, Discours médiatiques, institutionnels et politiques, Université Paul-Valéry Montpellier 3, directrice : Rachel Panckhurst.

Lamrani, Akila (2006). « Accessibilité scolaire et pédagogique pour les élèves déficients auditifs en collège », Mémoire professionnel pour le Certificat Complémentaire pour les Enseignements Adaptés et la Scolarisation des élèves en situation de handicap (2 CA-SH), Montpellier.

Langlais, Philippe, Patrick Drouin, Amélie Paulus, Eugénie Rompré Brodeur et Florent Cottin (2012). « Texto4Science : a Quebec French Database of Annotated Short Text Messages ». In : *Proceedings, LREC*. Istanbul, p. 1047–54.

Laporte, Éric (2010). « Le lexique-grammaire est-il exploitable pour le traitement des langues ? » In : *Les tables. La grammaire du francais par le menu. Mélanges en hommage à Christian Leclère.* Sous la dir. de Takuya Nakamura, Éric Laporte, Anne Dister et Cédrick Fairon. Cahiers du Cental. Louvain-la-Neuve : Presses universitaires de Louvain, p. 207–218.

– (2015). « The Science of Linguistics ». In : *Inference : International Review of Science* 1.2. url : http://inference-review.com/article/the-science-of-linguistics.

Laporte, Éric, Elsa Tolone et Matthieu Constant (2013). « Conversion of Lexicon-Grammar tables to LMF. Application to French. » In : *LMF. Lexical Markup Framework.* Sous la dir. de Gil Francopoulo. ISTE - Wiley, p. 57–187. url : https://hal-upec-upem.archives-ouvertes.fr/hal-00803800.

Lazar, Jan (2012). « Les anglicismes dans le discours électronique médié ». In : *Studia Romanica Posnaniensia Uam* 39.4.

– (2013). « L'équivalence terminologique dans le discours électronique médié ». In : *Roczniki Humanistyczne* LXI.8.
– (2014). « À propos des connecteurs textuels dans le discours électronique médié ». In : *Studi@ Naukowe* 23.
– (2017). *À propos des pratiques scripturales dans l'espace virtuel : entre Facebook et Twitter*. Ostrava : Faculté des Lettres de l'université d'Ostrava.

Le Marec, Joëlle (2004). « Les études d'usage ». In : *Evaluation des systèmes de traitement de l'information*. Sous la dir. de Stéphane Chaudiron. Paris : Hermès.

Lecomte, Alain, éd. (1990). *Actes du Colloque Ordre des mots dans le cadre des grammaires catégorielles*. Clermont-Ferrand, France : Clermont-Ferrand : Adosa.

Ledegen, Gudrun (2014). *Grand corpus de sms SMS La Réunion [corpus]. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy.* url : http://hdl.handle.net/11403/comere/cmr-smslareunion.

Lénaïs Maskens Louise-Amélie Cougnon, Sophie Roekhaut et Cédrick Fairon (2015). « Nouveaux médias et orthographe. Incompétence ou pluri-compétence ? » In : *Discours* 16.

Léon, Jacqueline (2010). « AAD69 : archéologie d'une étrange machine ». In : *Semen (en ligne)*. url : http://semen.revues.org/8823.
– (2015). *Histoire de l'automatisation des sciences du langage*. Langages. Lyon : ENS Éditions.

Liénard, Fabien (2005). « Langage texto et langage contrôlé. Descriptions et problèmes ». In : *Lingvisticæ Investigationes* XXVIII.1, p. 49–60.
– (2007). « Analyse linguistique et sociopragmatique de l'écriture électronique. Le cas du SMS tchaté ». In : *Actes du colloque CMT, La langue du cyberespace : de la diversité aux normes*. Sous la dir. de Jeannine Gerbault, p. 265–278.
– (2012). « TIC, Communication électronique écrite, communautés virtuelles et école ». In : *Études de linguistique appliquée, Les connaissances cachées développées par la lecture et l'écriture électronique extrascolaires* 166. Sous la dir. de Marie-Laure Elalouf, p. 143–155.

Ligia-Stela, Florea et Catherine Fuchs (2013). *Dictionnaire des verbes du français actuel. Constructions, emplois, synonymes*. Paris, Ophrys.

Luong, Clothilde (2015). « L'inscription de la complicité dans les SMS », Mémoire de Master 1, Gestion des connaissances, formations et médiations numériques, Université Paul-Valéry Montpellier 3, directrice : Rachel Panckhurst.

Maldider, Denis (1990). *L'inquiétude du discours, textes de Michel Pêcheux, choisis et présentés par Denise Maldidier.* Paris : Éditions des Cendres.

Marandin, Jean-Marie (1990). « Le lexique mis à nu par ses célibataires. Stéréotype et théorie du lexique. » In : *La Définition*. Sous la dir. de Jacques Chauraud et Francine Mazière. Paris : Larousse, Langue et Langage., p. 284–291.

– (1992). « La perception syntaxique. » In : *Le Gré des langues* 3, p. 64–91.

– (1993). « Les analyseurs syntaxiques. Équivoques et problèmes. » In : *TAL* 34.1, p. 125–153.

– (1997). « Dans le titre se trouve le sujet. L'inversion locative en français. » Mémoire d'habilitation à diriger des recherches, université Paris VII Denis-Diderot.

Marandin, Jean-Marie et Michel Pêcheux (1984). « Informatique et analyse du discours ». In : *Buscila* 1, p. 64–65.

Marcoccia, Michel (2000). « Communication et Organisation. La représentation du non-verbal dans la communication écrite médiatisée par ordinateur. » In : 18, p. 265–274.

– (2011). « "T'es où maintenant ?", les espaces de la conversation visiophoniques en ligne. » In : *Décrire la conversation en ligne*. Sous la dir. de Christine Develotte, Richard Kern et Marie-Noëlle Lamy. Lyon, ENS Éditions, p. 95–116.

– (2012). « Conversationnalisation et contextualisation : deux phénomènes pour décrire l'écriture numérique ». In : *Le Français dans le monde, Recherches et applications* 51, p. 92–105.

– (2016). *Analyser la communication numérique écrite*. Paris : Armand Colin.

Marty, Nicole (2005). *Informatique et nouvelles pratiques d'écriture.* Paris : Nathan, Repères pédagogiques.

Mazière, Francine (2015 (3e édition)). *L'analyse du discours*. Que Sais-je ? Paris : PUF.

McEnery, Tony et Andrew Hardie (2012). *Corpus Linguistics : Method, theory and practice*. Cambridge : Cambridge University Press.

Mela, Augusta (2004). « Linguistes et "talistes" peuvent coopérer : repérage et analyse des gloses. » In : *Revue Française de Linguistique Appliquée, « Linguistique et informatique : nouveaux défis », coor. Benoît Habert* 9.1.

– (2005). « Le repérage automatique des gloses de nomination seconde. » In : *Les marqueurs de la glose*. Sous la dir. d'Agnès Steuckardt et Aïno Niklas-Salminen. Publications de l'université de Provence.

Mela, Augusta, Mathieu Roche et Mohamed el Amine Bekhtaoui (2011). « Mixer les moyens pour extraire les gloses. » In : *Actes de la conférence Extraction et Gestion des Connaissances (EGC)*, p. 95–106.

Milner, Jean-Claude (1978). *De la syntaxe à l'interprétation : quantités, insultes, exclamations*. Paris, Seuil.

– (1989). *Introduction à une science du langage*. Paris : Des Travaux/Seuil.

Moïse, Claudine (2012). « Action et construction de la science. Dominique Caubet ou la théorie de la pelote ». In : *Dynamiques langagières en Arabophonies : variations, contacts, migrations et créations artistiques. Hommage offert à Dominique Caubet par ses élèves et collègues*. Sous la dir. d'Alexandrine Barontini, Christophe Pereira, Ángeles Vicente et Karima Ziamari. Colección Estudios de Dialectología Árabe 7. Zaragoza. : Zaragoza : Universidad de Zaragoza, Área de Estudios Árabes e Islámicos, p. 321–336.

– (2013a). « Lol non tkt on ta pas oublié. Rapports à la norme et valeurs de la faute dans l'écriture Sms (projet et corpus Sud4science). Réflexions sociolinguistiques ». In : *conférence plénière, Colloquium Si j'aurais su, j'aurais pas venu ! Linguistique des formes exclues : description, genre, épistémologie*. Université Libre de Bruxelles.

– (2013b). « Wesh trkl tkt ;) tu fou quoi ? La question de la norme dans l'écriture Sms : de la "faute" à l'indignation normative ». In : *Conférence invitée présenté à La circulation des discours*. Université Laval Québec.

Moïse, Claudine et Christina Schultz-Romain (2010). « Violence verbale et listes de discussions : les argumentations polémiques. » In : *Cahiers de l'institut de linguistique de Louvain, Du « terrain » à la relation : expériences de l'internet et questionnaires méthodologiques, Isabelle Pierozak (éd.)* 2.36, p. 113–133.

Morel, Étienne (2016). « Le bricolage plurilingue dans la communication par texto. Interprétations d'une pratique entre affiliation locale et aspiration

globale. » Thèse de doct. Institut des sciences du langage, Centre de linguistique appliquée, Université de Neuchâtel.

Mourlhon-Dallies, Florence et Jean-Yves Colin (1999). « Des didascalies sur l'internet ? » In : *Internet, communication et langue française*. Sous la dir. de Jacques Anis. Paris : Hermès, p. 13–30.

Moussa, Aïché (2014). « Camfranglais : Pratiques et usages plurilingues dans les SMS au Cameroun », mémoire de Master 1 Gestion des connaissances, formations et médiations numériques, Université Paul-Valéry Montpellier 3, directrice : Rachel Panckhurst.

Nakamura, Takuya, Éric Laporte, Anne Dister et Cédrick Fairon, éds. (2010). *Les tables. La grammaire du francais par le menu. Mélanges en hommage à Christian Leclère*. Cahiers du Cental. Louvain-la-Neuve : Presses universitaires de Louvain.

Orlando, Elisabetta (2012). « L'expression des émotions et des sentiments dans Twitter et les SMS : analyse comparée des usages, des formes et des objectifs », mémoire de Master 1, Gestion des connaissances, formations et médiations numériques, Université Paul-Valéry Montpellier 3, directrice : Rachel Panckhurst.

Ouellet, Pierre, Adel El-Zaïm et Hervé Bouchard (1994). « La représentation des actes de perception : le cas de paraître. » In : *Cahiers de Praxématique* 22, p. 135–156.

Parret, H. (1974). *Discussing Language : Dialogues with W. Chafe, N. Chomsky, A. J. Greimas, M.A.K. Halliday, P. Hartmann, G. Lakoff, S. M. Lamb, A. Martinet, J. Maccawley, S. K. Saumjan, J. Bouveresse*. De Gruyter Mouton.

Patel, Namrata, Pierre Accorsi, Diana Inkpen, Cédric Lopez et Mathieu Roche (2013). « Approaches of anonymisation of an SMS corpus ». In : *Proceedings, of CICLING (Conference on Intelligent Text Processing and Computational Linguistics), LNCS*. Sous la dir. de Springer Verlag. University of the Aegean, Samos, Greece, p. 77–88.

Paumier, Sébastien (2003). « De la reconnaissance des formes linguistiques à l'analyse syntaxique ». Thèse de doct. en Informatique linguistique, sous la direction de Maurice Gross et d'Éric Laporte, Université Paris-Est Marne-la-Vallée.

Paumier, Sébastien et Claude Martineau (2016). *Unitex 3.1. Manuel d'utilisation*. Université Paris-Est Marne-la-Vallée.

Paveau, Marie-Anne (2013). « Technodiscursivités natives sur Twitter : Une écologie du discours numérique ». In : *Revue internationale de sciences humaines et sociales appliquées, Epistémè* 9. Sous la dir. de F. Liénard, p. 139–176.

Pêcheux, Michel, Simone Bonnafous, Jacqueline Léon et Jean-Marie Marandin (1982). « Présentation de l'analyse automatique du discours (AAD69) : théorie, procédures, résultats, perspectives ». In : *Mots* 4, p. 95–124.

Péry-Woodley, Marie-Paule (1995). « Quels corpus pour quels traitements automatiques ? » In : *TAL* 36, p. 213–232.

Pierozak, Isabelle, éd. (2007). *« Regards sur l'internet, dans ses dimensions langagières. Penser les continuités et discontinuités. En hommage à Jacques Anis »*, *Glottopol*. 10. Université de Rouen.

Piolat, Annie, éd. (2005). *Lire, écrire, communiquer et apprendre avec Internet*. Marseille : Éditions Solal.

Plante, Pierre (1979). « Le DEREDEC, logiciel pour le traitement linguistique et l'analyse du contenu des textes ». Thèse de doct. Université du Québec à Trois-Rivières.

– (1990). « L'autonomie conceptuelle des faisceaux ». In : *ICO, Québec* 2.6.
– (1991). « La modélisation en faisceaux pour le parsage syntaxique. » In : *Actes du colloque ILN*. Nantes.
– (1993). « La localisation procédurale dans le langage FX ». In : *Actes du colloque ICO*. Montréal, p. 184–193.
– (1996). *L'Atelier FX. Version 7. L'environnement de programmation*. RDLC, Centre d'ATO, Université du Québec à Montréal.

Poirier, Damien (2011). « Des textes communautaires à la recommandation ». Thèse de doct. d'informatique à l'université d'Orléans, sous la direction d'Isabelle Tellier et de Patrick Gallinari.

Pollard, Carl et Ivan Andrew Sag (1994). *Head-Driven Phrase Structure Grammar*. CSLI Series. Chicago : University of Chicago Press.

Porto, Giulia (2014). « Écriture SMS et Néographie chez les jeunes de 11-30 ans », mémoire de Master 2, Sciences du Langage, Discours médiatiques, institutionnels et politiques, Université Paul-Valéry Montpellier 3, directrice : Rachel Panckhurst.

Pozzi, Federico Alberto, Elisabetta Fersini, Enza Messina et Bing Liu, éds. (2016). *Sentiment Analysis in Social Networks*. Elsevier.

Rastall, Paul R. (1979). « L'empirisme en linguistique ». In : *La linguistique* 15.2, p. 107–120.

Riou, Stéphane et Benoît Sagot (2016). *Étiquetage morpho-syntaxique du corpus FAVI [corpus]. D'après Hyeon Yun and Thierry Chanier (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]. Banque de corpus CoMeRe. Ortolang.fr : Nancy.* url : https://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2.

Roche, Mathieu (2004). « Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes ». Thèse de doct. Université Paris 11.

– (2011). « Fouille de Textes : De l'extraction des descripteurs linguistiques à leur induction », Habilitation à diriger des recherches, Université de Montpellier.

Sabah, Gérard, éd. (2006). *Compréhension automatique des langues et interaction*. Paris : Hermès Sciences.

Sagot, Benoît (2010). « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French ». In : *Proceedings, LREC, 2010*. Valletta, Malta. url : http://hal.inria.fr/inria-00521242/;~Project~Alexina:https://gforge.inria.fr/projects/alexina/..

Salmon, Gilly (2000). *E-moderating : the key to teaching and learning online*. London/Sterling : Kogan Page.

Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova et Federico Sangati (2015). « PARSEME — PARSing and Multiword Expressions within a European multilingual network. » In : *Proceedings of the 7th Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poznan, Poland.

Schiffrin, Deborah, Deborah Tannen et Heidi E. Hamilton, éds. (2001). *The Handbook of Discourse Analysis*. Oxford, Blackwell.

Segond, Frédérique (1990). « Grammaire catégorielle du français. Etude théorique et implémentation. Le système GraCE (Grammaire Catégorielle Eten-

due) ». Thèse de doct. École des Hautes Etudes en Sciences Sociales, Paris, France, 1990.
– (1994). « Traitement des verbes dans une grammaire catégorielle du français : le coup de GraCE ». In : *Cahiers de Praxématique* 22, p. 13–36. URL : https://praxematique.revues.org/1892.
– éd. (2002). *Multilinguisme et traitement de l'information*. Paris : Hermès Sciences.
SEGOND, Frédérique et Annie ZAENEN (1995). « Cherche Linguiste-informaticien désespérément. » In : *TAL* 36.
SMADJA, Frank (1993). « Retrieving collocations from text : Xtract ». In : *Computational Linguistics* 19.1, p. 143–177.
SOUCHARD, Maryse, Stéphane WAHNICH, Isabelle CUMINAL et Virginie WATHIER (1997). *Le Pen. Les Mots. Analyse d'un discours d'extrême-droite*. Paris : Le Monde éditions.
STABILE, Teresa et Elisabetta TORTORELLA (2008). « Le langage de SMS : entre italien et français », Dossier de Master, Université Paul-Valéry Montpellier 3, direction : Rachel Panckhurst.
STARK, Elisabeth (2011). « La morphosyntaxe dans les SMS suisse francophones : Le marquage de l'accord sujet – verbe conjugué. » In : *La communication par SMS en Suisse. Usages et variétés linguistiques. Linguistik Online, Adrian Stähli and Christa Dürscheid and Marie-José Béguelin (éds)* 48.4, p. 35–47.
– (2014). « Negation marking in French text messages. » In : *SMS Communication. A Linguistic Approach*. Sous la dir. de Louise-Amélie COUGNON et Cédrick FAIRON. Amsterdam/Philadelphia : John Benjamins, p. 191–216.
TAGG, Caroline (2012). *The Discourse of Text Messaging : Analysis of SMS communication*. London/New York, Continuum.
THURLOW, Crispin et Kristine MROCZEK, éds. (2011). *Digital Discourse, Language in the New Media*. Oxford : Oxford University Press.
TOLONE, Elsa et Benoît SAGOT (2011). « Using Lexical-Grammar Tables for French Verbs in a Large-Coverage Parser ». In : *Proceedings, 4th Language and Technology Conference, LTC, Human Language Technology. Challenges for Computer Science and Linguistics : Lecture Notes in Computer Science*. Sous la dir. de Zygmunt VETULANI.
TROUILLEUX, François (2015). « Gabriel G. Bès, linguiste empiriste ». In : *TAL* 56.1, p. 13–37.

Verine, Bertrand (2013). « Les verbes SMS, texto, texter et textoter dans le corpus sud4science. » url : https://praxiling.hypotheses.org/348.

– (2015). « C pa 1 sms, c 1 roman !! : le SMS est-il interprété comme un genre par ses usagers ? » In : *Actes du colloque « Interpréter selon les genres »*. Université Cadi Ayyad. Marrakech, Maroc. url : https://halshs.archives-ouvertes.fr/hal-01317800/document.

Véronis, Jean (1988). « Contribution à l'étude de l'erreur dans le dialogue homme-machine. » Thèse de doct. Université Aix-Marseille III.

Véronis, Jean et Émilie Guimier de Neef (2006). « Le traitement des nouvelles formes de communication écrite ». In : *Compréhension automatique des langues et interaction*. Sous la dir. de Gérard Sabah. Paris : Hermès Sciences, p. 227–248.

Vilariño, Darnes, David Pinto, Beatriz Beltrán, Saul León, Estaban Castillo et Mireya Tovar (2012). « Pattern Recognition ». In : Berlin/Heidelberg : Springer. Chap. A machine-translation method for normalization of SMS, p. 293–302. url : http://www.cs.buap.mx/~dpinto/research/MCPR2012/MCPR2012_Vilarino.pdf.

Vincent-Durroux, Laurence (2014). *La langue orale des jeunes sourds profonds*. Voix, Parole, Langage. Paris, Éditions de Boeck-Solal.

Vincent-Durroux, Laurence et Cécile Poussard (2014). « "Conception et utilisation d'un logiciel pédagogique, l'exemple de Macao" ». In : *Alsic (Apprentissage des Langues et Systèmes d'Information et de la Communication)* 17.1. url : http://alsic.revues.org/2698.

Volckaert-Legrier, Olga, Josie Bernicot et Alain Bert-Erboul (2009). « Electronic mail, a new written-language register : A study with French-speaking adolescents. » In : *British Journal of Psychology* 27.1, p. 163–181.

Vold Lexander, Kristin (2010). « Pratiques plurilingues de l'écrit électronique : alternances codiques et choix de langue dans les SMS, les courriels et les conversations de la messagerie instantanée des étudiants de Dakar, Sénégal. » Thèse de doct. University of Oslo, Faculty of Humanities.

Vouillon, Betty (2014). « J'ai hâte en tabernak, dude. of course, mec ! De quelques manières d'apostropher dans les SMS (projet SMS4science), une approche comparative des termes d'adresse français et québécois », Master

2, Sciences du Langage Discours médiatiques, institutionnels et politiques, Université Paul-Valéry Montpellier 3, directrice : Catherine Détrie.

WIGHAM, Ciara R. et Gudrun LEDEGEN, éds. (2017). *Corpus de communication médiée par les réseaux. Construction, structuration, analyse*. Humanités numériques. Paris : L'Harmattan.

WOOD, Claire, Nenagh KEMP et Beverly PLESTER (2013). *Text messaging and literacy : the evidence*. London, Routledge.

Glossaire

88milSMS corpus de plus de 88 000 SMS authentiques en français, http://88milsms.huma-num.fr/ (v1), https://hdl.handle.net/11403/comere/cmr-88milsms (v2). 141, 168, 177, 187, 188, 190, 196, 198, 205, 206, 208, 209, 211, 213, 216, 218, 219, 223, 224, 225, 226, 228, 230, 232, 233, 235, 238, 240, 242, 244, 249, 252, 253, 264, 268, 269, 270, 276, 277, 280

C2i compétences numériques. 16, 17, 18, 19, 21, 26, 92, 161, 164

CBD consultation de bases de données. 6, 7, 91

CIL correspondant informatique et libertés. 188, 189, 191, 252, 253, 277

CMC computer-mediated communication. 143, 144, 149, 275

CMM communication, médias, médiations numériques. 17, 18, 19

CMO communication médiée par ordinateur. 13, 18, 19, 20, 38, 50, 81, 141, 145, 146, 148, 149, 150, 151, 152, 154, 164, 244, 275

CoMeRe communication médiée par les réseaux. 218, 240, 252

COMUE communauté d'universités et établissements. V, 11, 18, 31, 98

CRCT congé pour recherches ou conversions thématiques. 45, 104, 131

DAJI Direction des Affaires Juridiques et Institutionnelles. 188, 189, 253

DAL dispositif d'assignation lexicale. 57, 91

DEM discours électronique médié. 13, 20, 38, 50, 81, 135, 136, 141, 142, 146, 148, 149, 150, 152, 153, 154, 155, 156, 158, 160, 161, 163, 244, 275

DGLFLF Délégation générale à la langue française et aux langues de France. 141, 145, 188, 190, 206, 210, 211, 251, 264, 265, 276, 278

DNM discours numérique médié. 13, 20, 38, 50, 81, 141, 142, 152, 153, 154, 155, 156, 158, 160, 161, 163, 235, 244, 275

EASA European Academic Software Award. 41, 42, 91, 101, 102, 103, 104, 105, 106, 107, 109, 110, 111, 113, 114, 115, 116, 117, 132, 135

EKMA European Knowledge Media Association. 101, 102, 101, 103, 102, 104, 105, 109, 111, 113, 114, 115, 116

ELEN eLearning exchange network. 121, 123, 125, 126, 127, 128, 129, 132, 135, 136

FOAD formation ouverte et à distance. 13, 18, 19, 38, 50, 91, 100, 101, 103, 105, 115, 117, 119, 131, 135, 142, 160, 235

Huma-Num très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales, http://www.huma-num.fr/. 219, 242

IDL industries de la langue. 17, 19, 26

INS items non standard. 225, 226, 228

INSO items non standard originaux. 225, 227, 228

LSF langues des signes française. 32, 200

MDD mediated digital discourse. 275

MED mediated electronic discourse. 161, 275

METICE multimédia, enseignement, technologies de l'information et de la communication éducatives. 29, 30, 31, 42, 43, 47, 80, 92, 98, 100, 101, 134, 276

MIAP mathématiques et informatique appliquées. 16, 17, 18, 19, 21, 26

MIASHS mathématiques et informatique pour les sciences humaines et sociales. 16

MSH-M Maison des Sciences de l'Homme de Montpellier. 43, 46, 175, 187, 188, 216, 249, 251, 252, 265

Ortolang outils et ressources pour un traitement optimisé de la langue, https://www.ortolang.fr/. 187, 219, 240, 242, 249, 253

PAO publication assistée par ordinateur. 18, 19, 21, 104

RISIF répertoire informatisé des structures interrogatives du français. 6, 7, 91

RÉEL réseau d'échanges pédagogiques en eLearning. 121, 123, 125, 126, 127, 128, 129, 132, 135, 136

Scatlex dispositif automatisé de classification lexicale pour la sous-catégorisation verbale. 39, 57, 58, 65, 67, 73, 80, 83, 85, 87, 91

SCV sous-catégorisation verbale. 60, 61, 65, 69, 72, 73, 74

SL sciences du langage. 18, 19, 27

sud4science projet de recherche visant à la collecte et à l'analyse d'un grand nombre de SMS, mené à Montpellier http://sud4science.org/. 43, 46, 47, 141, 165, 166, 168, 167, 175, 176, 186, 188, 191, 204, 212, 216, 230, 241, 249, 251, 252, 264, 276, 280

TAL traitement automatique du langage/des langues. 12, 16, 17, 18, 19, 20, 23, 24, 25, 26, 27, 31, 32, 73, 74, 77, 82, 83, 87, 91, 96, 97, 101, 109, 131, 135, 141, 142, 157, 161, 163, 164, 184, 186, 190, 206, 219, 223, 231, 233, 235, 236, 238, 265

TALN traitement automatique du langage naturel. 24, 26, 78, 87, 191, 249, 278

TALNE traitement automatique du langage naturel écrit. 12, 76, 142, 223

TICE technologies de l'information et de la communication éducatives. 13, 18, 19, 21, 28, 31, 38, 130, 131, 135, 142, 143, 235

UVPL unités verbales polylexicales. 57, 60, 61, 71, 75, 76, 77, 78, 79, 80, 87, 89, 91, 97, 168

Table des matières