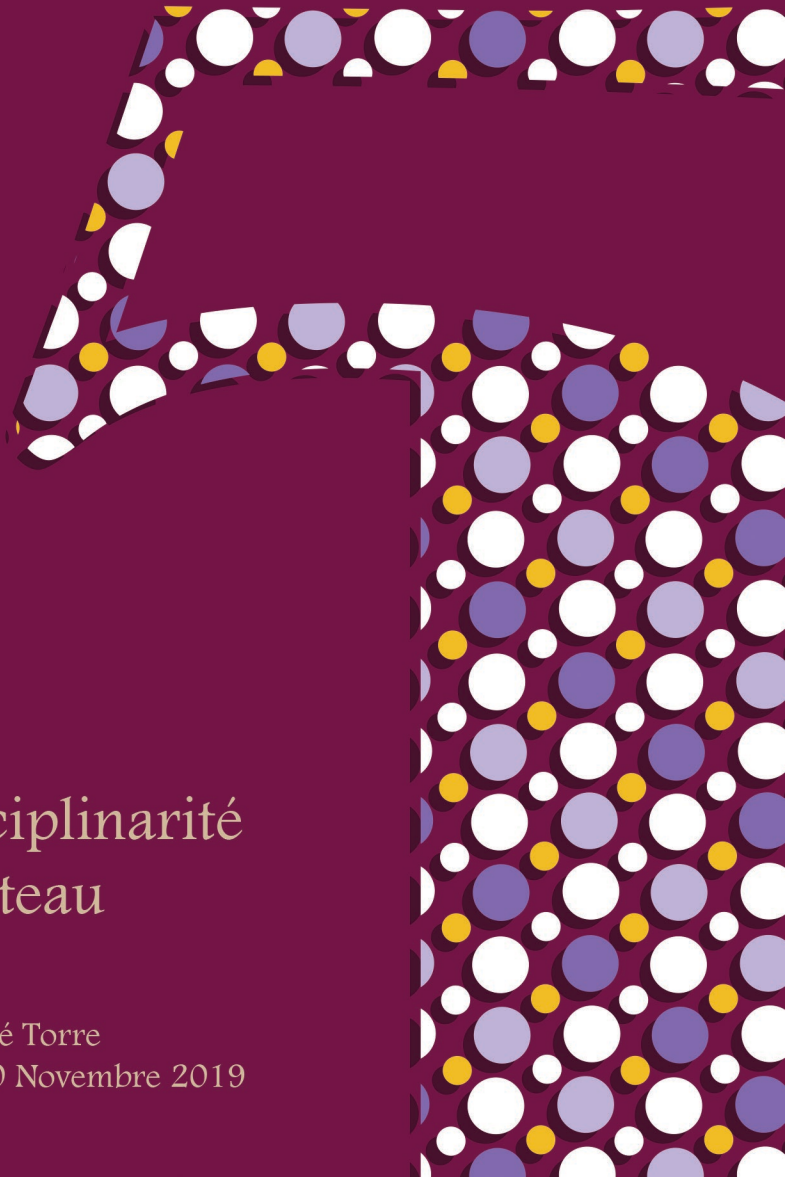
Maison des Sciences de l'Homme PARIS-SACLAY

MSH PARIS-SACLAY



5 ANS
d'interdisciplinarité sur un Plateau

DIRECTION : André Torre
COLLOQUE du 20 Novembre 2019

ÉDITION
André Torre
Directeur de la MSH Paris-Saclay

COORDINATION
Éric Valdenaire
Chargé de communication, MSH Paris-Saclay

SECRÉTARIAT DE RÉDACTION
Anne-Sophie Décriaud
Éditrice, MSH Paris-Saclay

ILLUSTRATIONS ET MAQUETTE
Léa Avril
Graphiste, MSH Paris-Saclay

ENTRETIENS
Propos recueillis par Sylvain Allemand
Journaliste, rédacteur en chef de *Paris-Saclay Le Média*

Maison des Sciences de l'Homme
PARIS-SACLAY

MSH PARIS-SACLAY

5 ANS D'INTERDISCIPLINARITÉ SUR UN PLATEAU

MSH PARIS-SACLAY
ÉDITIONS

61 avenue du Président Wilson, 94230 Cachan
www.msh-paris-saclay.fr

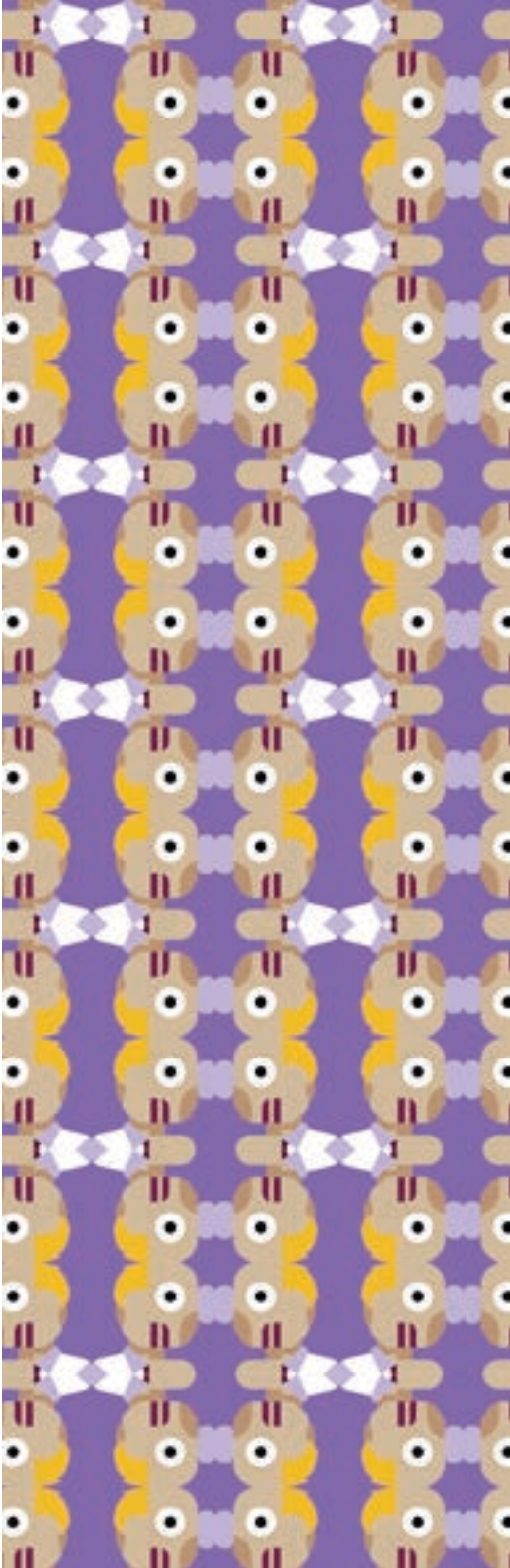
ISBN 978-2-490369-04-1

IOANA VASILESCU

La linguistique à l'heure de l'IA et des systèmes automatiques

Chargée de recherche CNRS en linguistique, Ioana Vasilescu travaille au Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Limsi, un laboratoire CNRS, associé à l'Université Paris-Sud), sur le traitement automatique de la communication parlée.

En 2017, elle a été lauréate d'un appel à projets Émergence avec le projet « HistorIA », qui se propose d'explorer les changements sonores avec l'intelligence artificielle. La même année, ce projet faisait l'objet d'un colloque sur le thème « Linguistique et *Big Data* ». L'année suivante, elle était lauréate d'un appel à projets Maturation pour le projet « HistorIA.2 ». Destiné à approfondir l'étude de l'évolution des langues avec les moyens de l'intelligence artificielle, il a notamment contribué à la réalisation et la mise en ligne d'un Atlas sonore des langues régionales.



Ioana VASILESCU
Chargée de recherche CNRS au Limsi (UPR 3251)

– Pour commencer, pouvez-vous rappeler comment vous en êtes venue à ce domaine de recherche au croisement de plusieurs champs d'études, la linguistique et le numérique ?

J'ai commencé par poursuivre des études en langues et littératures roumaine et française à Bucarest. Aussi, quand je suis arrivée en France, en 1996, ce fut avec un bagage à dominante philologique : en Roumanie, on privilégiait l'histoire des langues latines, la phonétique historique et la lexicographie dans une perspective étymologique (on s'attachait à étudier l'origine, l'évolution du sens et de la forme des mots). J'en ai gardé une passion pour l'évolution des langues et les lois qui la gouvernent. C'est tout naturellement que je me suis inscrite en Sciences du langage à l'université Lyon 2.

Mon ouverture sur les enjeux du langage automatique intervient à la fin des années 1990. En 1998, je me suis inscrite au master Ingénierie des langues proposé par l'Aupelf [Association des universités partiellement ou entièrement de langue française]. Ce master pour le moins exotique à l'époque (il était parmi les premiers à proposer ce croisement entre ingénierie et linguistique) a été pour moi comme une révélation. Des cours mettaient en relation langue et traitement informatique. Pour la première fois, j'ai eu un aperçu à la fois du fait qu'on étudiait la langue dans ce cadre applicatif et qu'il pouvait y avoir un impact de la variation langagière sur la performance de ces outils. Parmi les cours proposés, quelques-uns se tenaient au Limsi (les autres avaient lieu à

Marne-la-Vallée et à l'ISTN du CEA). J'ai eu grand plaisir à fréquenter ce laboratoire, où je ne demandais qu'à avoir l'opportunité de travailler.

Depuis, beaucoup de chemin a été fait en matière de linguistique et de traitement automatique. Pour ma part, je n'ai eu de cesse de m'intéresser aux systèmes issus du monde de l'ingénierie, non pas tant pour en devenir une spécialiste, mais pour faire de la recherche en linguistique. Mon sujet de thèse de doctorat porta sur la comparaison entre humains et systèmes automatiques dans l'identification des langues. Puis j'ai eu l'immense chance de décrocher un poste CNRS au... Limsi.

– *Sur quoi portèrent vos travaux de recherche ?*

J'ai poursuivi mes recherches sur la comparaison entre les humains et les systèmes automatiques, puis en portant davantage mon attention aux erreurs des systèmes de reconnaissance vocale et à la variation inhérente à la langue parlée. Progressivement, j'en suis arrivée à m'intéresser à la relation de la variation captée par ces systèmes comparée à l'évolution des langues, lorsqu'un élément de variation de stabilise et devient fait spécifique à une langue, adopté par tous les locuteurs. C'est ainsi qu'a pu s'instaurer un pont durable entre ma formation première, philologique, et le traitement automatique.

– *Quel défi cela représente-t-il en termes d'interdisciplinarité ?*

Les défis sont multiples : tout d'abord, travailler avec des chercheurs venant d'autres horizons disciplinaires ou professionnels que le sien suppose de trouver un « langage » commun. Il s'agit aussi de définir un thème de recherche qui puisse intéresser chacun. À l'Université Paris-Saclay, nous ne sommes que trois linguistes (tous chercheurs CNRS de la section 34, linguistique). C'est dire s'il nous faut trouver notre place. Au sein du Limsi, j'ai eu l'opportunité de travailler dans une équipe formidable en Traitement du langage parlé (TLP) avec des collègues dont les intérêts scientifiques allaient bien au-delà de la seule performance des systèmes. Nous avons pu développer des sujets portant à la fois sur le fonctionnement des langues et sur la relation entre spécificités linguistiques et performances automatiques. J'ai pu obtenir des résultats

fiables et pertinents pour mieux comprendre des faits linguistiques jusqu'alors étudiés majoritairement sur des corpus de taille restreinte, dits « de laboratoire ».

Un autre défi est propre à la reconnaissance de nos travaux au sein des sciences humaines. Ce n'est que depuis peu que les approches que je développe et qui impliquent de grands corpus et l'utilisation de systèmes automatiques en tant que véritables outils linguistiques sont reconnues par mes collègues linguistes comme véritablement « linguistiques ». Reste qu'il demeure difficile d'obtenir des financements purement SHS.

– *En quoi le* Big Data *a-t-il changé la donne par rapport à ce que pouvait recouvrir jusqu'ici l'ingénierie linguistique ?*

Lorsque je suis arrivée en France, mon laboratoire de recherche me mettait à disposition magnétophone et logiciels dont la licence coûtait cher, pour les besoins de mon travail de thèse. Créer et analyser son corpus de « laboratoire » était donc laborieux. Désormais, on dispose de quantités impressionnantes de données langagières, écrites et orales, sur Internet. À défaut, il est relativement facile de les produire, selon le besoin expérimental, et de les exploiter avec des outils libres. Un étudiant en master peut enregistrer les données de son travail de mémoire avec un simple smartphone et trouver des logiciels libres en tout genre sur Internet pour analyser ses données.

On ne peut prétendre faire abstraction de cette richesse : ces données existent, nous pouvons y avoir accès et fonder nos réflexions linguistiques théoriques sur l'existant.

> ***« (…) je n'ai eu de cesse de m'intéresser aux systèmes issus du monde de l'ingénierie, non pas tant pour en devenir une spécialiste, mais pour faire de la recherche en linguistique »***

– Sans incidence d'un point de vue méthodologique ?

Si. Passer du corpus restreint, de laboratoire, à des données massives implique une autre technique de travail, une autre manière d'aborder les faits de langue, et, donc, des résultats d'un type différent. Les grands corpus et les méthodes de travail scientifique associées enrichissent et valident des approches linguistiques précédentes. Il ne s'agit pas de nier tout le passé des données dites de « laboratoire », mais de reconnaître une nouvelle manière de faire de la linguistique. De ce point de vue, le Limsi a eu un rôle pionnier. Dès les années 2000, mes collègues ont proposé des études sur la variation linguistique en s'appuyant sur les grands corpus utilisés pour entraîner les systèmes de reconnaissance vocale. Précisons cependant que grand corpus ne veut pas forcément dire *Big Data*. Celui-ci signifie une absence totale de maîtrise des caractéristiques du corpus. Or, en ce moment, nous utilisons plutôt des centaines d'heures – ce qui est déjà beaucoup – et nous savons typer les caractéristiques de nos données (journalistiques, conversations spontanées etc.).

– Vous avez été lauréate de plusieurs appels à projets de la MSH, notamment pour les projets HistorIA (appel à projets Émergence 2017) et HistorIA 2 (appel à projets Maturation 2018). Dans quelle mesure ont-ils contribué à l'approfondissement et l'élargissement de vos recherches ?

Pour mémoire, le premier, HistorIA vise à mieux appréhender les mécanismes acoustiques et articulatoires qui régissent les changements sonores, à partir de données synchroniques, et ce en faisant appel à des outils qui relèvent de l'intelligence artificielle, notamment des systèmes de reconnaissance vocale. Il fait pour cela appel à de grands corpus oraux investigués avec des moyens numériques et statistiques. Ce projet a donné lieu en novembre 2017 à un colloque sur le thème « Linguistique et *Big Data* ».

HistorIA.2 vise à mener plus loin cette réflexion et à approfondir l'étude de l'évolution des langues avec les moyens de l'intelligence artificielle à partir d'une exploitation de données massives multi-sources (synchroniques, diachroniques, dialectales). Ce projet a notamment contribué à la réalisation et la mise en ligne de l'Atlas sonore des langues

régionales, dont les résultats permettent de répondre à un triple objectif : la sauvegarde du patrimoine linguistique, la validation de théories et le développement du *machine learning* pour la reconnaissance vocale.

S'il y a un premier enseignement à tirer de ces projets, c'est que les systèmes automatiques ne sont pas seulement des outils permettant de pré-traiter des données linguistiques, mais de véritables instruments de recherche permettant d'aller au plus profond des questions que les linguistes « classiques » se posent : comment évoluent les langues ? Comment d'une variation observable à un instant t on arrive à des faits linguistiques stables, à des changements linguistiques accomplis ?

> ***« (…) HistorIA vise à mieux appréhender les mécanismes acoustiques et articulatoires qui régissent les changements sonores, à partir de données synchroniques, et ce en faisant appel à des outils qui relèvent de l'intelligence artificielle (…) »***

– Dans quelle mesure la MSH aura-t-elle permis de faire contribuer la linguistique aux humanités numériques que, pour votre part, vous enseignez à Paris 3 Sorbonne Nouvelle – ENEAD ?

Comme vous le savez, la question des humanités numériques est vaste et encore ambiguë. Il reste encore beaucoup à explorer dans ce domaine, en gardant à l'esprit qu'il s'agit d'un enjeu sociétal. Je pense que nous sommes tous d'accord pour considérer que ces « humanités numériques » doivent rimer avec inter- et pluridisciplinarité. Elles appellent la valorisation d'outils informatiques pour le stockage et l'analyse des données spécifiques aux humanités. Pour ce qui est de la linguistique, la MSH Paris-Saclay peut avoir un rôle phare, en montrant qu'il y a des opportunités de collaboration féconde entre elle et les technologies du numérique. Si les activités « historiques » du laboratoire Limsi concernent d'abord le traitement automatique des langues, les projets labellisés par la MSH ont apporté la démonstration que la perspective inverse – une approche des systèmes automatiques des langues à partir des questions que se posent les linguistes – est non seulement possible,

mais encouragée dans l'environnement de recherche de Paris-Saclay. C'est d'ailleurs la volonté du laboratoire Limsi que de mettre lui-même en avant une approche interdisciplinaire des langues.

– *On mesure l'intérêt des plateformes numériques, mais aussi d'opérateurs de la télécommunication et autres pour vos travaux de recherche. Comment appréhendez-vous leur valorisation ?*

Mes premiers travaux avaient la valorisation pour finalité. Je cherchais à mobiliser la linguistique pour améliorer les systèmes automatiques et les rendre plus viables. Désormais, ils consistent davantage à montrer comment des outils et des méthodes qui viennent d'un domaine applicatif (celui de l'ingénierie linguistique, donc) peuvent être récupérés par la linguistique classique. Concrètement, je me sers des systèmes automatiques pour mesurer la variation dans les langues *via* des paramétrisations spécifiques. Ils permettent de traiter des centaines ou des milliers d'heures d'enregistrements audio en très peu de temps et, de ce fait, offrir une vision très globale d'un motif de variation orale.

Les réponses qu'on obtient, si elles ne débouchent pas toutes seules sur des logiciels ou des brevets, contribuent à améliorer ces systèmes en intégrant cette dernière (*via* ce qu'on appelle les dictionnaires de prononciation). Elles permettent ainsi de réduire les erreurs de transcription automatique. Elles sont évidemment précieuses à la recherche fondamentale sur les langues et leurs évolutions.

MSH PARIS-SACLAY

5 ANS D'INTERDISCIPLINARITÉ SUR UN PLATEAU

La construction d'un grand pôle scientifique sur le plateau de Saclay est avant tout comprise comme la création d'un fort potentiel de recherche technologique. Pourtant, les Sciences de l'Homme et de la Société ont un rôle majeur à y jouer, par leur volume et par leur place essentielle en termes d'activités et de dispositifs d'innovation.

La MSH Paris-Saclay, créée en 2015, apporte sa contribution à ce défi par son engagement au service des équipes du périmètre saclaysien. Le travail réalisé lui permet d'occuper une place centrale dans la promotion et l'organisation de leurs recherches interdisciplinaires, de développer une position d'interface entre les SHS et de s'ouvrir aux autres disciplines (sciences de la vie, sciences exactes, sciences de l'ingénieur).

Cet ouvrage a pour but de présenter le travail réalisé au cours de ces cinq premières années, à partir d'un bilan des recherches et d'interviews dans lesquels les trois directeurs successifs reviennent sur leur parcours. Dix chercheuses et chercheurs emblématiques des projets passés et en cours apportent également leurs témoignages, afin d'éclairer à la fois la diversité des thèmes de recherche et la variété des résultats obtenus.

ISBN 978-4-490369-04-1
EAN 9784490369041