École Normale Supérieure de Lyon

Les phénomènes sans échelle à la source de la révolution numérique

Présenté et soutenu publiquement le 22 juin 2023
pour l'obtention de

l'Habilitation à Diriger des Recherches

par

Jean-Philippe Magué

Garante :	Kris Lund	Ingénieure de Recherche, CNRS
Rapporteurs :	Jean-Pierre Chevrot	Professeur Émérite, Université Grenoble-Alpes
	François Pellegrino	Directeur de Recherche, CNRS
	Marcello Vitali-Rosati	Professeur, Univesité de Montréal
Examinateurs :	Éric Fleury	Directeur de recherche, INRIA
	Sabine Loudcher	Professeure, Université Lumière Lyon 2

Table des matières

Figures

Tableaux

Préambule

Une quête disciplinaire

Enfant, je voulais être magicien. J'étais fasciné par les effets des tours qui semblaient jouer à l'encontre des possibles, mais j'étais plus encore subjugué par l'existence du truc, par le fait que derrière le mystère se dévoile une explication. Cet attrait et cette curiosité du pourquoi et du comment ne m'ont jamais quitté et ont façonné mon esprit scientifique.

Lycéen, cette fascination était toute entière tournée vers l'astrophysique. Elle était partagée avec Brice Ménard, ami rencontré sur les bancs du lycée. Les discussions passionnées que nous avions alors, et que nous avons toujours, ont été déterminantes dans la construction de mon esprit scientifique. Le texte proposé ici lui doit d'ailleurs énormément. Notre trajectoire professionnelle était claire à nos yeux, toute tracée, nous étions déterminés à être astrophysiciens. Lui le sera. Pour ma part, cette trajectoire pensée rectiligne et dirigée vers un objectif fixe s'est transformée en une courbe traversant un éventail de disciplines scientifiques.

Étudiant, je suis entré à l'Université Paris XI, à Orsay, comme aspirant astrophysicien dans une filière, un DEUG, enseignant la physique, les mathématiques et l'informatique. Bien que lecteur depuis des années de magazines et livres de vulgarisation scientifique, cette entrée à l'université m'a ouvert les yeux sur la diversité des champs de recherche, sur la multiplicité des questions accrochant ma curiosité. Mon intérêt pour l'astrophysique ne s'est pas émoussé en tant que tel, mais il a perdu son unicité pour se retrouver parmi d'autres. Et au sein de ces autres, s'est détachée la cognition. J'ignorais alors jusqu'à l'existence du mot. La prise de conscience que les processus mentaux, leur fonctionnement, leur production par cerveau, puissent être l'objet d'une démarche scientifique m'a emporté. J'étais non seulement séduit par l'objet des *sciences cognitives*, mais également par leur pluriel :

les sciences cognitives. Ce pluriel était à la fois le reflet de la complexité l'objet, la cognition, mais aussi celui de la complexité de l'entreprise scientifique, de la nécessité de croiser les angles de vue, de construire des édifices théoriques mobilisant des concepts issus de disciplines différentes. Je trouvais illuminant les contenus des programmes des DEA de sciences cognitives, affichant neurosciences, psychologie, informatique, philosophie ou encore linguistique. C'est le moment où la puissance de l'interdisciplinarité m'est apparue de manière flagrante.

J'ai ainsi bifurqué. L'informatique étant le dénominateur commun entre le DEUG dans lequel j'étais inscrit et les DEA auxquels j'aspirais, c'est la voie pour laquelle j'ai opté : j'aborderai les sciences cognitives via l'intelligence artificielle. Le magistère d'informatique que j'ai alors suivi à Grenoble aura été une étape marquante. Il s'en est fallu de peu pour que l'informatique théorique prenne le dessus et me détourne des sciences cognitives. Mais j'ai surtout acquis une maitrise et une compréhension fines des outils numériques, une forme de littératie qui m'accompagne depuis et qui structure toute mon activité scientifique. Ce magistère d'informatique était une formation à la recherche, par la recherche. Les stages étaient une composante importante et m'ont permis de confirmer une bonne fois pour toutes mon désir d'embrasser une carrière académique. Ils ont également été l'occasion de donner une coloration sciences cognitives à ce cursus. En licence, j'ai fait un stage sous la direction de Bernard Amy très ancré dans l'intelligence artificielle, portant sur l'explicitation des connaissances d'un réseau de neurones (sujet brûlant d'actualité 25 ans plus tard). En maîtrise, j'ai souhaité profiter de mon stage pour me rapprocher d'une autre discipline dans le giron des sciences cognitives, la linguistique. J'ai travaillé dans une équipe de traduction automatique sous la direction de Gilles Serasset. Il y avait quelque chose de renversant à pouvoir utiliser un ordinateur pour modéliser la langue, expliciter des structures latentes dans une chaîne de caractères. Le pont avec les structures cognitives était là flagrant.

Mon année en DEA de sciences cognitives à l'Université Lyon 2, a été une explosion intellectuelle, à la hauteur des attentes que j'y avais placées. La diversité des enseignements m'a alors embarqué sans retour possible dans l'aventure de l'interdisciplinarité. Il était évident que l'articulation de discours d'origines disciplinaires variées les enrichissait les uns les autres. Dans un certain mouvement de balancier entre intelligence artificielle et linguistique, mon mémoire, dirigé Par Hélène Paugam-Moisy, a porté sur la

modélisation de processus visio-spatiaux à l'aide de réseaux de neurones (Paugam-Moisy et al., 2002)[1]. Mais l'aspect le plus important de cette année ont été les cours de linguistique par Jean-Marie Hombert. Ses cours portaient sur l'origine du langage; ils sautaient d'une échelle temporelle à l'autre et étaient eux-mêmes profondément interdisciplinaires, mobilisant non seulement la linguistique, mais aussi l'anthropologie, l'archéologie, les neurosciences, la modélisation informatique...Délaissant l'intelligence artificielle pour me tourner pour de bon vers la linguistique, j'ai entrepris ma thèse sous sa direction.

J'ai réalisé mon doctorat en sciences cognitives à l'Université Lyon 2 dans le laboratoire Dynamique du Langage. Pouvoir être apprenti chercheur dans ce laboratoire a été une chance formidable. Le travail que j'y ai réalisé doit beaucoup à son directeur d'alors, François Pellegrino, qui a secondé mon directeur de thèse lorsque celui a été appelé à prendre des responsabilités au CNRS, à Paris. Le séjour de six mois que j'ai fait pendant mon doctorant à Université de Californie à Berkeley, temple de la linguistique cognitive, a été une étape majeure de ma maturation intellectuelle. Ma thèse portait sur les changements sémantiques, sur l'évolution au cours du temps du sens des mots. Puisque le sens réside notamment dans nos têtes, la manière dont il change ne peut qu'être contrainte par les structures et mécanismes cognitifs qui le portent. Mon objectif, en grande partie atteint (Magué, 2005), était de produire des modèles informatiques démontrant cette proposition. Le changement linguistique est longtemps resté un mystère pour qui s'y intéressait : comment résoudre le paradoxe entre l'observation de langues qui changent (par exemple, le latin devenu français) et l'observation de langues qui ne changent pas (une génération parle la même langue que celle de ses grands-parents, mais aussi que celle de ses petits-enfants). La réponse est venue dans les années 60, en comprenant que la clé est la variation à l'intérieur même d'une population : cette variation est à la fois l'origine du changement et sa manifestation. La compréhension de ces mécanismes a été pour moi une étape majeure : le langage est à la fois un objet cognitif et un objet social, et comprendre ses dynamiques (ce qui était mon objectif) nécessite de se plonger dans les dynamiques sociales de la population des locuteurs. Le changement linguistique est la manifestation macroscopique, émergente, d'une myriade d'interactions linguistiques (Magué, 2002, 2006a,

[1]Les références citées dans ce préambule correspondent à l'ensemble de mes travaux publiés.

2006b).

Ainsi s'est passée ma rencontre avec la *Science des Systèmes Complexes*. Une nouvelle forme d'interdisciplinarité, mais un pluriel différent. À la différence des sciences cognitives, la science est au singulier et c'est l'objet qui est, d'apparence, pluriel. Le programme scientifique n'est pas de multiplier les points de vue épistémologiques pour cerner un objet, mais de viser une approche unifiée pour rendre compte de phénomènes issus d'une diversité de champs scientifiques. Toutes ces formes de décloisonnement résonnent en moi. Parti des sciences exactes et expérimentales, de la physique, en direction des sciences humaines et sociales, la linguistique, les frontières disciplinaires m'apparaissent à la fois fragiles et vaines.

Ma compagne était en thèse en même temps que moi. Nous n'avons pas accepté l'injonction tacite du milieu académique qui veut que l'on sacrifie sa vie personnelle au profit de sa vie professionnelle et notre premier enfant est né pendant nos années de thèse.

Nos thèses terminées, nous sommes partis à Chicago. L'objectif de mon postdoc avec Salikoko Mufwene au département de linguistique de l'Université de Chicago était inscrit dans cette vision du langage comme système social, dynamique et complexe, et était de proposer des modèles multiagents de la compétition entre variants linguistiques au sein d'une population (Magué, 2007). Mais entre la difficulté de trouver des financements et celle de mener de front vie familiale et vie scientifique, s'instille un doute qui vite grandit. Je renonce au milieu académique et, de retour en France, à la naissance de notre deuxième enfant, j'occupe un poste d'ingénieur dans une entreprise éditrice de logiciels. Cette expérience lève mes doutes : le milieu industriel ne me nourrit pas intellectuellement, je dois retourner vers le monde académique. Je postule et suis recruté sur un poste d'ingénieur de recherche contractuel au laboratoire ICAR, à l'ENS Lettres et Sciences Humaines.

J'intègre l'équipe pilotée par Serge Heiden qui développe le logiciel de textométrie TXM. Le travail que nous menons est à la fois théorique, en cherchant à modéliser ce qu'est un texte et la manière dont on l'appréhende, et pratique puisqu'il faut développer le logiciel (Heiden et al., 2010 ; Loiseau et al., 2009). Ce faisant, j'intègre les *humanités numériques*, une nouvelle entreprise interdisciplinaire, avec ses propres modalités. Ici, pas d'objet commun, pas de cadre épistémologique commun, mais une destinée commune à l'ensemble des sciences humaines et sociales de s'emparer des opportuni-

tés méthodologiques et épistémologiques portées par la vague numérique qui traverse nos sociétés. Si je découvre les humanités numériques à cette occasion, c'est à la manière d'un monsieur Jourdain puisque, rétrospectivement, ma pratique de la linguistique a dès le début été nourrie par les opportunités offertes par le numérique. Lorsque l'ENS Lettres et Sciences Humaines (ENS-LSH) ouvre au concours un poste de maître de conférences en humanités numériques, je postule et je l'obtiens.

La pratique du métier

Mal armé

L'obtention de ce poste est un aboutissement. L'aboutissement, auréolé d'un prestige certain, d'un parcours débuté comme un rêve d'enfant, marqué par des bifurcations et des opportunités, des échecs et des succès.

L'obtention de ce poste est aussi un commencement. Ce poste était particulier puisque la personne recrutée devait prendre la direction d'un service d'appui à la recherche de l'ENS-LSH qui offrait à ses laboratoires (en lettres et sciences humaines) une expertise et des compétences numériques pour leurs projets en humanités numériques. J'ai donc pris en octobre 2009 la tête d'une équipe d'ingénieures, quelques mois à peine avant la fusion des deux ENS lyonnaises, effective au 1er janvier 2010. Je ne connais alors rien, ou si peu, aux enjeux du numérique spécifiques à l'histoire, aux lettres, aux sciences politiques…Je ne connais rien, ou si peu, aux enjeux qui sous-tendent la fusion entre les deux ENS, aux conséquences de placer ici plutôt que là les humanités numériques dans le nouvel organigramme en cours de constitution. Je n'étais pas seul à penser l'avenir de la structuration des humanités numériques dans la nouvelle ENS, les ingénieures impliquées et la direction apportaient leurs contributions, mais j'avais à jouer un rôle central auquel je n'étais pas préparé. Les premières années qui ont suivi ma prise de fonction ont été dures, mais passionnantes, j'ai beaucoup appris.

Transmettre

L'obtention de ce poste a également été pour moi un commencement du point de vue de l'enseignement. Dès ma thèse, d'abord comme vacataire puis comme ATER, j'ai eu l'opportunité de me présenter devant des étudiants

pour enseigner. Ces expériences m'avaient permis de découvrir le plaisir à transmettre. Mais c'est à travers mon poste que j'ai pu apprécier l'enseignement dans toutes ses facettes et toute sa profondeur. L'essentiel de mon activité d'enseignement est à destination d'étudiants et d'étudiantes en lettres, en sciences humaines ou en sciences sociales, et concerne des thématiques liées au numérique : la ligne directrice de mes enseignements est la transmission de ma littératie numérique. La compréhension des technologies numériques et des enjeux liés à leurs pouvoirs transformateurs ou destructeurs est un prérequis à leur analyse, à la formulation de discours critiques à leur égard et, par conséquent, à être un esprit éclairé au 21e siècle. Je m'attèle donc à cette tâche de transmission avec ardeur. J'ai créé de nombreux cours, allant de l'initiation à la programmation à de l'analyse de données poussée, de l'histoire d'internet aux recherches actuelles en humanités numériques. Dans mes cours, j'applique tant les technologies numériques aux sciences humaines et sociales que les sciences humaines et sociales aux technologies numériques.

Cette activité d'enseignement s'est accompagnée au fil des années de prises de responsabilités. J'ai participé activement au montage de deux masters. D'abord le master architecture de l'information, notamment avec Jean-Michel Salaün et Benoît Habert, qui, s'appuyant sur les sciences de l'information, le design et le numérique, formait des étudiantes et des étudiants aux techniques et aux enjeux de la structuration et de la circulation de l'information sur le Web (Habert et al., 2012 ; Magué, 2014 ; Magué & Mabillot, 2015 ; Mille & Magué, 2012). Le master humanités numériques ensuite, en collaboration avec de nombreux collègues de l'université Lyon 2 (notamment Sabine Loudcher), de l'Université Lyon 3 (Bruno Bureau) et de l'enssib (Agnieszka Tona), qui est un master exigeant pour avoir l'originalité de ne prendre que des étudiantes et des étudiants en double diplôme, de manière à compléter leur master disciplinaire par une formation en humanités numériques. Je suis aujourd'hui responsable du master pour l'ENS de Lyon. C'est dans ces deux masters, architecture de l'information et humanités numériques, qu'étaient inscrits la majorité de la quinzaine d'étudiants et d'étudiantes dont j'ai dirigé le mémoire (les autres étant en sciences du langage, traitement automatique des Langues, sciences sociales ou informatique).

Laisser des traces

Les premières années après l'obtention de mon poste de maître de conférences, ma recherche était principalement guidée par mon activité de structuration des humanités numériques à l'ENS, et portait sur des questions d'édition et de publication électroniques. L'enjeu était d'éviter que chaque laboratoire SHS de l'ENS déploie ses propres technologies pour soutenir ses projets, en mutualisant des compétences, des développements informatiques et des infrastructures numériques (Beaugiraud et al., 2011 ; Château-Dutier et al., 2015, 2016 ; Magué, 2011). De 2010 à 2015 j'ai, pour la première fois, codirigé avec Benoît Habert la thèse d'Adrien Barbaresi intitulée *Ad hoc and general-purpose corpus construction from web sources*. En parallèle, j'ai continué à nourrir une réflexion sur le langage comme système complexe (Loiseau et al., 2011).

Cet aspect de mes recherches est redevenu central à partir de 2014, peu après la naissance de notre troisième enfant, quand j'ai entrepris de m'orienter vers la sociolinguiste computationnelle. La place sans cesse croissante prise par les technologies numériques dans nos vies nous conduit à laisser des traces de nos interactions sociales en grande quantité, traces qui plus est linguistiques lorsque nous nous exprimons sur les médias socionumériques. Ces traces sont autant de données potentielles pour revisiter avec une méthodologie renouvelée le programme de la sociolinguiste variationniste qui interroge les liens entre variation et changement linguistiques. Sur la base de cette argumentation, j'ai initié puis piloté le projet *SoSweet, une sociolinguistique de Twitter* financé par l'ANR à hauteur de 650 000 € en collaboration avec Márton Karsai, Djamé Seddah et Jean-Pierre Chevrot. Le consortium réunissait une quinzaine de participants issus de quatre laboratoires. Ce projet a donné lieu à plus d'une soixantaine de communications scientifiques impliquant plus de 80 personnes (Abitbol et al., 2018 ; Abitbol, Chevrot et al., 2017 ; Abitbol, Karsai, Chevrot et al., 2017a, 2017b ; Abitbol, Karsai, Magué et al., 2017 ; Chevrot et al., 2015, 2019 ; Levy Abitbol et al., 2017, 2018 ; Magué, 2018 ; Magué, Fleury et al., 2015 ; Magué, Quignard et al., 2015a, 2015b ; Magué et al., 2020 ; Mangold et al., 2017 ; Tarrade et al., 2022 ; Thibert & Magué, 2016 ; Thibert, Magué et al., 2016 ; Thibert, Zeynaligargari et al., 2016). Dans le cadre de ce projet, j'ai bénéficié de deux délégations successives de 6 mois, à l'INRIA puis au CNRS, que j'ai passées à l'IXXI, dans l'équipe INRIA DANTE dirigée par Éric Fleury. Le projet SoSweet m'a également donné l'opportuni-

té de codiriger, avec Nathalie Rossi-Gensane, une seconde thèse, menée par Clément Thibert et intitulée *Sociolinguistique des médias sociaux : étude de la variabilité linguistique sur Twitter* (thèse qui n'est malheureusement pas arrivée à son terme pour raisons médicales). Dans le prolongement de ce projet, je codirige, avec Jean-Pierre Chevrot, la thèse de Louise Tarrade intitulée *Approche computationnelle du changement linguistique sur Twitter : population, communautés, individus*.

Depuis 2021, je co-coordonne avec Marc Allassonnière-Tang le projet MACDIT[2], financé par le LabEx Aslan. Ce projet s'inscrit également dans une vision du langage comme système social, dynamique et complexe. Son objectif est d'étudier les interactions entre les niveaux individuels et collectifs de la variation et du changement linguistique. Nous demandons d'une part comment l'individuel construit le collectif (plus précisément comment les conventions linguistiques collectives sont construites à travers les interactions interindividuelles) et, d'autre part, comment le collectif construit l'individuel (plus précisément, comment les conventions collectives influent sur les trajectoires linguistiques des individus). Dans le cadre de ce projet, je codirige avec Denis Vigier la thèse de Jean-Baptiste Chaudron intitulée *Approches computationnelles de la variation et du changement linguistique*.

De l'individuel au collectif et réciproquement

J'ai également perçu et recherché ces influences réciproques du collectif et de l'individuel à travers les rôles et responsabilités que j'exerce dans diverses structures. Je conçois ces prises de responsabilités à la fois comme une implication au sein d'un collectif avec comme objectif d'apporter ma contribution et comme l'opportunité de nourrir ma progression intellectuelle des influences de ce collectif.

Ainsi, depuis 2016, je suis membre du comité de pilotage du LabEx Aslan (Advanced Studies on Complexities of Language). De 2016 à 2019, j'étais le coresponsable de l'axe *Penser la complexité linguistique*. Depuis 2020, je suis coresponsable de l'axe *Modelling and Digital Humanities*. Ce LabEx, dirigé de 2011 à 2018 par François Pellegrino et par Kris Lund depuis, a un rôle central dans la structuration des sciences du langage à Lyon et m'a participa-

[2]Modèles multiagents et données de médias sociaux : dynamiques collectives et trajectoires individuelles dans les populations linguistiques

tion m'a permis d'habiter pleinement cette communauté et de contribuer à y faire vivre des approches numériques et quantitatives.

Je me suis également investi dans le département *éducation et humanités numériques* de l'ENS de Lyon dirigé par Karine Bécu-Robinault et dont je suis le directeur adjoint depuis 2018. Ce rôle s'inscrit dans la continuité de la mission initialement liée à mon poste de faire vivre les humanités numériques à l'ENS.

Enfin, depuis 2016 je suis membre du Comité de direction de l'IXXI, l'Institut Rhônalpin des Systèmes complexes. En 2020, j'en suis devenu l'un des directeurs adjoints, avec Claire Lesieur, alors qu'il était dirigé par Pierre Borgnat. Depuis 2021, je suis l'adjoint de Patrice Abry. L'IXXI est un lieu exceptionnel. Il fédère dans l'ensemble de la région des chercheuses et des chercheurs convaincus que les croisements disciplinaires sont un moteur sans pareil pour faire germer des idées nouvelles et explorer des territoires scientifiques nouveaux. Avec d'autres, j'ai contribué, et je m'y applique encore, à faire entrer les sciences humaines et sociales dans le périmètre de l'Institut. Le projet SoSweet est directement issu de ces efforts.

J'ai souhaité que ces investissements professionnels ne soient pas menés au détriment d'une vie personnelle et familiale. En 2019, j'ai demandé et obtenu un temps partiel et ai travaillé à 90% pendant 3 ans.

Introspection (inter)disciplinaire

Un bilan introspectif de cette trajectoire scientifique pourrait se résumer en une question existentielle, un *Qui suis-je ?* disciplinaire. Dans sa pratique classique, l'interdisciplinarité se joue à l'échelle d'un groupe, chacun ayant sa provenance disciplinaire, son point de vue, un socle conceptuel sur lequel s'appuyer pour nouer avec les autres des interactions afin d'apporter sa pierre disciplinaire à un édifice collectif. Je ne vis pas l'interdisciplinarité ainsi, mais à une l'échelle individuelle, portant en moi-même une pluralité de points de vue. Je me sens riche de tout ce que j'ai acquis à travers toutes les disciplines que j'ai traversées, mais en manque d'une origine disciplinaire ferme, d'un socle solide à partir duquel affirmer mon point de vue. Bien qu'officiellement linguiste, ma formation ne m'a apporté que quelques poignées d'heures de cours en linguistiques, qui plus est sur des thématiques périphériques à la discipline. J'ai appris sur le tas, en fonction de mes intérêts

du moment et ma vision générale de la linguistique est restée longtemps lacunaire. La crise de légitimité qui a suivi mon entrée en fonction fut forte.

Au fil des années, j'ai toutefois dépassé ce sentiment d'illégitimité pour assumer mon positionnement intrinsèquement interdisciplinaire. J'y ai trouvé au moins deux forces. Pour illustrer la première, je me permets un détour la manière dont l'analyse des réseaux sociaux a contribué à la compréhension de la diffusion des changements linguistiques. Dans les années 70 le couple de sociolinguistes Lesley et James Milroy s'est intéressé à trois quartiers de Belfast, chacun ayant ses propres spécificités linguistiques, chacun correspondant à une communauté linguistique. Ils se sont particulièrement penchés sur les liens sociaux qui structuraient ces communautés, et ont identifié d'une part des personnes très centrales à ces communautés linguistiques, reconnues comme telles et n'interagissant majoritairement qu'avec des personnes du quartier, des membres de leur communauté. Ils ont identifié d'autre part des personnes qui, à l'inverse, étaient plus périphériques et entretenaient des relations avec des personnes extérieures à la communauté. D'un côté, des personnes très ancrées dans leur environnement, à l'identité marquée, de l'autre des personnes naviguant d'un monde à l'autre, avec une appartenance identitaire bien moins affirmée. Les spécificités linguistiques propres à chaque communauté sont les plus fortes chez ces personnes centrales qui revendiquent ainsi leur appartenance à la communauté. Les locuteurs périphériques marquent moins ces spécificités. Cela leur permet d'être les porteurs du changement. Ce sont eux qui amènent les innovations. Ce sont les passeurs, qui transmettent les traits linguistiques d'une communauté linguistique à une autre. À l'échelle de la communauté, vus comme périphériques, ces derniers pourraient passer pour secondaires. Lorsqu'on regarde les choses à l'échelle du dessus, cette périphéricité apparait comme autant de ponts et leur rôle devient crucial. Transposé dans le monde académique, je suis périphérique dans bien des disciplines, mais de fait central dans la circulation d'idées, de concepts, de méthodes, etc., entre disciplines.

Outre ce rôle de passeur, la deuxième force que je tire de mon interdisciplinarité incarnée est d'être en position de synthèse. De pouvoir jeter sur le monde un regard transverse, nourri d'influences multiples. « *Les deux pôles possibles de la connaissance sont de savoir presque tout sur presque rien ou de savoir presque rien sur presque tout* » m'a dit un jour Kris Lund. J'ai depuis longtemps opté pour le second pôle et vécu des périodes de doutes

en étant trop focalisé sur *savoir presque rien* alors que la partie importante de la phrase est *sur presque tout*.

Ce qui suit

Le texte que je propose dans ce document est le fruit de ma trajectoire bien que je n'y synthétise pas mes recherches passées. J'ai préféré profiter de cette occasion pour développer des idées sur la nature de la révolution numérique qui maturent depuis longtemps, mais que je n'avais pas mises par écrit jusqu'alors. L'argument que je construis dans ce texte s'ancre dans les humanités numériques, les sciences du langage, les sciences sociales, la science de la complexité et les sciences cognitives : l'interdisciplinarité en est une caractéristique majeure.

S'il ne revient pas sur mes recherches passées, ce texte est toutefois accompagné de quatre articles publiés qui les illustrent.

Le premier (Mille & Magué, 2012) marque le début de ma réflexion sur les technologies numériques comme objet social ayant des conséquences sociales. Nous y discutons la spécificité du fait documentaire et des processus éditoriaux sur le web.

Le second (Abitbol et al., 2018) est issu du projet SoSweet. Il analyse à partir de données riches, complexes et massives la manière dont la variabilité linguistique observée sur Twitter est socialement structurée. Il illustre les idées, la démarche et l'interdisciplinarité que j'ai promues en montant le projet SoSweet.

Le troisième (Magué et al., 2020) propose un contre-point : à partir du même matériau, le corpus collecté dans le projet SoSweet, nous développons une approche purement linguistique et purement qualitative pour analyser le rôle tenu par les émojis et les émoticônes dans la segmentation des énoncés. Cet article montre que ma revendication d'une interdisciplinarité incarnée ne m'empêche pas m'inscrire dans les canons de ma discipline.

Enfin, le dernier (Tarrade et al., 2022) se penche sur les innovations lexicales sur Twitter et met en avant mon rôle de directeur de thèse. Pointer des pistes et des risques, échanger, discuter et même débattre, alterner entre la construction stratégique de l'ensemble de la thèse et la résolution du problème technique, guider, accompagner, être acteur et témoin privilégié de l'éclosion d'un chercheur ou d'une chercheuse. Je prends ce rôle très à cœur.

1 Qu'y a-t-il de révolutionnaire dans la révolution numérique ?

« J'ai donc ouï dire qu'il existait près de Naucratis, en Égypte, un des antiques dieux de ce pays, et qu'à ce dieu les Égyptiens consacrèrent l'oiseau qu'ils appelaient ibis. Ce dieu se nommait Theuth. C'est lui qui le premier inventa la science des nombres, le calcul, la géométrie, l'astronomie, le trictrac, les dés, et enfin l'écriture. Le roi Thamous régnait alors sur toute la contrée ; il habitait la grande ville de la Haute-Égypte que les Grecs appellent Thèbes l'égyptienne, comme ils nomment Ammon le dieu-roi Thamous. Theuth vint donc trouver ce roi pour lui montrer les arts qu'il avait inventés, et il lui dit qu'il fallait les répandre parmi les Égyptiens. Le roi lui demanda de quelle utilité serait chacun des arts. Le dieu le renseigna ; et, selon qu'il les jugeait être un bien ou un mal, le roi approuvait ou blâmait. On dit que Thamous fit à Theuth beaucoup d'observations pour et contre chaque art. Il serait trop long de les exposer. Mais, quand on en vint à l'écriture :

« Roi, lui dit Theuth, cette science rendra les Égyptiens plus savants et facilitera l'art de se souvenir, car j'ai trouvé un remède pour soulager la science et la mémoire. »

Et le roi répondit :

« Très ingénieux Theuth, tel homme est capable de créer les arts, et tel autre est à même de juger quel lot d'utilité ou de nocivité ils conféreront à ceux qui en feront usage. Et c'est ainsi que toi, père de l'écriture, tu lui attribues, par bienveillance, tout le contraire de ce qu'elle peut apporter.

Elle ne peut produire dans les âmes, en effet, que l'oubli de ce qu'elles savent en leur faisant négliger la mémoire. Parce qu'ils

> auront foi dans l'écriture, c'est par le dehors, par des empreintes étrangères, et non plus du dedans et du fond d'eux-mêmes, que les hommes chercheront à se ressouvenir. Tu as trouvé le remède, non point pour enrichir la mémoire, mais pour conserver les souvenirs qu'elle a. Tu donnes à tes disciples la présomption qu'ils ont la science, non la science elle-même. Quand ils auront, en effet, beaucoup appris sans maître, ils s'imagineront devenus très savants, et ils ne seront pour la plupart que des ignorants de commerce incommode, des savants imaginaires au lieu de vrais savants. »
>
> *Platon,* Phèdre

1.1 Organisation documentaire, rapport au savoir et transformation sociale

La prise de Tolède

En 1085, les troupes menées par Alphonse VI, roi de León, de Castille et de Galice, conquièrent Tolède sous domination Maure depuis le 8e siècle. Tolède est alors un centre culturel majeur du monde arabo-musulman. Après sa prise, l'occident chrétien découvre dans ses bibliothèques un grand nombre de textes qu'il ne connaît pas : tant des textes de philosophes arabes que de textes perdus de penseurs grecs. Une grande entreprise de copie et de traduction de ces textes de l'arabe vers le latin et les langues vulgaires se met alors en place, permettant la circulation de ces textes dans toute l'Europe. Ces textes sont étudiés et commentés, ce qui conduit à l'écriture de nouveaux textes qu'il faut à leur tour copier : la production documentaire croît de manière vertigineuse de la prise de Tolède à la Grande Peste au 14e siècle (Bozzolo & Ornato, 1980).

L'augmentation du nombre de manuscrits pose des problèmes nouveaux d'organisation. La question du repérage d'un manuscrit dans une bibliothèque ne se pose pas de la même manière lorsque le nombre de ceux-ci est décuplé. On voit apparaître dès le 12e siècle des formes de catalogage dans les bibliothèques. Gasparri (2009) note en effet que *« parmi les fonctions de l'armarius, le moine en charge de la bibliothèque, figurent aussi la mise*

à jour de l'inventaire des livres, leur examen et leur recension deux ou trois fois par an (ce qui est considérable), leur conservation dans des armoires saines, non surchargées : donc une réglementation stricte, ne laissant rien au hasard, laquelle n'a pas pu aller sans la réalisation d'un inventaire ou au moins de registres sommaires.»

Mais quand bien même les bibliothèques s'organisent pour permettre au lecteur de s'y retrouver dans des collections devenues bien plus vastes, ce dernier fait face à un problème majeur : il y a trop de livres pour qu'ils puissent tous être lus. L'érudit se trouve dans une situation de surcharge informationnelle qui va conduire au développement de tout un ensemble de dispositifs et de pratiques permettant d'appréhender cette abondance d'information. C'est explicitement en réponse à cette abondance que Pierre Lombard, au milieu du 12e siècle, justifie la démarche qui l'a conduit à l'écriture de son *Livre des Sentences*, recueil d'extraits des écrits des Pères de l'Église. Il explique qu'il souhaite rassembler *«en un volume court les avis des Pères [...] afin qu'il ne soit pas nécessaire de consulter une abondance de livres pour le chercheur, pour qui la brièveté des extraits rassemblés offre sans effort ce qu'il cherche»* (cité par Hamesse, 1995).

L'apparition de cette lecture fragmentaire est rendue possible par les florilèges et autres compilations à l'image du Livre des Sentences, mais pas seulement. Le mouvement d'organisation qu'ont connu les bibliothèques pour le repérage des manuscrits s'est décliné au sein même des livres : la structuration du texte en sections (avec des titres) et en paragraphes à fait son apparition (sous l'impulsion des Cisterciens notamment), permettant de compléter le texte par tables et index et donc de retrouver rapidement le passage recherché.

Pour aller encore plus loin dans cette approche non linéaire du texte, celui-ci se voit accompagné d'outils : en 1240, Hugues de Saint Cher, à la tête d'une armée de 500 moines et après près de 10 ans d'un travail acharné, termine la concordance de Saint Jacques : toutes les occurrences de tous les mots de la Bible sont relevées, triées et listées. On peut, en un seul coup d'œil, repérer tous les usages d'un mot (Bartko, 2003).

C'est ainsi l'activité même de lecture, sa pratique et ses objectifs qui sont bouleversés. La *ruminatio*, la lecture qui se pratiquait au Haut Moyen Âge, laisse place à la *lectura* (Hamesse, 1995) : alors qu'elle était une activité monacale, tournée vers la sagesse, la lecture devient une activité inscrite dans l'enseignement, tournée vers le savoir. De lente, articulée et méthodique, elle

devient silencieuse et fragmentaire. Il ne s'agit plus de lire méditativement un livre de la première à la dernière page (ils sont trop nombreux pour cela), mais de sauter d'un passage à l'autre, d'un livre à l'autre, de comparer et de commenter. C'est un nouveau rapport au savoir, la scolastique, qui émerge suite à la réintroduction soudaine et massive de la pensée grecque dans l'Europe médiévale. Pour la concilier avec la pensée chrétienne, on développe une approche critique : on commente, on explique, on débat. Les pratiques d'enseignement sont profondément changées également. Ce nouveau rapport au savoir est institutionnalisé : la connaissance quitte les monastères pour les universités naissantes.

Le 12e siècle pour éclairer le 21e

Le processus qui s'est enclenché suite à la prise de Tolède n'est pas sans rappeler celui que nous sommes en train de vivre. L'événement clé de notre époque, notre Tolède, c'est la création d'Internet, du Web, et plus généralement des technologies numériques. La production de documents, qu'ils soient textuels ou audio-visuels, est une activité devenue banale qui rythme le quotidien du plus grand nombre. Soudain, nous faisons face à une quantité de documents sans commune mesure avec ce que nous avions l'habitude de traiter. Les sociétés médiévales ont réagi en opérant un changement dans leur rapport au document. Elles ont modifié simultanément l'organisation des systèmes de documents (les bibliothèques) et l'organisation des documents eux-mêmes (avec l'introduction des sections, titres et paragraphes) et, par conséquent le rapport entre le lecteur et le document : la structuration du document a proposé de nouvelles affordances qui ont transformé l'expérience de lecture et le vécu du lecteur. Le rapport du lecteur au contenu du livre, et donc plus généralement le rapport du lecteur au savoir en a été transformé également. Ces transformations documentaires et épistémiques se sont accompagnées de transformations sociales sur la circulation et l'accès au savoir par instauration de nouvelles institutions. Le bouleversement a été majeur et a touché autant l'intime que le collectif.

Nous sommes à notre tour en train de vivre un tel changement, nos sociétés redéfinissant leurs rapports au document qui, devenu numérique et en réseau, appelle à de nouvelles formes d'organisation (Pédauque, 2006, 2007 ; Salaün, 2012). Notre rapport au document est transformé par des pratiques nouvelles. En s'intercalant entre les lecteurs et les documents, al-

gorithmes et statistiques offrent aux premiers des regards sur les seconds qui produisent des savoirs nouveaux dont la circulation prend des formes sans cesse renouvelées : open-access, réseaux sociaux... Comme à l'époque de l'apparition des premières universités médiévales, nous assistons à des transformations dans la transmission des savoirs. Les technologies numériques, en s'immisçant dans la pédagogie produisent des MOOCS, des jeux sérieux, etc., et reconfigurent les rôles des enseignants et des apprenants. À travers la redéfinition du rôle, de la place et du regard que portent nos sociétés sur les documents, c'est une redéfinition du rapport au savoir qui s'écrit. Et c'est précisément ce que sont les humanités numériques : la manière dont nous nommons ce nouveau rapport au savoir en train de se constituer. Les humanités numériques sont la scolastique d'aujourd'hui.

Big data, humanités numériques, data science, e-science, société de l'information, architecture de l'information, data journalism, e-learning... Tous ces termes s'inscrivent derrière l'idée que les technologies numériques ont provoqué une rupture, rupture qui touche aux manières dont nos sociétés organisent les connaissances, les produisent, les font circuler... L'apparition des technologies numériques marque un moment de bascule qui définit un avant et un après, et la profusion de ces termes témoigne d'un besoin, d'une nécessité, de nommer et décrire les caractéristiques de cet après.

Toutefois, si pointer le fait que nous sommes dans une phase de transition de régime documentaire peut nous mettre sur la voie du caractère révolutionnaire de notre époque, cela ne suffit pas. Il faut également expliciter les spécificités de cette transition pour qui la distingue d'autres transitions déclenchées par une hausse soudaine de la production documentaire ayant marqué l'Histoire : la transition médiévale des 11^{e} et 12^{e} siècles déjà discutée en est une, la citation de Platon en exergue montre que l'invention de l'écriture en est une autre, l'invention de l'imprimerie l'a été aussi et nous verrons dans la section 2.2 que la première moitié du 19^{e} siècle en est une autre. Pourtant, le discours dominant se cantonne principalement à pointer l'explosion informationnelle comme caractéristique du présent.

1.2 Ce qui n'est pas révolutionnaire

Le récit des Big Data

Le terme *big data* a particulièrement marqué la décennie passée. Il a quitté les cénacles industriels et académiques pour faire la une des magazines grand public. On a largement disserté sur la révolution en cours alimentée par ces données massives, commenté les fortunes tout aussi massives amassées par des firmes et leurs dirigeants qui ont su construire des services avec et prospérer sur ces données. On a largement alerté sur les dangers qui pèsent sur les libertés individuelles, ces données pouvant être utilisées à des fins de surveillance et de contrôle social.

Qu'est-ce qui confère à ces données leur caractère spécial ? Le mot lui-même, *big data*, semble donner réponse à cette question. Big. Ces données sont grosses, massives. Pour autant, la question reste entière : qu'est-ce que cela signifie pour des données d'être massives ? Par rapport à quoi sont-elles massives ? Qu'est-ce que cela signifie d'acquérir, de traiter, d'analyser beaucoup de données ? Et, surtout, qu'est-ce que cela change ?

Une large partie des définitions qui ont été proposées pour *big data* s'inscrivent dans une lignée initiée par Laney (2001), avant même que le terme ne se soit imposé[1]. L'accent est mis sur une tension entre les données auxquelles les entreprises doivent faire face et les outils à leur disposition pour les traiter : «*Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.*» Laney identifie trois caractéristiques de ces données, auxquelles il est fréquent de se référer par l'expression les *3 V* :

- Volume : C'est bien la première caractéristique de ces données. L'émergence du e-commerce dans les années 90 conduit à collecter des données dans des proportions qui dépassent de loin ce que le commerce traditionnel autorise : «*The lower cost of e-channels enables an enterprise to offer its goods or services to more individuals or trading partners, and up to 10x the quantity of data about an individual transaction may be collected thereby increasing the overall volume of data to be managed*». Pour faire face à cette avalanche de données, les en-

[1] Le terme Big Data semble naître dans la Silicon Valley dans la seconde moitié des années 90. Il apparaît dans les littératures industrielle et académique au tournant des années 2000 (Diebold, 2012)

treprises doivent mettre en place des infrastructures particulières et adopter des stratégies de gestion de leurs données.

- Velocity : Le cycle des données, de leur acquisition, leur traitement et leur analyse à leur réutilisation s'accélère. Là encore, des innovations technologiques sont nécessaires pour s'adapter : «*[r]ecognizing that data velocity management is much more than a physical bandwidth and protocol issue, enterprises are implementing architectural solutions*».

- Variety : Les systèmes de gestion de bases de données traditionnels, conçus pour emmagasiner des données au format strictement défini, sont mis à mal par des données de provenances diverses, aux formats variés et parfois faiblement structurées : «*[a]ttempts to resolve data variety issues must be approached as an ongoing endeavor*».

Sans dénier à Laney une certaine pertinence de son analyse, le choix d'allitérer les caractéristiques définitoires a certainement contribué à leur popularité. Ainsi a-t-on vu apparaître un 4ème V (Value), un 5 ème (Veracity), puis 6, 7, 10... Tom Shafer (2017) en recense 42.

Le noyau dur des définitions de *big data* reste le volume des données ou, plus précisément, une tension entre le volume des données et les capacités computationnelles disponibles pour les traiter. Le tableau 1.1 donne les versions successives de la première phrase d'article *big data*[2] dans la Wikipédia anglophone. Depuis la création de l'article, en 2010, c'est cette tension qui est systématiquement mise en avant. Le récit sur lequel la conception des Big Data est construite est le suivant : nous vivons une époque caractérisée par une explosion sans précédent de la production de données, tant et si bien que nos technologies pour traiter ces données, malgré leurs évolutions rapides, sont poussées dans leurs retranchements. L'étalon de mesure de la taille des données, ce qui permet d'affirmer qu'elles sont volumineuses, massives, est nos technologies : les *big data* sont trop grosses pour nos technologies. Ou du moins, à la limite de leurs capacités.

[2] https ://en.wikipedia.org/wiki/Big_data

Table 1.1 : *Historique de la première phrase de la page anglophone Big Data de Wikipédia. Les modifications mineures et les actes de vandalisme ont été exclus.*

Date de la première apparition	Définition
21 avril 2010	The term Big data from software engineering and computer science describes datasets that grow so large that they become awkward to work with using on-hand database management tools
27 octobre 2010	Big data are datasets that grow so large that they become awkward to work with using on-hand database management tools
08 janvier 2012	In information technology, big data consists of data sets that grow so large that they become awkward to work with using on-hand database management tools
28 mai 2012	In information technology, big data is a loosely-defined term used to describe data sets so large and complex that they become awkward to work with using on-hand database management tools
07 septembre 2012	In information technology, big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools
14 décembre 2012	In information technology, big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
04 avril 2013	Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
05 août 2013	Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
04 juin 2014	Big data is a blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications

Date de la première apparition	Définition
28 juillet 2014	Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications
09 septembre 2014	Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications
22 février 2015	Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate
01 mars 2016	Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate
12 septembre 2016	Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them
17 novembre 2017	Big data is data sets that are so voluminous and complex that traditional data-processing application software are inadequate to deal with them
28 août 2018	Big data is a term used to refer to the study and applications of data sets that are so big and complex that traditional data-processing application software are inadequate to deal with them
24 octobre 2018	Big data is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with
26 mars 2019	Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software
22 mars 2022	Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software
05 janvier 2023	Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software

Il est incontestable que le rythme actuel de la croissance de la quantité de données est faramineux. Les estimations de la quantité mondiale de données stockées au cours des dernières décennies montrent qu'elle suit une croissance exponentielle. La figure 1.1 illustre cette explosion en combinant des données construites avec des méthodologies différentes. Hilbert et López (2011) ont estimé l'évolution de la capacité totale de stockage entre 1986 et 2007 en analysant 25 technologies, tant analogiques que numériques. Ils parviennent à la conclusion que l'Humanité est passée de 2.6 exaoctets (Eo) [3] en 1986 à 295 Eo en 2007, soit un doublement tous les 40 mois. Le cabinet *International Data Corporation* (IDC) publie régulièrement depuis 2008 des rapports dans lesquels il estime la taille de l'*univers numérique* définit comme «*information that is either created, captured or replicated in digital form*» (IDC, 2008). En 2007, IDC évaluait la taille de l'univers numérique à 281 Eo (un nombre très proche de celui avancé par Hilbert et López (2011)), et à 59 000 Eo en 2020, soit 59 zettaoctets (Zo) : un doublement tous les 20 mois environ.

D'après le récit standard des *big data*, cette profusion de données se heurte aux limites de technologies qui n'ont pas été pensées pour de telles masses. Il a donc fallu concevoir de nouvelles générations de technologies (Chen, 2014; Dean & Ghemawat, 2004; White, 2012).

Les Big Data au regard de l'Histoire

Le rythme de production de connaissances

Le récit standard des Big Data affirme donc que le début du 21e siècle a pour caractéristique quasi définitoire d'être une époque marquée par la croissance exponentielle de la production de données qui stimule la création de nouvelles technologies. Mais est-ce vraiment propre à notre ère ? Qu'en est-il d'autres périodes de l'Histoire ?

La figure 1.2 présente le nombre de livres publiés par an et par habitant de 1500 à 1997 (Fink-Jensen, 2015). Bien que les échelles de temps soient différentes, sa similarité avec la figure 1.1 est flagrante. Au cours des cinq siècles passés, la quantité de "données", ou du moins d'information, mesurée par le nombre de livres publiés, a crû exponentiellement. La figure 1.3 montre,

[3] 1 exaoctet correspond à 10^{18} octets, soit 1 million de téraoctets, la taille typique d'un disque dur qui équipe en 2023 un ordinateur de milieu de gamme.

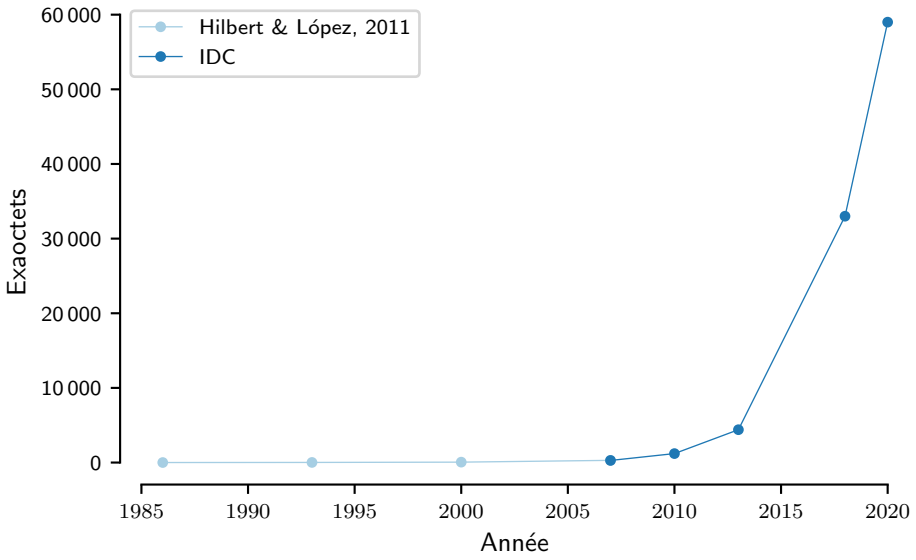


Figure 1.1 : *Estimation de la quantité de données produites en exaoctet de 1986 à 2020. Les données de 1986 à 2007 sont issues de Hilbert et López (2011), celles de 2007 à 2020 de IDC (2008, 2012, 2014, 2018, 2020).*

quant à elle, la quantité de manuscrits produits entre les 6^{e} et 15^{e} siècles (Buringh & Zanden, 2009). Le même phénomène est à l'œuvre, la même croissance exponentielle.

Le caractère exponentiel de la croissance de données n'est donc pas une spécificité de l'époque contemporaine due aux technologies numériques, mais au contraire une constante de l'histoire de l'Humanité, au moins depuis l'apparition de l'écriture si ce n'est depuis l'apparition du langage, liée au caractère cumulatif du savoir. La tension entre quantité de données et puissance de calcul pointée par les définitions standards des big data n'a rien de spécifique à l'époque contemporaine non plus. La quantité de données, informations ou connaissances gérées par l'Humanité a toujours été à la limite des technologies disponibles et a toujours été un moteur d'innovation technologique et de transformations sociales. C'est précisément ce que montre les suites de la prise de Tolède.

Néanmoins, c'est bien dans la masse des données que réside, par un effet de seuil, l'élément dont ce texte vise à démontrer qu'il est à l'origine de la révolution numérique, c'est-à-dire des transformations sociales en cours

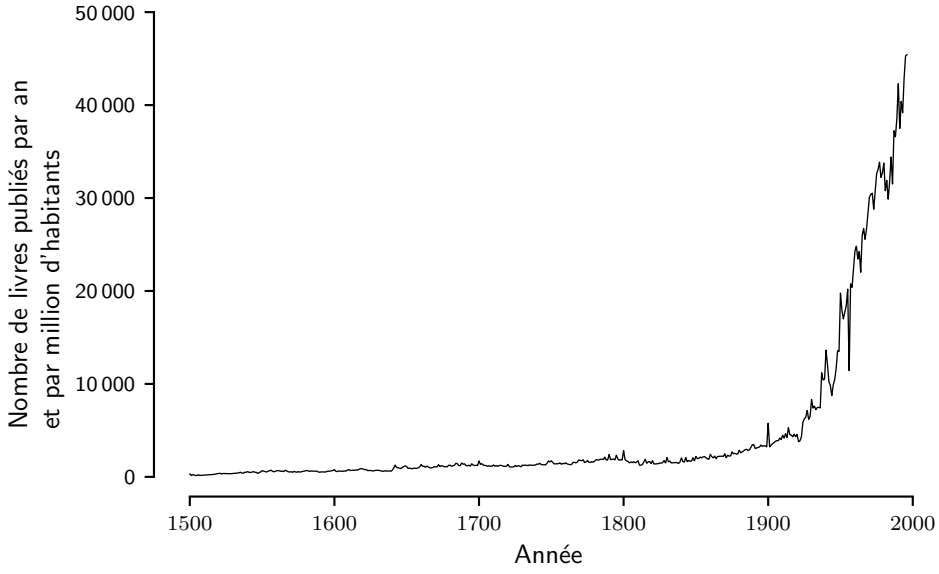


Figure 1.2 : *Nombre de livres publiés par an et par million d'habitants. D'après Fink-Jensen (2015).*

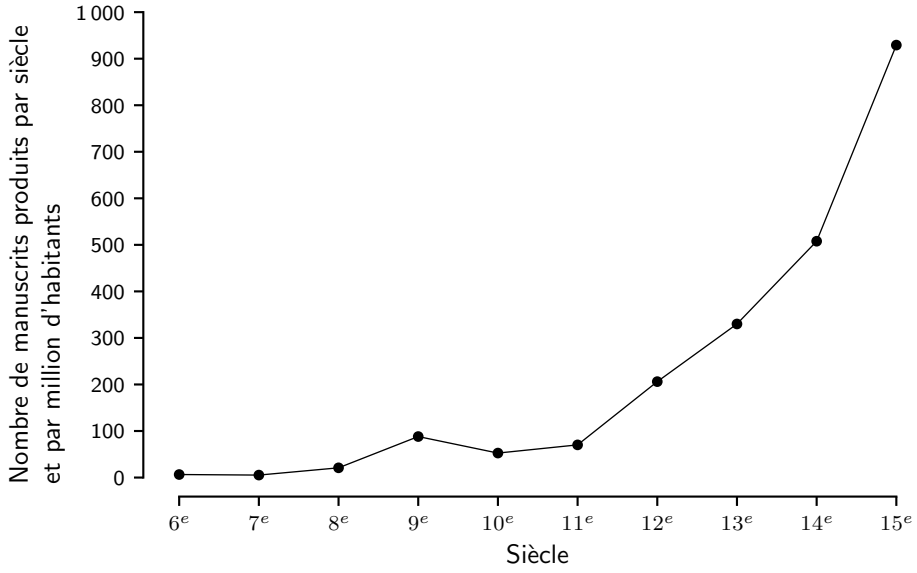


Figure 1.3 : *Production de manuscrits par siècle en Europe entre le 6e et le 15e siècle. D'après Buringh et Zanden (2009).*

provoquées par les technologies numériques. Pour pointer cet élément, un nouveau saut dans l'histoire s'impose.

1.3 Phénomènes circonscrits, phénomènes étalés

Le 18 mai 1896, à Paris, dans la salle des séances du Bureau des Longitudes, les directeurs des principaux instituts d'Europe et d'Amérique du Nord chargés de l'établissement des éphémérides étaient réunis à l'occasion d'une conférence dont l'objectif était l'établissement d'un standard international pour les constantes astronomiques. Parmi les présents, Simon Newcomb, alors Professeur à l'Université Johns Hopkins et directeur du *US Nautical Office* était déjà renommé pour la précision de ses calculs sur les positions apparentes des étoiles et les trajectoires des planètes. Et ce fut donc de manière très consensuelle que les calculs qu'il avait publiés l'année précédente (Newcomb, 1895) furent élevés au rang de standards par la recommandation de leur usage à l'échelle internationale.

La précision des prévisions de Newcomb tenait à la fois à la qualité de ses observations, à ses développements théoriques, mais également à la précision de ses calculs numériques (Campbell, 1924). Faire des calculs pour déterminer la position des étoiles ou des planètes à la fin du 19^{e} siècle recouvrait une réalité bien différente de la même activité au début du 21^{e} : point d'ordinateur auquel déléguer les fastidieux calculs qui devaient être conduits à la main. Des outils existaient cependant, au premier rang desquels le logarithme, introduit au 17^{e} siècle par Napier 1614. En transformant les multiplications en addition et les divisions en soustractions, il est d'un recours inestimable. Mais si aujourd'hui le logarithme se trouve à portée de clic, il fallait, jusqu'au milieu du 20^{e} siècle, consulter des tables : des livres dans lesquels étaient donnés les logarithmes des nombres de 1 à 10 000, voire de 1 à 100 000. Ces tables avaient une structure particulière. Les logarithmes de, par exemple, 4.275, 42.75, 427.5 et 4275 étant respectivement 0.630 936 119 1, 1.630 936 119 1, 2.630 936 119 1 et 2.630 936 119 1, il était naturel que ces 4 informations soient données ensemble, au même endroit. Dans ces tables, les nombres ne sont pas listés selon l'ordre numérique, mais selon un ordre apparenté à l'ordre alphabétique : l'ouvrage débute par les nombres commençant par le chiffre 1, puis ceux commençant par le chiffre 2...

En 1881, Simon Newcomb débute un article intitulé *« Note on the Frequency of Use of the Different Digits in Natural Numbers »* par la remarque suivante :

> *« That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithm tables, and noticing how much faster the first pages wear out than the last ones. »*
>
> (Newcomb, 1881)

Les pages des tables de logarithmes ne s'usent pas toutes de la même manière, témoignant du fait qu'elles sont plus souvent utilisées pour trouver le logarithme d'un nombre débutant par un 1 que d'un nombre commençant par un 2, d'un nombre débutant par un 2 que d'un nombre commençant par un 3...

Cette remarque ne suscita que peu d'intérêt et tomba rapidement dans l'oubli jusqu'à ce que, plus d'un demi-siècle plus tard, Frank Benford, un ingénieur de General Electric, propose un article intitulé *« The Law of Anomalous Numbers »* débutant par les mots suivants :

> *« It has been observed that the pages of a much used table of common logarithms show evidences of a selective use of the natural numbers. The pages containing the logarithms of the low numbers 1 and 2 are apt to be more stained and frayed by use than those of the higher numbers 8 and 9. Of course, no one could be expected to be greatly interested in the condition of a table of logarithms, but the matter may be considered more worthy of study when we recall that the table is used in the building up of our scientific, engineering, and general factual literature. There may be, in the relative cleanliness of the pages of a logarithm table, data on how we think and how we react when dealing with things that can be described by means of numbers. »*
>
> (Benford, 1938)

La dernière phrase de cette citation témoigne d'une intuition brillante. La démonstration de Benford de différence de fréquence des chiffres en

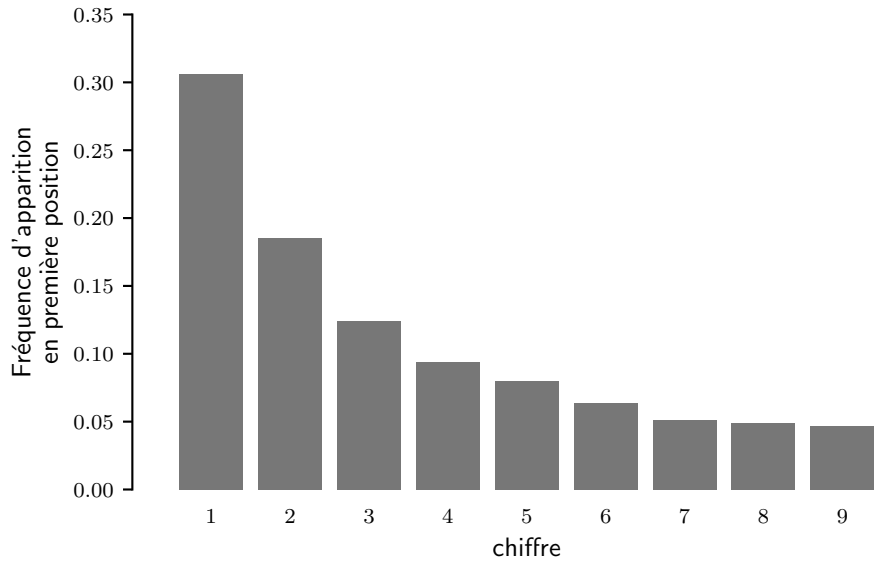


Figure 1.4 : *Fréquences d'apparition des chiffres en première position des nombres observées par Benford (1938)*

première position des nombres ne se limita pas au constat d'une usure variable des pages des tables de logarithme. Benford rassembla en effet 20 229 nombres aux origines variées comme des tailles de rivières, des populations de villes ou encore des taux de mortalité (voir table 1.2) La caractéristique commune de tous ces nombres était d'être produits par des phénomènes naturels. La fréquence d'apparition des chiffres en première position observée par Benford est donnée à la figure 1.4. Plus de 30% de ces nombres débutent par un 1, 18.5% par un 2, moins de 5% d'entre eux par un 9. Un ensemble de nombre est dit vérifier la *Loi de Benford* si la fréquence F_a d'observer le chiffre a en première position est :

$$F_a = log(\frac{a+1}{a})$$

L'observation de Benford est de prime abord étonnante : comment la nature peut-elle préférer à d'autres les nombres débutant par certains chiffres ? Pour tâcher de comprendre ce qui se joue ici, prenons deux exemples, un qui ne suit pas la loi de Benford alors que l'autre oui. Le premier de ces deux jeux de données a été rassemblé par Francis Galton (1886) et donne pour 930 individus, entre autres, la taille et la genre. Nous ne nous intéresserons ici qu'a la taille des hommes, exprimée en centimètres. Le second donne le nombre

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Table 1.2 : *Fréquences d'apparition des chiffres en première position des nombres observées par Benford (1938)*

de locuteurs de chacune des 6934 langues recensées par le *Summer Institute of Linguistics* (SIL) (Eberhard et al., 2022). Les fréquences d'apparition en première position des chiffres pour ces jeux de données sont indiquées à la figure 1.5.

Les tailles des individus ne suivent pas la loi de Benford. Sur les 481 hommes qui constituent ce jeu de données, un d'entre eux mesure plus 2 mètres (le premier chiffre de sa taille est donc 2) tandis que le reste mesure moins (le premier chiffre de leur taille étant donc 1). Le contraire aurait été surprenant : comment aurait-on pu avoir un ensemble d'individus dont les tailles commencent par des 3, des 4... ? Il aurait fallu que certains d'entre eux mesurent 30 ou 40 centimètres ou bien 3 ou 4 mètres. Changer d'uni-

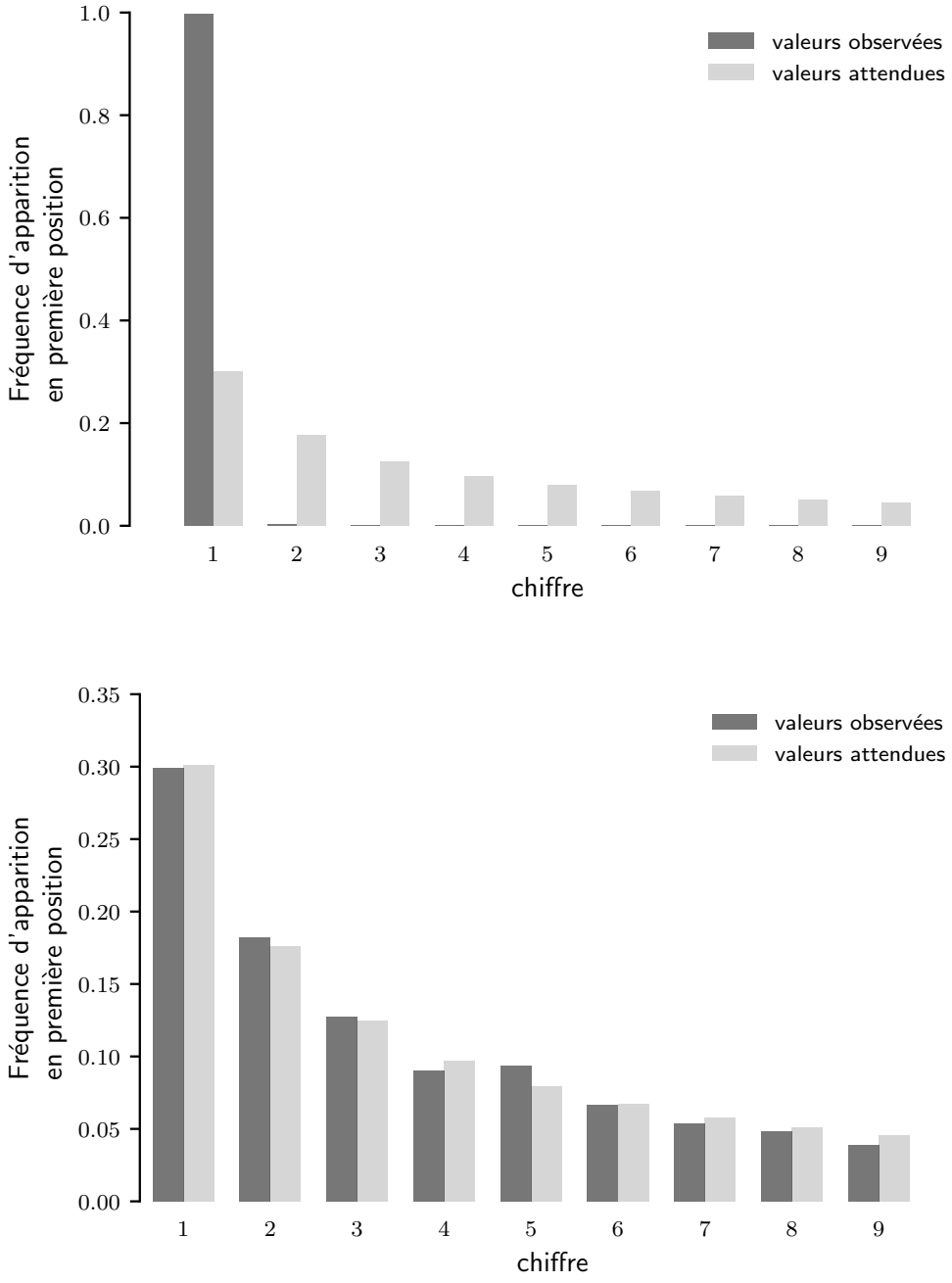


Figure 1.5 : *Fréquences d'apparition en première position des chiffres observées et attendues selon la loi de Benford pour les tailles d'individus relevées par Galton (1886) (en haut) et les nombres de locuteurs par langue donnés par Eberhard et al. (2022) (en bas)*

té ne change rien au problème. Dans les données originales de Galton, les tailles étaient données en pouces. Tous les hommes mesurent entre 60 et 79 pouces ; aucun ne fait 20 ou 200 pouces. La loi de Benford ne dépend pas du système d'unité choisi.

À l'inverse, le nombre de locuteurs par langue suit remarquablement fidèlement la loi de Benford. Un petit tiers des langues du monde ont un nombre de locuteurs commençant par 1 : on recense 31 langues comptant 1 locuteur ; 90 en comptant entre 10 et 19 ; 170 entre 100 et 199 ; 553 entre 1000 et 1999 ; 628 entre 10 000 et 19 999 ; 393 entre 100 000 et 199 999 ; 153 entre 1 000 000 et 1 999 999 ; 49 entre 10 000 000 et 19 999 999 ; 5 entre 10 000 000 et 19 999 999 ; et enfin 2 ayant plus d'un, mais moins de deux, milliards de locuteurs. Dans les deux tiers de langues restants, certaines ont un nombre de locuteurs commençant par 2 (environ 18% du total), par 3 (13%)... Peu de langues ont un nombre de locuteurs débutant par 9 (environ 3%), mais on en trouve.

L'examen de ces deux exemples conduit à une première conclusion : les tailles des individus ne peuvent pas suivre la loi de Benford car ces tailles sont trop similaires les unes aux autres, leurs premiers chiffres sont, peu ou prou, les mêmes (figure 1.6). Pour pouvoir exhiber la loi de Benford, les mesures d'un phénomène doivent pouvoir prendre des valeurs suffisamment différentes pour que les premiers chiffres puissent être 1, 2,..., 9. Autrement dit, ces valeurs doivent pouvoir s'étaler sur un ordre de grandeur, souvent plusieurs : dans le cas des langues du monde, les nombres de locuteurs s'étalent sur plus de 9 ordres de grandeur, des langues aux portes de la disparition, qui ne comptent plus qu'un locuteur, aux plus vivaces qui en comptent plus d'un milliard (figure 1.6).

L'étalement des données sur plusieurs ordres de grandeurs est étroitement lié à la loi de Benford, mais ne suffit pas à l'expliquer (pourquoi tous les chiffres en première position ne sont-ils pas équiprobables ?). Nous reviendrons sur ces liens et sur l'origine de cette loi au chapitre 4.

Cette discussion sur la loi de Benford et sur le fait qu'un jeu de données puisse la suivre ou non nous permet de distinguer deux types de phénomènes naturels. D'un côté ceux qui, à l'instar des tailles d'individus, sont circonscrits dans une gamme de valeurs restreintes, autour d'une valeur centrale. Ces phénomènes ont alors une échelle caractéristique, toutes leurs manifestations sont à peu près du même ordre de grandeur. Et de l'autre côté les phénomènes qui, comme le nombre de locuteurs des langues du

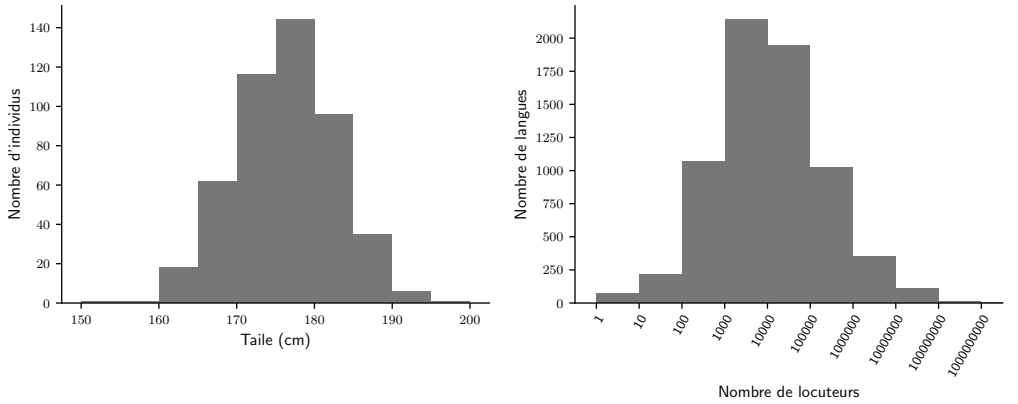


Figure 1.6 : *Distribution de la taille des individus d'après Galton (1886) (à gauche) et du nombre de locuteurs par langue donnés par Eberhard et al. (2022) (à droite, l'axe des abscisses est logarithmique)*

monde, n'ont pas d'échelle caractéristique mais s'étalent à travers plusieurs ordres de grandeur sans valeur centrale.

1.4 Les phénomènes étalés à l'origine de la révolution numérique

L'ensemble de l'argumentaire développé dans ce texte repose sur l'opposition entre ces deux types de phénomènes, circonscrits ou étalés, avec ou sans échelle ou, plus particulièrement sur notre capacité à les observer. Le point clé repose sur l'effort à fournir pour observer et caractériser ces deux types de phénomènes.

À la fin du 19^{e} siècle, le Royaume-Uni comptait 8 521 952 hommes de plus de 20 ans (University of Portsmouth, 2004). Galton en en mesurant 481, soit environ 0.005%, a obtenu une distribution relativement précise de la taille de ses concitoyens. A contrario, échantillonner 0.005% des 6934 langues du monde n'aurait pas de sens, l'échantillon ne contiendrait même pas une langue. La figure 1.7 montre les distributions obtenues en échantillonnant 0.05%, 0.5%, 5% et 50% des langues. On constate qu'il faut au moins 5% des langues pour que, visuellement, l'histogramme de l'échantillon ressemble à celui de la figure 1.6. Même les échantillons de 5% et 50% ne comprennent,

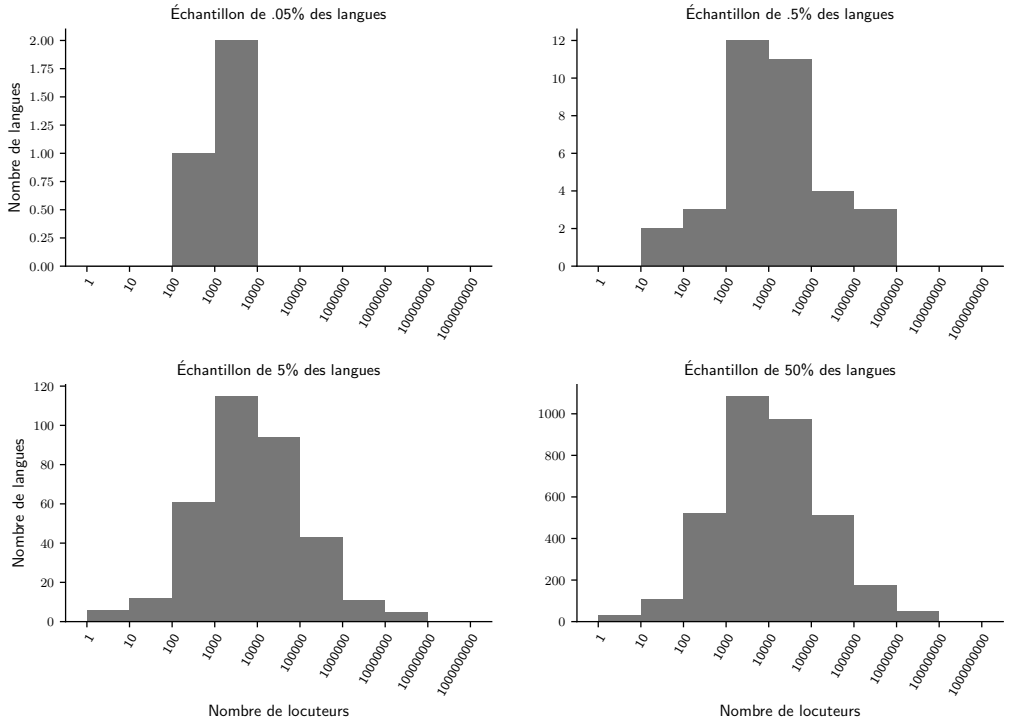


Figure 1.7 : *Distribution du nombre de locuteurs par langue à partir d'échantillons 0.05%, 0.5%, 5% et 50% des langues (d'après Eberhard et al. (2022), les axes des abscisses sont logarithmiques)*

dans cet exemple, aucune langue de plus d'un milliard de locuteurs. Ces échantillons ne sauraient être représentatifs : un discours sur les langues du monde construit sur échantillon qui ne contiendrait ni l'anglais ni le mandarin ne pourrait être qu'erroné.

Alors qu'un petit échantillon permet une représentation fidèle des phénomènes circonscrits, les phénomènes étalés nécessitent de grands échantillons, de l'ordre de la taille de la population. Par conséquent, l'observation et l'analyse des phénomènes sans échelle nécessitent des ressources bien plus grandes que l'observation et l'analyse des phénomènes avec échelle. Benford a mis des années pour rassembler son jeu de données. L'obtention du nombre de locuteurs par langue est le fruit de dizaines d'années de travail par des centaines de linguistes.

Cette différence entre les deux types de phénomènes étant établie, l'argument développé dans ce texte peut être résumé ainsi :

- Parce qu'ils sont difficiles à observer *à la main*, les phénomènes sans échelle sont longtemps restés invisibles, à l'exception de quelques cas perçus à l'époque comme anecdotiques. L'arsenal conceptuel développé par les statistiques des 18^{e} et 19^{e} siècles, au premier rang duquel la loi normale, a été largement construit pour les phénomènes avec échelle. Ces concepts et outils statistiques ont forgé une vision du monde, et notamment du monde social, dans laquelle la notion de catégorie est centrale. Par rétroaction, ces catégories ont eu un effet structurant sur les sociétés, les transformant en profondeur. Cet aspect constitue le chapitre 2.

- Outre les ressources nécessaires pour construire, manipuler et analyser des échantillons représentatifs des phénomènes sans échelle, ceux-ci nous sont invisibles pour des raisons plus profondes : notre cognition, et particulièrement nos systèmes de perception et de catégorisation ont été façonnés par la sélection naturelle pour appréhender les phénomènes avec échelle. Conceptualiser les phénomènes sans échelle nécessite donc un effort d'abstraction supplémentaire. Cela sera développé dans le chapitre 3.

- L'avènement des technologies numériques réduit considérablement le coût de l'observation et de l'analyse des phénomènes sans échelle. Là où le microscope nous a permis de découvrir le monde à petite échelle, le télescope à grande échelle, les technologies numériques sont un instrument nous permettant d'observer le monde à travers les échelles. Le chapitre 4 détaillera ces points et montrera qu'il en ressort que les phénomènes sans échelle, loin d'être anecdotiques, sont partout, omniprésents. Dans le chapitre 5, nous verrons que les technologies ne nous permettent pas seulement d'observer les phénomènes sans échelle, mais qu'elles œuvrent à les développer dans les sphères sociales. Il s'en dégage une nouvelle vision du monde qui, comme pour les statistiques, a un effet social rétroactif affaiblissant les systèmes de catégories structurant les sociétés pour laisser place à de nouvelles formes d'affirmations de l'identité.

2 L'essor de la statistique et sa rétroaction sociale

2.1 Le milieu qu'il faut prendre[1]

Lorsque Charles II, roi d'Espagne, meurt en 1700, il laisse son trône sans successeur conduisant les grandes puissances européennes à la guerre qui durera jusqu'en 1713. En octobre 1707, alors qu'ils sont en route pour rejoindre Portsmouth après avoir tenté sans succès de prendre le port de Toulon, 22 navires de la flotte britannique sont pris dans une tempête. Le 22 octobre, alors qu'ils se croyaient au large d'Ouessant, quatre de ces navires se fracassent sur les îles Sorlingues, au large des Cornouailles. Selon les estimations, entre 1400 et 2000 marins périrent. Outre le mauvais temps, il fut établi que la cause principale de ces naufrages était l'incapacité des marins à déterminer avec précision leur position.

Suite à cette tragédie, la dernière d'une longue série, en 1714, le parlement britannique vota une loi, le *Longitude Act*, instaurant les *Commissioners for the Discovery of the Longitude at Sea* et visant à promouvoir des solutions au problème de la détermination de la longitude, en offrant des prix à quiconque proposerait une méthode permettant de gagner en précision (Dunn et al., 2014). En effet, alors que l'estimation de la latitude est rendue possible par la mesure de l'élévation des astres, celle de la longitude est autrement ardue.

Dans cette quête de la longitude, plusieurs approches vont être suivies au cours des décennies suivantes. *In fine*, le problème sera résolu au début du 19^{e} siècle par le développement d'horloges suffisamment précises pour conserver sur les navires l'heure d'un méridien de référence, notamment celui de Greenwich, permettant ainsi de comparer l'observation de positions d'astres avec celles prévues au même instant à ce méridien de référence.

[1] Le titre de cette section est emprunté à Desrosières (1993)

Mais tout au long du 18e, c'est par l'observation de la Lune que de larges progrès seront réalisés.

Observer, se tromper, et observer encore

Tobias Mayer, un astronome allemand, fut de ceux qui firent gagner en précision la détermination de la longitude par la qualité de ses observations de la Lune. Il toucha d'ailleurs en 1765, à titre posthume, un prix de 3000 Livres Sterling. Mais si ses résultats eurent un impact certain, c'est par ses méthodes qu'il effectua un bond conceptuel radical. Depuis la révolution scientifique initiée, entre autres, par Copernic, Galilée, Bacon, Kepler ou Newton, c'est la combinaison de l'observation du monde, de sa mesure et de la raison qui rend possible sa compréhension. Or, les mesures répétées d'un même phénomène, incluant nécessairement une part d'erreur, diffèrent de l'une à l'autre et appellent donc à des méthodes pour traiter cette variation. De là naîtra la statistique. Avant le milieu du 18e siècle, cette variation dans les mesures inspirait méfiance, la crainte étant que ces erreurs se combinent et s'amplifient les unes les autres (Stigler, 1986). Ainsi le formule Euler en 1748 : *« [...]par la combinaison de deux ou plusieurs équations, les erreurs d'observation et du calcul se peuvent multiplier[...] »* (Académie Royale des sciences, 1769, p. 102).

La bascule fondamentale opérée par Mayer (1750) a été de comprendre que non seulement cette crainte d'amplification n'avait pas lieu d'être, mais surtout de comprendre que, au contraire, la répétition et la combinaison des observations est bénéfique : *«he approached his problem with the conviction that a combination of observations increased the accuracy of the result»* (Stigler, 1986, p. 28). Dans les décennies qui suivirent, le principe de combinaison de mesures s'est rapidement développé. Ainsi, en 1784, Jean Bernoulli III notait :

> *« Quand on a fait plusieurs observations d'un même phénomène, et que les résultats ne sont pas tout à fait d'accord entre eux, on a coutume alors de prendre le milieu entre tous les résultats, parce que de cette manière les différentes erreurs se répartissent également dans toutes les observations, l'erreur qui peut se trouver dans le résultat moyen devient aussi moyenne entre toutes les erreurs »*

(cité par Armatte, 2006)

La synthèse de Gauss-Laplace

Les méthodes employées par Mayer restent de l'ordre de la technique de calcul, dépourvues des justifications théoriques nécessaires. Au cours des dernières décennies du 18^{e}siècle, les travaux de mathématiciens comme de Moivre, Legendre ou Bernoulli posèrent les bases de cette justification. Elle fut pleinement achevée au début du 19^{e} siècle par les résultats Gauss et Laplace (Armatte, 2006 ; Hald, 1998 ; Stigler, 1986, 1999).

La contribution de Gauss (1809) fut de formuler la distribution suivie par les erreurs lors d'observations répétées :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.1}$$

L'année suivante, Laplace 1810 publia la démonstration du théorème central limite qui affirme que toute somme de variables aléatoires (comme, par exemple, des erreurs de mesure) converge vers la distribution établie par Gauss.

Observer et mesurer un phénomène, c'est intercaler entre lui et soi un ou plusieurs instruments et ainsi laisser place à de multiples incertitudes. Mesurer l'élévation dans le ciel d'une étoile au moyen d'un sextant nécessite de viser l'horizon, viser l'étoile et de lire l'angle sur les graduations. Chacune de ces étapes est source d'erreurs qui nous empêchent d'accéder à la vraie position de l'étoile. Ce que, de Mayer à Gauss et Laplace, la statistique naissante de la fin du 18^{e} et du début du 19^{e} nous a enseigné, c'est qu'en répétant ces mesures, celles-ci se distribuent comme prédit par la courbe de Gauss, la gaussienne. La moyenne de ces mesures nous donne alors accès à une vérité sur le phénomène observé, avec d'autant plus de précision que le nombre de mesures répétées sera grand. Établie sous le nom de *loi des erreurs*, cette construction intellectuelle est en soi remarquable (figure 2.1).

2.2 Le type moyen

Du relevé des récoltes dès le néolithique aux registres des baptêmes, des mariages et des décès ensuite (en France, depuis au moins l'Ordonnance de Villers-Cotterêts en 1529), les états ont besoin de connaître leur territoire et

Figure 2.1 : *Billet de dix Deutsche Marks à l'effigie de Gauss et de la gaussienne*

leur population pour les administrer. Quand le mot *Statistik* fut forgé en allemand au 18e siècle par l'économiste Gottfried Achenwall, il signifie science de l'état. Au début du 19e en France, en Angleterre, aux États-Unis et dans d'autres pays, la pratique s'institutionnalisa pleinement avec la création de structures administratives chargées de collecter et de publier des nombres décrivant les nations (Desrosières, 1993). Il en découla ce que Hacking (1982) appelle une *avalanche of numbers* : *«During the years 1820-1840 the rate of increase in the printing of numbers appears to be exponential [...]»*

Ainsi, lorsque Adolphe Quetelet, mathématicien de formation et astronome royal en Belgique, commença au cours des années 1820 à s'intéresser aux mécanismes qui régissent les sociétés, il avait à sa disposition tout le matériel nécessaire. Acquis à la puissance de la démarche scientifique, l'objectif de Quetelet était limpide :

> *«to lay the groundwork for a social physics, to conduct a rigorous, quantified investigation of the laws of society that might some day stand with astronomers' achievements of the previous century.»*
>
> (Stigler, 1986, p.170)

Ce que la loi des erreurs enseigne à Quetelet, c'est que si l'on mesure plusieurs fois la taille d'une même personne, on obtient des valeurs différentes, que celles-ci se distribuent selon l'équation établie par Gauss et qu'en prenant la moyenne de ces mesures on obtient une valeur dont on peut dire qu'elle est la taille de la personne. Mais quelle conclusion Quetelet peut-

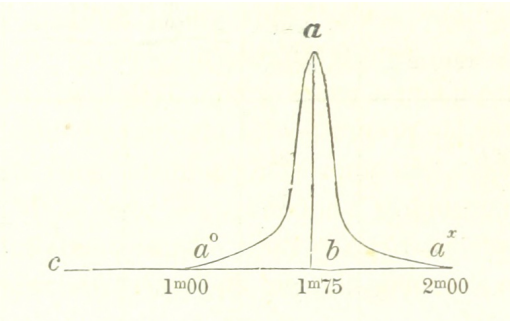


Figure 2.2 : *Distribution obtenue à partir de la taille de 25 878 volontaires aux États-Unis. D'après citequeteletAnthropometrieOuMesure1870*

il tirer lorsqu'après avoir mesuré *plusieurs personnes*, il constate que leur taille se distribuent également selon l'équation de Gauss (figure 2.2) ? En quoi l'écart de chaque taille à la moyenne des tailles est-elle une erreur ? L'avancée conceptuelle majeure que Quetelet a su réaliser en transposant la statistique naissante des sciences de la nature aux sciences de la société a été de transposer la notion d'erreur d'une mesure à l'autre à celle de variation d'un individu à l'autre. C'est un exemple typique de la puissance de l'interdisciplinarité. C'est parce qu'il avait un pied dans plusieurs communautés qu'il a pu faire les *analogies* nécessaires (Desrosières, 1993). Ce sont les personnes à la périphérie de plusieurs communautés, et non celles centrales à une seule, qui permettent la circulation des idées (Granovetter, 1973).

Pour pouvoir assimiler erreur et variation, Quetelet postula l'existence de ce qu'il appela l'*Homme moyen* qui abstrait les contingences individuelles et dont chaque individu dérive. Chaque aspect quantifiable des individus (comme la longueur des bras) et des populations (comme le taux de crime) contribue à cet Homme moyen. Quetelet en fait le concept central à partir duquel il tenta de construire une *physique sociale* qui décrit le fonctionnement des sociétés :

> *« Ainsi, les phénomènes moraux, quand on observe les masses, rentreraient en quelque sorte dans l'ordre des phénomènes physiques; et nous serions conduits à admettre comme principe fondamental dans les recherches de cette nature que plus le nombre d'individus que l'on observe est grand, plus les particularités individuelles, soit physiques, soit morales, s'effacent*

> *et laissent prédominer la série des faits généraux en vertu desquels la société existe et se conserve.* »
>
> (Quetelet, 1835, p.12)

Pour Quetelet, il n'y a pas un Homme moyen mais plusieurs, selon la nation considérée, le genre, le groupe d'âge...Chaque catégorie de population peut être appréhendée à travers son idéal type sur lequel s'applique le déterminisme de la physique sociale en ne laissant plus de place au libre arbitre :

> *« Tout procède d'année en année avec une constance et une régularité telles, que les effets des volontés individuelles peuvent être considérés comme à peu près complètement neutralisés. Les seules causes morales qui exercent une action sensible sur le cours des choses n'émanent plus des individus; elles appartiennent au peuple et à ses coutumes, dont les individus subissent à tout instant les influences comme autant de nécessité...quand on aura mieux reconnu la neutralisation des particularités individuelles (libre arbitre) dans la production des phénomènes sociaux, et la permanence des mêmes effets sous l'influence des mêmes causes, on sentira que la science sociale doit rentrer désormais dans les sciences de l'observation et en suivre toutes les phases.* »
>
> (Quetelet (1846), cité par Lottin (1911))

2.3 Mettre les gens dans des cases

Quantifier le fait social

Il y a une différence fondamentale entre mesurer un phénomène naturel comme peuvent le faire les astronomes ou mesurer un phénomène social comme s'y sont appliqués Quetelet ainsi que les sociologues et les démographes qui l'ont suivi. Dans le cas des phénomènes naturels, les catégories, par exemple les électrons ou les protons, précèdent l'observation et la mesure. Dans le cas des phénomènes sociaux, c'est l'observation statistique, scientifique, qui, par nécessité, crée la catégorie, par exemple les riches ou les pauvres. Desrosières (2013) propose d'utiliser deux verbes distincts,

selon que l'on applique l'analyse à des phénomènes naturels ou sociaux : mesurer la nature, quantifier la société. Plus précisément, *« quantifier, c'est convenir puis mesurer »*. Ce que l'on veut mesurer n'est pas donné d'avance et il faut donc se mettre d'accord *a priori* sur les contours du phénomène étudié (par exemple, la pauvreté), avant de pouvoir l'analyser. Toujours selon Desrosières (2013), *« L'usage du verbe quantifier attire l'attention sur la dimension, socialement et cognitivement créatrice, de cette activité. »* (p.11)

L'essor de la quantification des sociétés dès la première moitié du 19e siècle est donc accompagné de la création d'un vaste ensemble de catégories créant les observables analysés :

> *« As Frege taught us, you can't just print numbers. You must print numbers of objects falling under some concept or other. The avalanche of printed numbers brought with it a moraine of new concepts. »*
>
> (Hacking, 1982, p.292)

Cette opération de création de catégories n'est pas neutre. Elle n'est pas une simple opération technique mise en œuvre par les démographes et sociologues dans la construction de leurs discours scientifiques. Au contraire, elle produit une rétroaction sur la société observée : *« society and the statistics that measure and describe it are mutually constructed »* (Saetnan et al., 2010, p.1). Cette coconstruction est directement liée au fait que le développement de la statistique, en tant que discipline, est intimement lié au développement des états modernes (Foucault, 2004) et de leur besoin de connaître leur population pour construire leurs politiques :

> *« As layers of classification system become enfolded into a working infrastructure, the original political intervention becomes more and more firmly entrenched. In many cases, this leads to a naturalization of the political category, through a process of convergence. »*
>
> (Bowker et Star, 2008, p.196)

Les catégories socioprofessionnelles

L'analyse des catégories socioprofessionnelles conduites par Desrosières et Thévenot (2002) illustre cette réification et les rétroactions de l'observation des sociétés vers l'organisation des sociétés qui en découlent. Cette nomenclature des activités professionnelles en France est le fruit de l'évolution depuis le début du 19^{e}siècle de la description du monde socioprofessionnel par les statisticiens (Desrosières, 1977). Elle est formalisée par l'Institut National de la Statistique et des Études Économiques (INSEE) une première fois en 1954 (sous le nom de Catégories socioprofessionnelles, CSP) puis en 1982 (sous le nom de Professions et catégories socioprofessionnelles, PCS). Organisée hiérarchiquement, elle comporte des catégories comme, par exemple, *Cadres et professions intellectuelles supérieures* ou *Employés*. La nomenclature des activités professionnelles est initialement un système de représentations statistiques construit par les démographes.

Cette nomenclature est également un système de représentations politiques. L'État s'appuie sur ces catégories pour développer ses analyses du pays et par conséquent pour construire ses politiques publiques. À travers ces processus, ces catégories s'affranchissent d'un rôle purement descriptif. Ainsi, la catégorie *Cadre*, qui se construit à l'issue des grèves de 1936, s'institutionnalise autour de plusieurs syndicats et se voit à la Libération dotée d'un régime spécial de retraite (Boltanski, 1982).

Enfin, cette nomenclature est un système de représentations cognitives. À travers les discours politiques, ou encore les discours médiatiques ou publicitaires, ces catégories circulent dans l'espace public. Les citoyens construisent des représentations mentales de ce que représentent ces catégories, représentations qui peuvent diverger de l'observation statistique en étant idéalisées ou caricaturées. Dans une étude empirique, Desrosières et Thévenot (2002) ont demandé à des sujets de décrire en termes d'âge, de niveau d'étude, de revenus, etc., ce qu'ils et elles imaginent être un cadre typique :

> *« Les cadres choisis en exemple sont beaucoup plus diplômés [...], plus fréquemment parisiens, plus souvent dans la quarantaine, exerçant dans le secteur privé des fonctions commerciales, de marketing ou de publicité. Les PME sont sous-représentées, à l'inverse de grandes entreprises dont le renom est assuré (et dont le nom est souvent précisé par les participants). Le cadre*

Figure 2.3 : *Publicité pour la gamme d'appareils électroménagers Moulinex de 1960. La femme et l'homme représentés sont des stéréotypes de la ménagère et du cadre et l'ensemble véhicule de multiples injonctions sur les bons comportements à adopter pour revendiquer une appartenance à ces catégories.*

exemplaire est donc stylisé de la manière suivante : HEC, IBM, marketing, BMW. »

(p.57)

Ces représentations mentales influencent directement les comportements des gens. Elles sont ce à quoi chaque individu s'identifie ou s'oppose. Elles acquièrent un pouvoir normatif en indiquant comment il faut se comporter ou pas, ce qu'il faut posséder ou pas, pour justifier l'appartenance aux catégories dont on se revendique. La publicité est très largement basée sur ces mécanismes (figure 2.3). Le glissement des représentations de la statistique vers le cognitif établit une rétroaction entre l'observation de la société et sa structuration.

2.4 Francis Galton, stigmatisation sociale et eugénisme

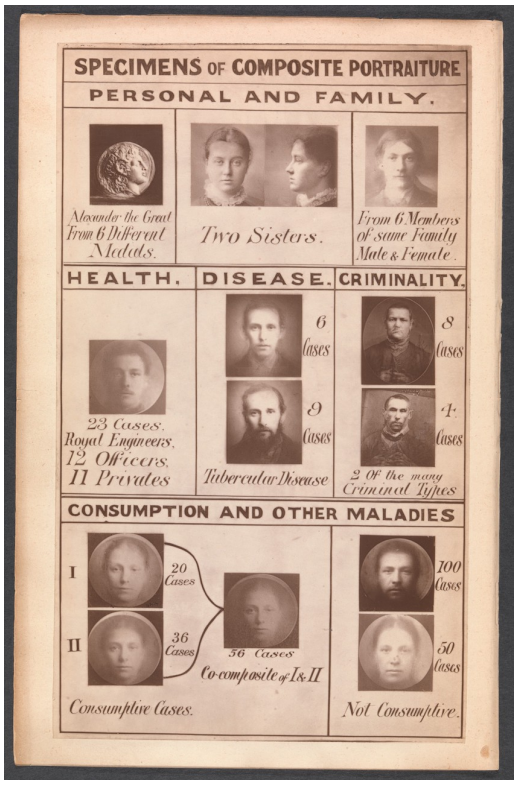
Le descendant le plus direct de l'entreprise anthropométrique de Quetelet de caractériser les populations humaines fut Francis Galton. Un des fondateurs des statistiques modernes par l'introduction de nombreux outils et concepts (régression, corrélation, déviation standard...), Galton nourrissait une admiration totale pour la distribution gaussienne :

> *« I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.* »
>
> (Galton, 1889, p.66)

À partir des années 1870, Galton s'est employé à développer une technique de composition de portraits en superposant plusieurs photographies sur la même plaque. Son objectif était de dégager le visage moyen et caractéristique de différentes catégories de personnes (Cryle & Stephens, 2017). Mais pas n'importe quelles catégories : parmi les premières auxquelles Galton s'attèle, on compte les criminels, les tuberculeux (figure 2.4a) ou encore les juifs (figure 2.4b).

Galton était le cousin de Darwin, et c'est par lui que la théorie de l'évolution a acquis un versant quantitatif et statistique, notamment par son étude de l'hérédité de diverses caractéristiques dans les populations humaines (Galton, 1889).

Mais quand Quetelet voyait dans la variation des fluctuations autour de la moyenne dont il fallait s'abstraire, Galton y voyait la source d'une possible évolution. Et quand Quetelet, à travers son *homme moyen*, voyait la



(a)

(b)

Figure 2.4 : *Portraits composites de Galton. a : d'après (Galton, 1883). b : d'après Pearson (1924).*

valeur idéale dans la moyenne, Galton voyait cette valeur idéale dans les extrêmes supérieurs. Le portrait composite de personnes en bonne santé de la figure 2.4a est construit à partir d'ingénieurs royaux et de soldats ayant en commun leurs *« bodily and mental qualifications required for admission into their select corps, and their generally British descent »* (Galton, 1883, p.14). Du portrait qui en résulte, Galton affirme que *« [t]his face and the qualities it connotes probably gives a clue to the direction in which the stock of the English race might most easily be improved. »* (Galton, 1883, p.14)

Galton fit de la théorie de Darwin un funeste projet politique, l'Eugénisme, dont il est à l'origine du nom :

> *« The aim of Eugenics is to bring as many influences as can be reasonably employed, to cause the useful classes in the community to contribute more than their proportion to the next generation »*
>
> (Galton, 1909, p.38)

2.5 La construction d'une normalité

Si les positions de Quetelet furent discutées, et même contestées, l'empreinte qu'il laissa est profonde et son influence eut des ramifications multiples : la physique statistique de Maxwell (Porter, 1986), la sociologie de Durkheim (Siracusa, 2017) ou encore la philosophie politique de Marx (Wells, 2017).

La généralisation de la loi des erreurs initie la prise de conscience que cette distribution est omniprésente. À partir des années 1870, sa dénomination sous le terme de *loi normale* se répand peu à peu (Stigler, 1999), d'abord sous les plumes de Peirce (1873), Lexis (1877) et Galton (1877). C'est finalement sous l'influence de Karl Pearson que le terme s'impose définitivement.

Ce que traduit l'adoption de l'adjectif *normal* est que si d'autres types de distributions sont parfois observées, celles-ci sont vues comme des déviations, qui nécessitent explication :

> *« If a series of measurements, physical, biological, anthropological, or economical, not of the same object, but of a group of objects of the same type or family, be made, and a curve be*

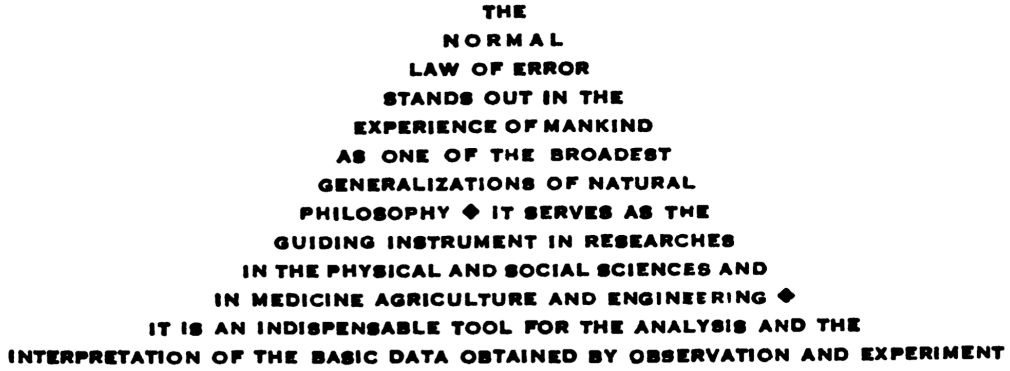


Figure 2.5 : *Calligramme conçu par Youden publié initialement dans Wallis et Roberts (1956).*

> *constructed by plotting up the number of times the measurements fall within a given small unit of range to the range, this curve may be termed a frequency curve. As a rule this frequency curve takes the well known form of the curve of errors, and such a curve may be termed a normal frequency curve. The latter curve is symmetrical about its maximum ordinate. Occasionally, however, frequency curves do not take the normal form, and are then generally, but not necessarily, asymmetrical. Such abnormal curves arise particularly in biological measurements ;[...]* »
>
> (Pearson, 1894, p.329)

Mais plus important encore, ce que traduit l'usage du substantif *loi*, est que cette distribution est élevée au rang de principe organisateur fondamental de la Nature (voir figure 2.5).

En tant que discipline, la statistique est née de la prise de conscience provoquée par le constat, répété tout au long du 19^{e} siècle, que nombre de phénomènes, si ce n'est tous, dans tous les domaines, présentent une variabilité similaire en se répartissant autour d'une valeur moyenne, d'une échelle caractéristique, selon l'équation proposée par Gauss.

Dans le cas des phénomènes sociaux, l'observation statistique des populations a nécessité leur structuration en catégories, lesquelles sont passées du statut de représentations statistiques, à celui de représentations politiques et à celui de représentations sociales. Parce que dans ces catégories

la variabilité se distribue autour d'une valeur moyenne, ou d'une manière plus générale, autour d'un individu typique, elles ont été l'objet du développement de formes d'essentialisme. Cet individu typique, ou du moins la représentation sociale de cet individu, fût-elle idéalisée ou caricaturale, est devenu pôle d'attraction ou de répulsion dans la constitution des identités. La découverte de la loi normale au 19^{e} siècle a modelé les sociétés du 20^{e}.

Des deux types de phénomènes, avec ou sans échelle, les premiers sont, à la fin du 19^{e} siècle, omniprésents tant dans les observations que dans les théories et les seconds totalement absents. La raison de cette invisibilité tient à deux facteurs : d'une part nos capacités cognitives sont telles que nous sommes directement aptes à percevoir et conceptualiser les phénomènes avec échelle, mais pas les phénomènes sans échelle; d'autre part, les technologies disponibles au 19^{e} siècle n'avaient pas la puissance requise pour observer et analyser les phénomènes sans échelle. Le prochain chapitre traite du premier point, le suivant du second.

3 Nos capacités perceptives et cognitives face aux phénomènes sans échelle

3.1 Perception des phénomènes sans échelle

L'ingénieur français Léon Lalanne est le premier à avoir proposé, en 1843, d'utiliser des échelles logarithmiques pour des représentations graphiques (Funkhouser, 1937). Comme pointé par Field (1917) :

> *«[T]he logarithmic method is peculiarly effective when the data are essentially relative; when they exhibit a tendency to increase or decrease at a fixed relative rate; or when significant proportionalities between different series of data are to be demonstrated. Incidentally it serves to economize space, and thus permits the inclusion of very diverse magnitudes in the same figure.»*
>
> (p.841)

Mais qu'est-ce qui fait que cette représentation est «*peculiarly effective*» ? Les figures 3.1 et 3.2 donnent, en fonction de leur rang une fois classées de la plus parlée à la moins parlée, le nombre de locuteurs des 6934 langues du monde (Eberhard et al., 2022). La figure 3.1 utilise une échelle linéaire, tandis que la figure 3.2 une échelle logarithmique. L'anglais est la langue la plus parlée, avec 1 452 471 410 locuteurs. De l'autre côté du spectre 14 langues n'avaient plus, en 2022, qu'un seul locuteur. Ce sont les cas les plus extrêmes parmi les 40 % des langues considérées en danger (à ce stade, le terme est un euphémisme). Neuf ordres de grandeur séparent les langues les moins parlées de celles les plus parlées, il n'y a pas de langue typique, moyenne ; le nombre de locuteurs par langue est un phénomène sans échelle.

Visuellement, sur la figure 3.1, à part les quelques centaines premières, ce sont plus de 90% des langues qui paraissent n'avoir peu ou pas de locu-

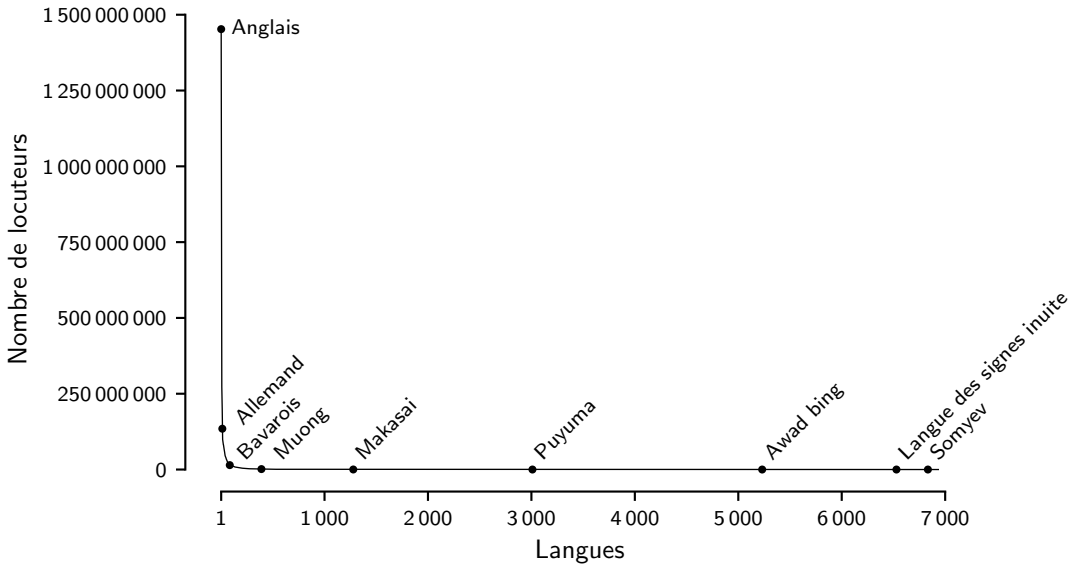


Figure 3.1 : *Représentation linéaire du nombre de locuteurs par langue (Eberhard et al., 2022). Outre l'anglais, les langues indiquées ont respectivement 10, 100, ..., 100 000 000 fois moins de locuteurs que l'anglais.*

teurs. Le muong, une langue du nord du Viêt Nam, le makasai, parlé au Timor oriental, le puyuma, à Taïwan, l'awad bing, en Papouasie-Nouvelle-Guinée, la langue des signes inuite et le somyev, parlé à la frontière entre le Nigeria et le Cameroun, semblent avoir le même nombre de locuteurs. Pourtant, si le muong possède 1000 fois moins de locuteurs que l'anglais, il en a 100 000 fois plus que le somyev. L'échelle linéaire utilisée sur la figure 3.1 ne rend pas compte de ces tailles relatives, alors que l'échelle logarithmique de la figure 3.2 le fait. Pourquoi ? Le principe de l'échelle linéaire est que, le muong étant 1000 fois moins parlé que l'anglais et 100 000 plus que la somyev, la distance entre l'axe des abscisses et le point représentant l'anglais devrait être 1000 fois plus grande que la distance entre cet axe et le point représentant le muong, qui devrait de même être 100 000 fois plus grande que la distance aux abscisses du point représentant le somyev. Sur un écran d'ordinateur [1], la distance entre deux points s'exprime par le nombre de pixels qui les séparent. Si la distance entre l'axe des abscisses et le point représentant le somyev était de 1 pixel, le point pour muong serait à 100 000 pixels de l'axe

[1] Les technologies d'impression rendent l'argument tout à fait équivalent sur papier.

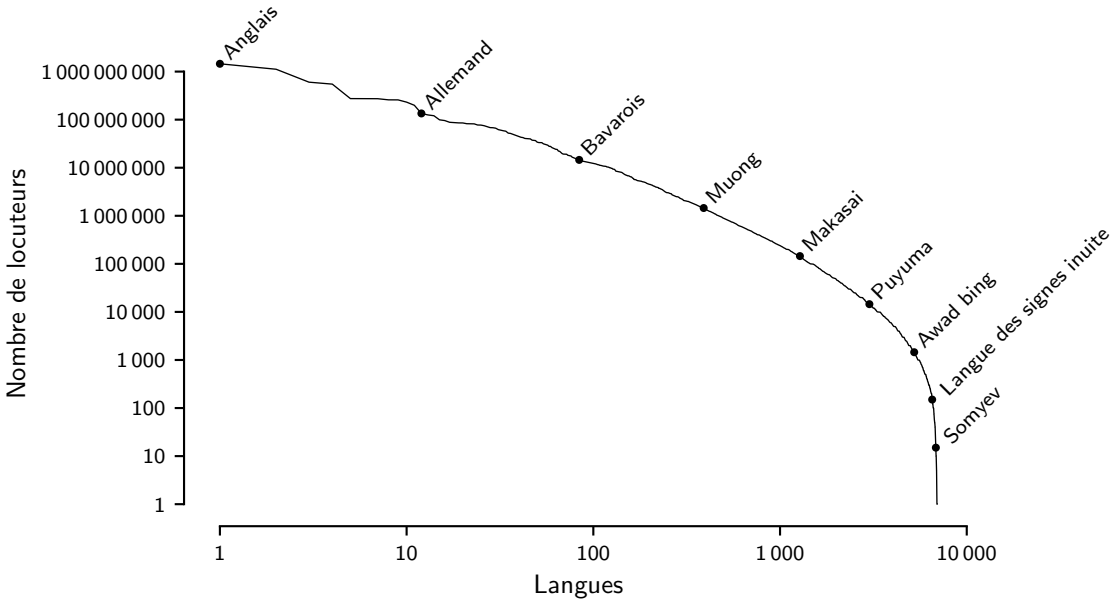


Figure 3.2 : *Représentation logarithmique du nombre de locuteurs par langue (Eberhard et al., 2022). Outre l'anglais, les langues indiquées ont respectivement 10, 100, ..., 100 000 000 fois moins de locuteurs que l'anglais.*

et le point pour l'anglais à 100 000 000 pixels. La norme d'écran la plus exigeante, le 8K UHD impose une résolution de 7680 pixels en hauteur par 4320 en largeur. Loin, très loin, des 100 000 000 pixels nécessaires dans notre exemple. Les écrans les plus performants ont une densité d'une centaine de pixels par centimètre. Avec une telle densité, pour pouvoir afficher avec une échelle linéaire l'ensemble de l'éventail de nombre de locuteurs des langues du monde, il serait nécessaire de disposer d'un écran de 1 000 000 cm, soit 10 km de haut.

Mais ces calculs masquent le fond du problème car, quand bien même nous aurions à notre disposition un écran de 10 km de haut, nous n'y verrions rien. Notre champ de vision maculaire mesure 18°. Pour qu'un écran de 10 km de haut y entre, il faudrait se tenir à une distance de 30 km. À une telle distance, le fin tracé de la courbe (0.1 mm si la courbe ne fait qu'un pixel d'épaisseur) nous serait imperceptible. On pourrait imaginer un tracé plus épais, pour qu'il soit visible : la résolution visuelle de notre œil étant de l'ordre d'une minute d'arc, il faudrait que le tracé fasse une dizaine de

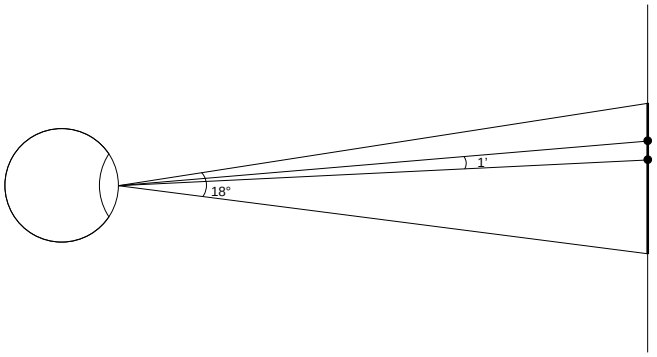


Figure 3.3 : *Champ visuel et résolution spatial de l'œil humain. La vision binoculaire humaine offre un champ de vision de 180° horizontalement. Sur une large périphérie de ce champ, nous ne sommes capables de discerner que des masses de couleurs. Au centre, le champ de vision maculaire, nous offre une vision plus précise, avec une résolution angulaire d'une minute d'arc, soit 1/60^{e} de degré.*

mètres d'épaisseur, soit 100 000 pixels. L'expérience perceptuelle face à un tel graphique serait alors sensiblement similaire à celle ressentie en regardant la figure 3.1 à une trentaine de centimètre. À quoi bon alors avoir un tel écran...

La limite fondamentale est notre œil. Avec un champ de vision de 18° et une résolution angulaire d'une minute d'arc, nous ne sommes capables de discerner que 18 × 60 = 1080 "positions" dans notre champ visuel (figure 3.3). Il nous est impossible dès lors de voir simultanément des objets dont les tailles différeraient de plus de trois ordres de grandeur.

Ces limites à nos capacités perceptives, illustrées ici par le pouvoir de résolution de notre vision, se retrouve dans toutes nos modalités sensorielles. Nous ne sommes capables de percevoir directement que des phénomènes qui s'étalent sur un nombre limité d'ordres de grandeur. Visuellement

toujours, notre œil est capable de détecter des différences d'intensités lumineuses de 1 à 5000, soit trois ordres de grandeurs, alors que les scènes naturelles peuvent contenir des rapports de 1 à 1 000 000, six ordres de grandeurs (Radonjić et al., 2011). Notre oreille est capable de percevoir des différences de fréquences de 20 Hz à 20 000 Hz, soit 3 ordres de grandeurs. Tactilement, les vibrations ne nous sont accessibles que jusqu'à 10 Hz.

Du point de vue des intensités sonores, les choses sont différentes. Nous sommes capables de les percevoir sur un spectre de l'ordre de 100 dB, soit un rapport de 1 à 10 000 000 000, ou 10 ordres de grandeur. Mais l'utilisation du décibel comme unité, défini à partir du logarithme de l'intensité de l'onde sonore, pointe une caractéristique de nos systèmes perceptifs qui freine encore notre appréciation de l'étalement d'un phénomène à travers les ordres de grandeur.

En psychophysique, la loi de Weber-Fechner relie la sensation subjective d'un stimulus à son intensité (Gescheider, 1997). Cette sensation, $\Psi(I)$, n'est pas linéaire mais proportionnelle au logarithme de l'intensité physique, I, du stimulus :

$$\Psi(I) = k \log(I)$$

Même lorsque l'intensité physique du stimulus peut s'étaler sur plusieurs ordres de grandeur, le fait que nos systèmes perceptifs en calculent le logarithme fait que notre expérience subjective reste dans un empan limité. Ces limites perceptives ne sont pas le seul obstacle à notre appréhension des phénomènes étalés. Nos capacités cognitives de catégorisation imposent des limites encore plus fondamentales.

3.2 Les propriétés de notre système de catégorisation

Conception classique des catégories

À observer notre environnement, un constat s'impose : bien que chaque objet qui le compose soit unique, il partage avec d'autres un certain nombre de similitudes qui nous conduisent à conclure qu'ils appartiennent à la même catégorie. Nous passons notre existence à faire et refaire ce même constat. Il nous est impossible de percevoir un objet ou un événement, d'accomplir une action ou de ressentir une émotion sans leur accorder une appartenance à

une catégorie. Il n'est pas d'entité qui n'appartienne à une catégorie. Cette systématicité a conduit la tradition philosophique occidentale depuis Aristote à réifier les catégories et à considérer l'appartenance d'un objet à une catégorie comme étant une propriété intrinsèque de l'objet, ou plutôt comme étant équivalente au fait que l'objet possède les propriétés caractéristiques de la catégorie. Selon cette conception dite classique des catégories, étant donné une catégorie CAT et un objet o, soit o possède la ou les propriétés caractéristiques de CAT et appartient de fait à CAT, soit il ne les possède pas et n'appartient pas à CAT. La possession des propriétés caractéristiques de CAT est une condition nécessaire et suffisante pour appartenir à CAT. Par exemple, étant donné la catégorie des objets pesant plus de 100 kg, un objet donné appartiendra à cette catégorie si et seulement si il a la propriété de peser plus de 100 kg.

Une des conséquences de cette conception des catégories est que la question de l'appartenance à une catégorie est une question propre aux objets et indépendante d'un quelconque observateur qui formulerait des jugements d'appartenance. Toujours selon cette conception, lorsque l'on dit que l'on catégorise un objet de telle ou telle manière, il y a abus de langage : on émet seulement une hypothèse, laquelle pouvant être soit vraie, si l'on a correctement identifié les propriétés caractéristiques de la catégorie et correctement vérifié leur présence chez l'objet, soit fausse si l'une des deux étapes n'a pas été correctement effectuée. Mais l'appartenance à proprement parler est antérieure à nos jugements, elle ne dépend que des propriétés de l'objet, et constitue une vérité absolue qu'il nous appartient d'atteindre en découvrant les propriétés nécessaires et suffisantes.

Les catégories peuvent, éventuellement, être organisées hiérarchiquement. L'exemple princeps est la manière dont Aristote définit l'être humain dans *De Anima*. La catégorie HUMAIN est une sous-catégorie de ANIMAL. Les propriétés caractéristiques d'une catégorie sont celles de la catégorie qui lui est immédiatement supérieure, plus un trait distinctif qui constitue la spécificité de la catégorie. Le trait distinctif qui distingue l'humain des autres animaux est sa rationalité[2]. La condition nécessaire et suffisante pour appartenir à la catégorie HUMAIN est d'appartenir à la catégorie ANIMAL *et* d'être rationnel.

Cette conception a prévalu jusqu'à la moitié du 20^{e} siècle sans être re-

[2]Nous reprenons ici le raisonnement d'Aristote sans pour autant y adhérer, en reconnaissant certaines formes de rationalité chez d'autres espèces animales que l'humain.

mise en question. Pourtant, depuis, cette conception a acquis le statut de théorie, peut donc être falsifiable et a été mise en défaut. Les premières brèches ont été ouvertes par Wittgenstein (1953) et les travaux de Rosch ont conduit à une théorie alternative, plus résistante à la vérification empirique, la théorie du prototype[3].

Ressemblance de famille

À deux reprises, le philosophe d'origine autrichienne Ludwig Wittgenstein a bouleversé la philosophie du langage (et même l'ensemble de notre système de pensée). Le premier de ces bouleversements est exposé dans son unique ouvrage publié de son vivant, le Tractatus Logico-Philosophicus Wittgenstein (1922). Cette monographie s'inscrit dans la révolution de la logique et de la formalisation des fondements des mathématiques opérée au début du 20e siècle, illustrée entre autres par Frege ou Russell. Le Tractatus Logico-Philosophicus se veut une réponse à la question *« en quoi les propositions de la logique se distinguent-elles de toutes les autres propositions du langage ? »* (Marconi, 1998). La réponse apportée par Wittgenstein l'a conduit à définir le sens d'une expression comme étant ses conditions de vérité, posant, ce faisant, les jalons de la sémantique formelle. Convaincu que toutes les questions philosophiques reposaient sur l'ambiguïté de leur formulation en langue naturelle et que ses travaux résolvaient ce problème, Wittgenstein considéra, après la publication du Tractatus Logico-Philosophicus, que toutes les questions philosophiques avaient trouvé réponses et que la philosophie était par conséquent achevée. Il se retira alors dans les Alpes autrichiennes où il devint instituteur, puis jardinier.

Le second bouleversement provoqué par Wittgenstein est une autocritique exemplaire, synthétisée dans son ouvrage posthume *Philosophical Investigations*. Toujours convaincu que les problèmes philosophiques naissent de la confusion de leur formulation, Wittgenstein y récuse le recours à la logique pour rendre compte de la signification des énoncés, argumentant que le sens des mots réside dans leurs usages. En développant cette position, Wittgenstein jette bas la conception classique des catégories. Illustrant son

[3] Si Wittgenstein et Rosch sont les plus représentatifs de la mise à mal de la conception classique des catégories, ils n'en sont pas pour autant les seuls. Voir Lakoff (1987) pour un historique détaillé de l'évolution des idées qui a conduit à la remise en cause de la conception classique des catégories.

propos avec la catégorie JEU, il part à la recherche des propriétés caractéristiques de cette catégorie, partagées par tout jeu :

> «*Consider, for example, the activities that we call "games". I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all? — Don't say: "They* must *have something in common, or they would not be called 'games'", but* look and see *whether there is anything common to all. — For if you look at them, you won't see something that is common to* all*, but similarities, affinities, and a whole series of them at that. To repeat: don't think, but look! — Look, for example, at board-games, with their various affinities. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball-games, much that is common is retained, but much is lost. — Are they all '*entertaining*'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball-games, there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck, and at the difference between skill in chess and skill in tennis. Think now of singing and dancing games; here we have the element of entertainment, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way, can see how similarities crop up and disappear.*
>
> *And the upshot of these considerations is: we see a complicated network of similarities overlapping and criss-crossing: similarities in the large and in the small.*»
>
> (Wittgenstein, 1953, §66)

À travers l'exemple de la catégorie JEU, Wittgenstein nous montre que l'hypothèse selon laquelle l'appartenance à une catégorie repose sur la vérification d'une condition nécessaire et suffisante, ou autrement dit, que tous les membres d'une catégorie partagent des propriétés en commun (et qu'ils sont les seuls), n'est pas fondée. Au contraire, il ressort de cette analyse que

si des similitudes sont bien observables entre les membres d'une catégorie telle que JEU, chacune d'entre elle ne concerne qu'une partie des membres de la catégorie. Le ciment qui unit les membres au sein d'une même catégorie ne peut être une ou plusieurs propriétés vérifiées simultanément par tous, mais plutôt un ensemble de propriétés, vérifiées chacune par un nombre variable de membres. Pour caractériser cette structuration des catégories, Wittgenstein introduit le terme de *ressemblances de famille* :

> *«I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family — build, features, colour of eyes, gait, temperament, and so on and so forth an overlap and criss-cross in the same way. — And I shall say: 'games' form a family.»*
>
> (Wittgenstein, 1953, §67)

Structure des catégories et structure des taxonomies

Si la contestation de Wittgenstein de la conception classique des catégories se situe sur un terrain philosophique, la psychologue américaine Eleanor Rosch a articulé la sienne sur le terrain empirique (Rosch, 1973, 1975a, 1975b, 1977, 1978; Rosch & Mervis, 1975; Rosch, Mervis et al., 1976; Rosch, Simpson & Miller, 1976). Selon la conception classique des catégories, la structure interne d'une catégorie est homogène : ses constituants en sont tous membres pour la même raison, en vertu du fait qu'ils vérifient les propriétés qui définissent l'appartenance à la catégorie. Il n'y a donc aucune raison pour qu'un ou plusieurs des éléments d'une catégorie se différencient des autres ou acquièrent un statut particulier; tous les membres d'une catégorie sont situés sur un même pied d'égalité. De la même manière, au sein d'une taxonomie, une hiérarchie de catégories, la conception classique ne privilégie aucun niveau par rapport aux autres : Milou est autant un FOX TERRIER, qu'un CHIEN, qu'un CANIDÉ, qu'un MAMMIFÈRE, qu'un VERTÉBRÉ ou qu'un ANIMAL. Ce sont ces deux prédictions de la théorie classique des catégories, l'homogénéité interne des catégories et l'homogénéité des taxonomies, que les travaux de Rosch ont falsifiées.

Structure interne des catégories

En multipliant les paradigmes expérimentaux, Rosch et ses collègues ont montré que la structure interne des catégories n'est pas homogène. Il existe au sein des catégories un gradient de typicité. Les membres (ou plutôt les sous-catégories) d'une catégorie ne sont pas tous également représentatifs de la catégorie, certains étant plus typiques que d'autres. Par exemple, dans la catégorie FRUIT, la pomme est un meilleur exemplaire que l'olive. Les paradigmes empiriques exhibant un gradient de typicité, ou effet de prototype, c'est-à-dire mettant en lumière l'hétérogénéité de la structure interne des catégories, sont variés et les résultats sont robustes d'un paradigme à l'autre : lorsque l'on demande aux sujets de produire des exemples d'une catégorie, les plus typiques sont plus fréquemment cités (Rosch, 1975b; Rosch, Simpson & Miller, 1976). Lorsque l'on demande d'évaluer la véracité de phrases telles que *la pomme est un fruit*, les temps de réponse sont d'autant plus courts que l'exemplaire est représentatif de la catégorie proposée (Rosch, 1973; Rosch, Simpson & Miller, 1976); lorsque l'on demande d'évaluer directement la typicité d'exemplaires d'une catégorie, les réponses des sujets sont hautement corrélées entre elles et avec les résultats d'expériences telles que celles décrites ci-dessus. Par ailleurs, un effet de prototype apparaît aussi dans les jugements de similarité entre membres d'une catégorie par une asymétrie des réponses : si a et b sont deux exemplaires d'une même catégorie et que a est plus typique que b, alors b est jugé plus similaire à a que a à b (Rosch, 1975b). De même, les propriétés d'un exemplaire ont d'autant plus tendance à être généralisées à l'ensemble de la catégorie que celui-ci est typique (Rosch, 1975a; Tversky & Gati, 1978; Tversky, 1977). Enfin, lors de l'apprentissage de catégories artificielles, les exemplaires les plus typiques sont les plus rapidement appris (Rosch, Simpson & Miller, 1976), ce qui est également le cas d'un point de vue développemental chez les enfants pour les catégories naturelles (Rosch, 1973).

L'effet de prototype que l'on observe dans les catégories peut être assorti d'un gradient d'appartenance. Alors que la conception classique des catégories prédit que la question de l'appartenance à une catégorie attend une réponse tranchée, ce n'est que rarement le cas. Labov (1973) a proposé à des sujets des dessins de vaisselle tels que ceux de la figure 3.4. Dans l'une des conditions de son expérience, il était simplement demandé aux sujets de nommer le dessin. Alors que l'ensemble des sujets s'accordaient à caté-

Figure 3.4 : *Exemples de stimuli utilisés par Labov (1973)*

goriser le stimulus 1 comme étant une CUP et qu'une majorité catégorisait le stimulus 4 comme étant un BOWL, les stimuli intermédiaires étaient nettement moins consensuels, indiquant que la frontière entre les catégories CUP et BOWL n'est pas franche comme le prédit la conception classique, mais graduelle, floue, certains objets n'étant ni totalement une tasse, ni totalement un bol.

Niveau de base des taxonomies

Le second point étudié par Rosch découle directement des observations de Wittgenstein : les catégories sont caractérisées par un faisceau de similitudes, un ensemble de propriétés plus ou moins partagées par leurs membres. Au sein d'une taxonomie, une catégorie possède, pourvu qu'elle ne soit pas la plus générale ou une des plus spécifiques, une catégorie superordonnée, la catégorie qui lui est immédiatement supérieure, et des catégories subordonnées, les catégories qui lui sont immédiatement inférieures. Ainsi, la catégorie JEU a comme catégorie superordonnée par exemple LOISIR, qui est aussi la catégorie superordonnée de catégories comme SPORT ou PHILATÉLIE. Les catégories subordonnées de JEU sont JEU DE CARTES, JEU DE BALLON...

Le faisceau de similitudes d'une catégorie est composé de ceux de ses catégories subordonnées. Ainsi, parmi les propriétés pouvant être possédées par les membres de la catégorie JEU figurent les propriétés pouvant être possédées par les membres de la catégorie JEU DE CARTES, celles pouvant être possédées par les membres de la catégorie JEU DE BALLON...Certaines des propriétés d'une catégorie, les plus générales, sont fréquentes dans l'ensemble de ses catégories subordonnées, alors que d'autres sont plus spécifiques à certaines. Par exemple, la propriété *divertissant* est possédée par beaucoup de jeux, quelle que soit leur catégorie subordonnée, alors que la propriété *demande de la réflexion* est plus fréquente dans la catégorie subordonnée JEU DE CARTES que dans la catégorie JEU DE BALLON.

Sur ces considérations, Rosch, Mervis et al. (1976) ont fait l'hypothèse

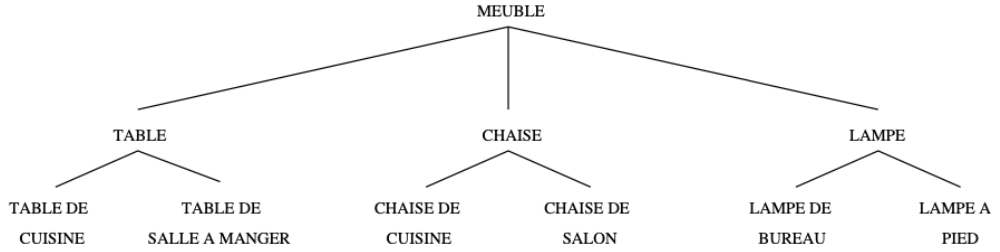


Figure 3.5 : *Une des neuf taxonomies étudiées par Rosch, Mervis et al. (1976)*

qu'il existe dans les taxonomies un niveau plus saillant que les autres et privilégié lors de la catégorisation. Ce niveau, le niveau de base, est celui auquel se réalise le compromis entre deux heuristiques guidant la catégorisation d'un objet : (1) faire en sorte que l'objet soit similaire aux autres membres de la catégorie et (2) qu'il soit différent des membres des catégories "contrastives" (les catégories contrastives d'une catégorie sont les autres catégories subordonnées de sa catégorie superordonnée). La figure 3.5 présente l'une des neuf taxonomies étudiées par Rosch, Mervis et al. (1976). Étant donné, par exemple, un stimulus représentant une table de cuisine, l'heuristique (1) pousse à le catégoriser dans la catégorie TABLE DE CUISINE, catégorie avec les membres de laquelle le stimulus partagera un grand nombre de propriétés. Mais cette catégorisation va à l'encontre de l'heuristique (2) : dans le cadre de notre exemple la catégorie contrastive de TABLE DE CUISINE est TABLE DE SALLE À MANGER. Or, une table de cuisine n'est pas fondamentalement différente d'une table de salle à manger. Pour satisfaire l'heuristique (2), il conviendrait de catégoriser le stimulus dans la catégorie MEUBLE. Ainsi serait maximisée la dissimilarité entre le stimulus et les membres des catégories contrastives de MEUBLE, des catégories d'artefacts différents de MEUBLE, tels que VÉHICULE ou OUTIL. Mais catégoriser le stimulus à ce niveau va à l'encontre de l'heuristique (1). L'équilibre entre ces deux tendances inverses se trouve à un niveau intermédiaire de la catégorie, le niveau de base. L'hypothèse de ce niveau de base a été vérifiée empiriquement de différentes manières.

Antérieurement aux travaux de Rosch, l'anthropologue Brent Berlin et ses collaborateurs avaient déjà proposé des éléments en faveur d'un niveau de base dans les taxonomies (Berlin et al., 1966, 1973, 1974). Ces travaux ont porté sur les classifications zoologique et botanique des locuteurs du tzeltal, langue parlée au Mexique dans la région du Chiapas. Berlin et ses collaborateurs ont montré que, bien que les tzeltal aient une connaissance poussée

des plantes et des animaux de leur environnement, en étant capables de distinguer les familles, les genres, les espèces et les variétés[4], le nom usuel pour dénommer une plante ou un animal est le nom correspondant au genre de la plante ou de l'animal. À ce niveau, la classification des tzeltal est largement concordante avec la classification scientifique, alors qu'aux autres niveaux les divergences se font plus grandes. De plus, le niveau du genre est le premier à être appris par les jeunes Tzeltal (Stross, 1969). Tous ces points supportent l'idée que le genre constitue le niveau de base des taxonomies biologiques des locuteurs du tzeltal.

Rosch, Mervis et al. (1976) ont, dans un article très riche relatant onze expériences, apporté les preuves empiriques de l'existence et de l'utilité cognitive d'un niveau de base dans les taxonomies. Confirmant la pertinence de l'heuristique (1), ils ont montré que le niveau de base est le plus abstrait des niveaux auxquels on observe une grande cooccurrence des propriétés des objets. C'est aussi le niveau le plus abstrait pour lequel on dispose de programmes moteurs permettant d'interagir avec l'ensemble des objets de la catégorie : alors que l'on a des programmes moteurs associés aux chaises en général, aucun n'est commun à l'ensemble des meubles. Quant aux programmes moteurs pour des catégories plus spécifiques, les chaises de cuisine ou les chaises de salon, ils ne diffèrent pas ou très peu de ceux des chaises en général. Il en va de même pour la similarité des formes des membres des catégories : le niveau de base est le plus abstrait auquel les membres présentent des formes similaires. Nous avons à ce niveau une perception gestaltique, c'est-à-dire que nous percevons les objets comme des touts psychologiquement plus saillants que la somme de leurs parts. Cela a pour conséquence que le niveau de base est le niveau le plus abstrait auquel il est possible de procéder à de l'imagerie mentale (Kosslyn et al., 1978 ; Shepard & Metzler, 1971). Le niveau de base est le niveau auquel s'effectue par défaut la catégorisation, le premier auquel un objet est reconnu comme étant membre d'une catégorie. C'est le niveau auquel apparaissent les premiers mots, tant du point de vue de l'acquisition des langues (les premiers mots appris par les enfants désignent les catégories situées au niveau de base), que du point de vue de l'évolution des langues : Rosch, Mervis et al. (1976) ont montré que le lexique de la Langue des Signes Américaine ne

[4] Les êtres vivants se décomposent en 5 règnes. Chacun de ces règnes est décomposé en embranchements, eux-mêmes décomposés en classes, puis en ordres, en familles, en genres, en espèces et finalement en variétés.

recouvre pas tous les niveaux des taxonomies et que le niveau de base est le plus lexicalisé. Cette observation suggère que les termes correspondant au niveau de base sont les premiers à entrer dans le lexique des langues, ceux des catégories superordonnées et subordonnées ne les enrichissant qu'ultérieurement. Enfin, les termes correspondant aux catégories situées au niveau de base ont tendance à être plus courts que ceux pour les autres catégories. Les catégories de niveau de base sont fondamentales à la cognition. Lakoff et Johnson (1998) retiennent quatre critères pour caractériser ces catégories de niveau de base :

- C'est le plus haut niveau auquel il nous est possible de former une image mentale représentant l'ensemble de la catégorie.
- C'est le plus haut niveau auquel nous avons une perception gestaltique.
- C'est le plus haut niveau auquel nous avons des programmes moteurs pour agir avec les objets de la catégorie.
- C'est le niveau auquel s'organise la plupart de notre savoir.

3.3 Nos capacités de catégorisation face aux phénomènes étalés

En nous dotant d'un mécanisme de catégorisation exhibant un effet de prototype, l'Évolution nous a façonné pour appréhender les phénomènes avec échelle, les phénomènes qui suivent une distribution gaussienne : le gradient de typicité nous permet d'avoir une appréciation mentale directe de l'écart à la moyenne. Il n'y a rien d'étonnant à cela : en tant qu'êtres humains, nous dévions plus ou moins d'une taille moyenne, d'une échelle caractéristique. C'est à cette échelle que nous appréhendons le monde, que se vit notre expérience du monde. Quand bien même nous sommes vulnérables à des virus ou des séismes, notre survie dépend avant tout d'évènements et de décisions qui se déroulent à notre échelle.

Il en résulte que lorsque nous sommes face à un phénomène sans échelle, il y a une incompatibilité fondamentale entre ce phénomène et nos capacités cognitives. Nous sommes forcés de voir une échelle caractéristique là où il n'y en a pas. Prenons trois exemples pour illustrer ce qui se joue.

Le premier sont les cours d'eau. Qu'ils soient mesurés par leur largeur ou leur débit, les cours d'eau présentent des dimensions très variables : quand le plus petit filet d'eau a un débit de 1 cm^3/s, l'Amazone charrie 219 000 m^3/s d'eau ; plus de 10^{11} fois plus. Pour se rendre compte de la différence, on peut imaginer qu'une personne qui serait 10^{11} fois plus grande qu'une personne moyenne mesurerait 170 millions de km, à peu près la distance de la Terre au Soleil. Comme l'illustre la figure 3.6, nous ne conceptualisons pas tous les cours d'eau au sein de la même catégorie, mais découpons ce continuum de taille en plusieurs catégories : RUISSEAU, RIVIÈRE, FLEUVE. Le deuxième exemple (figure 3.7) montre le même processus. Des plus petits hameaux comptant une poignée d'habitants aux plus grandes mégalopoles qui en comptent des dizaines de millions, les agglomérations s'étalent sur 7 ordres de grandeur. Là encore, nous découpons ce spectre en différentes catégories selon leur taille. Le troisième exemple (figure 3.8) montre que le même processus est encore à l'œuvre lorsque l'on considère des criques, des baies et des golfes dont les littoraux peuvent faire autant quelques mètres que plusieurs milliers de kilomètres.

Comment pourrait-il en être autrement ? Dans une catégorie qui englo-berait du hameau jusqu'à la mégalopole, ces cas extrêmes ne pourraient être qu'atypiques. Pour autant, si on s'accorde à dire qu'une personne d'1m30 ou de 2m30 est atypique, c'est beaucoup moins défendable pour les hameaux et les mégalopoles. Dit autrement, si cela a du sens de parler d'atypicité pour les valeurs extrêmes des phénomènes ayant une taille caractéristique, pour les phénomènes sans échelle qui n'ont pas de valeur moyenne autour de laquelle se distribue la variation, cela n'en a pas. Nos capacités cognitives for-çant un gradient de typicité dans nos catégories, nous sommes contraints de former des catégories plus resserrées, car *in fine* nous devons concep-tualiser certains fleuves, certaines villes ou certaines criques comme plus typiques que d'autres. C'est pourquoi certains fleuves se retrouvent affublés de l'adjectif *côtier* : ils sont par leur taille trop éloignés du prototype de la catégorie.

On pourrait objecter, à raison, que nous sommes capables de construire des catégories qui regroupent les manifestations d'un phénomène sans échelle. C'est le cas par exemple de la catégorie COURS D'EAU qui regroupe tous les cours d'eau, des plus petits aux plus grands. Mais cette catégorie, hyperonyme des catégories RUISSEAU, RIVIÈRE et FLEUVE, n'est pas le niveau de base. Ce n'est pas à travers elle que nous percevons les cours d'eau. Elle

(a) *Un ruisseau*

(b) *Une rivière*

(c) *Un fleuve*

Figure 3.6 : *Catégorisation des cours d'eau en fonction de leur taille*

(a) *Un hameau*

(b) *Un village*

(c) *Une ville*

Figure 3.7 : *Catégorisation des agglomérations en fonction de leur taille*

(a) *Une crique*

(b) *Une baie*

(c) *Un golfe*

Figure 3.8 : *Catégorisation des baies en fonction de leur taille*

n'est accessible qu'à travers un effort de conceptualisation et d'abstraction supplémentaire. Et d'ailleurs, pour nos deux autres exemples, on est bien en peine d'assigner une étiquette linguistique à la catégorie hyperonyme de HAMEAUX, VILLAGE et VILLE, ou à la catégorie hyperonyme de CRIQUE, BAIE et GOLFE. Nos niveaux de base, ceux à travers lesquels se passe notre expérience du monde, ne nous permettent pas d'appréhender les phénomènes sans échelle.

3.4 Observer les phénomènes sans échelle

La méthode hypothético-déductive, à la base de la révolution scientifique, consiste à observer un phénomène, proposer une théorie à son propos, formuler des hypothèses à partir de la théorie, confronter les hypothèses à l'observation, et, éventuellement, revoir la théorie. Si nos capacités cognitives ont été façonnées par l'Évolution pour permettre de voir et conceptualiser les phénomènes avec échelle et, qu'à l'inverse, elles nous masquent les phénomènes sans échelle, nous n'avons pas à être surpris que la Science se soit initialement concentrée sur l'observation et la théorisation des premiers.

Pour observer les phénomènes sans échelle, proposer des théories, formuler et surtout tester des hypothèses, nous ne pouvons pas nous contenter de nos sens. Il nous faut pour cela un instrument, comme le microscope qui permet d'observer à petite échelle ou le télescope à grande échelle. Mais il ne s'agit pas ici d'un instrument optique. Dans le prochain chapitre, nous allons voir que l'observation de phénomènes sans échelle passe par la manipulation de quantités de données qui dépassent, encore une fois, nos capacités cognitives. Dans le chapitre 1 nous nous demandions en quoi les *Big data* sont-elles *big* : elles sont *big* par rapport à nos capacités cognitives, parce qu'elles nous permettent de voir ce qui nous est autrement invisible.

4 La découverte des phénomènes sans échelle

4.1 La loi de Pareto

En 1895, l'anthropologue allemand Otto Ammon, dans son ouvrage *Die Gesellschaftsordnung und ihre natürlichen Grundlagen*[1], dédia tout un chapitre à une description de la répartition des richesses en développant une analyse statistique des revenus se basant sur les déclarations des contribuables du royaume de Saxe. Dès les premières lignes, il prit soin d'informer le lecteur de nécessaires précautions liées à une telle entreprise :

> *« La statistique est une belle chose; mais il est difficile de la comprendre correctement, plus difficile encore de l'employer avec intelligence. Par suite d'abus, la statistique est devenue suspecte, et on dit d'elle qu'elle est propre à démontrer tout ce qu'on veut. »*
>
> (p.177)

Commentant la figure montrant le nombre de contribuables touchant un revenu donné qu'il produisit à partir ces données (figure 4.1), il poursuivit :

> *« Pour faciliter la comparaison, j'ai reproduit dans la figure [4.1] la courbe de Galton au moyen d'une ligne pointée*[2]*.*
>
> *La courbe de Galton est établie sur la formule de probabilité de Gauss, et celle-ci s'applique à tous les groupes dans lesquels des cas particuliers s'écartent d'une moyenne et deviennent* d'autant plus rares que l'écart est plus accentué*. C'est le cas en général pour les revenus, par conséquent la courbe obtenue doit*

[1] *L'ordre social et ses bases naturelles*, selon la traduction de Muffang de 1900 dont sont issues les citations.

[2] Nommée courbe des aptitudes sur la figure.

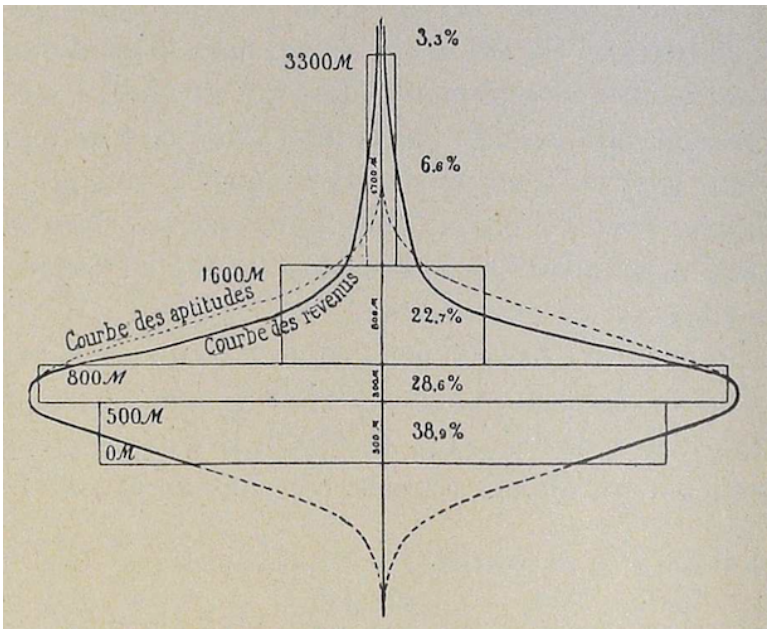
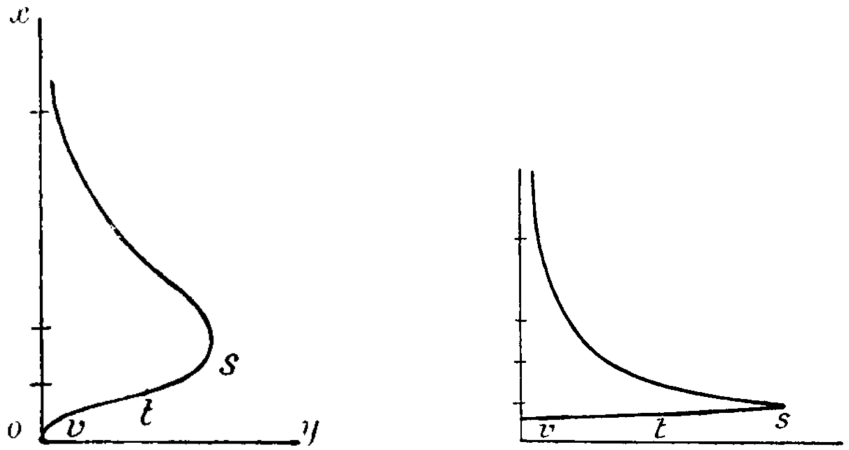


Figure 4.1 : *Répartition des revenus dans le royaume de Saxe en 1890 d'après Ammon (1895). Les revenus sont en ordonnée. En abscisse, on trouve vers la gauche le nombre de contribuables déclarant un revenu donné, vers la droite leur pourcentage par rapport à la population totale.*

> *se rapprocher de celle de Galton, sans qu'il faille y chercher un sens mystérieux et profond.*
>
> *Les légères différences entre la courbe des revenus et la courbe des aptitudes s'expliquent en partie par la méthode d'après laquelle sont recueillies les statistiques des revenus.* »
>
> (p.180-181)

Que ce soit parce que ses données étaient incomplètes, ou plus certainement parce qu'aveuglé par l'esprit de son temps (que nous avons décrit au chapitre 2), Ammon ne pouvait être plus dans l'erreur en plaquant la distribution des revenus sur la loi normale. Il sera contredit deux ans plus tard par l'économiste italien Vilfredo Pareto (1897) :

> « *A première vue, la courbe de la répartition des revenus ressemble à la courbe des probabilités, bien connue sous le nom*



(a) (b)

Figure 4.2 : *Les deux points de vue identifiés par Pareto (1897) sur la forme de la courbe de la répartition des revenus.*

> *de "courbe des erreurs". On pourrait donc supposer que la répartition des revenus est simplement l'effet du hasard [...]. Les riches auraient eu les gros lots. Il n'en est rien.* »
>
> (p.315)

À rebours de ses contemporains, Pareto reconnut que cette distribution diffèrait de la loi normale attendue par défaut :

> « *Certains auteurs, en se laissant guider par des conceptions théoriques, donnent à la partie inférieure de la courbe la forme s t v, Fig. [4.2a]. La statistique ne nous fournit aucune indication en ce sens. Il est donc fort probable que la partie s t v est très écrasée, et que la courbe réelle affecte une forme analogue à celle qu'indique la Fig. [4.2b].* »
>
> (p.314)

À travers son analyse de la distribution des revenus, Pareto mit en évidence pour la première fois un phénomène n'ayant pas d'échelle caractéristique. Contrairement à ce qu'affirmait Ammon (qui aurait dû mieux se plier à l'avertissement qu'il donnait de l'usage des statistiques), il n'y a pas de revenu moyen autour duquel les revenus des individus se répartissent, certains

étant un peu plus riches et d'autres un peu plus pauvres. Pareto fût même plus précis dans sa description de cette distribution :

> *« Traçons deux axes AB et AC. Sur AB portons les logarithmes de x [les revenus], sur AC les logarithmes de N [le nombre de contribuables ayant des revenus supérieurs à x]. Nous sommes tout de suite frappés du fait que les points ainsi déterminés, ont une tendance très marquée à se disposer en ligne droite. »*
>
> (p.304)

La distribution que Pareto exposa ainsi est une *loi de puissance*. Si x est le revenu et N le nombre d'individus ayant le revenu x, on a[3] :

$$N(x) = Ax^{-\alpha} \tag{4.1}$$

La différence entre l'analyse d'Ammon et celle de Pareto tient dans la formulation "*les cas particuliers [...] deviennent d'autant plus rares que l'écart est accentué*" utilisée par Ammon dans la citation de la page 4.1. Ammon conclue à une distribution des revenus qui suit une loi normale décrite par l'équation de Gauss (équation 2.1 page 25). Selon cette équation la fréquence d'observation des cas extrêmes décroît exponentiellement, donc très rapidement. À partir des données d'Ammon (table 4.1), on peut estimer que la distribution gaussienne qu'il a tracée présente une moyenne d'environ 740 Marks et un écart-type d'environ 450 Marks. Avec ces valeurs, la probabilité qu'il y ait un contribuable déclarant plus de 9600 Marks est de l'ordre de 10^{-88}, un nombre si petit qu'il est difficile de se le représenter. Cette probabilité est 10 000 milliards de milliards de fois plus petite que celle de gagner systématiquement au Loto en jouant un milliard de milliards de fois chaque nanoseconde pendant un milliard de milliards d'années (sachant que l'Univers n'est âgé que de 13.7 milliards d'années). Pourtant, toujours d'après les données recueillies par Ammon, 0.7 % des contribuables ont déclaré de tels revenus. La distribution normale décroît donc beaucoup trop rapidement pour rendre compte d'un phénomène tel que la distribution des revenus.

À l'inverse, la distribution de Pareto décroît beaucoup moins rapidement et permet de rendre compte de l'étalement des salaires sur plusieurs ordres

[3] Pareto s'intéressait à distribution *cumulative*, N étant le nombre d'individus ayant un revenu *supérieur ou égal* à x. Cela ne change rien à l'argument.

Tranche de revenu	Nombre de contribuables
moins de 500 Marks	546 138
de 500 à 800 Marks	401 439
de 800 à 1600 Marks	318 125
de 1600 à 3300 Marks	91 124
de 3300 à 9600 Marks	36 841
plus de 9600 Marks	10 402

Table 4.1 : *Nombre de contribuables par tranche de revenu dans le Royaume de Saxe en 1890. D'après Ammon (1895)*

de grandeur. Cet étalement ne touche pas que les salaires, mais leur fréquences également : si les salaires les plus hauts sont plusieurs ordres de grandeur plus élevés que les salaires les plus bas, ils sont aussi plusieurs ordres de grandeur plus rares que les salaires les plus bas.

Cette différence de vitesse de décroissance nous permet d'avancer une définition mathématique des phénomènes avec et sans échelle. Un phénomène a une échelle si sa distribution décroît au moins aussi vite qu'une exponentielle, maintenant ainsi les valeurs possibles autour d'une valeur typique, l'échelle du phénomène (l'exemple le plus représentatif étant la distribution gaussienne). Un phénomène est sans échelle si sa distribution décroît moins vite que l'exponentielle, permettant ainsi des valeurs possibles s'étendant sur plusieurs ordres de grandeur, sans valeur typique (l'exemple le plus représentatif étant la loi de puissance, mais d'autres entrent également dans ce cadre, comme la distribution log-normale, la distribution de Lévy...). Ces distributions sont dites *heavy tailed*, *fat tailed* ou encore *long tailed*, car leur queue, c'est-à-dire l'étendue des valeurs extrêmes, n'est pas négligeable à l'inverse des distributions décroissant exponentiellement (figure 4.3).

4.2 Les phénomènes sans échelle sont difficiles à observer

Les travaux pionniers

Pour mener à bien son analyse de la distribution des revenus, Pareto s'est appuyé sur des données des administrations fiscales de Grande-Bretagne,

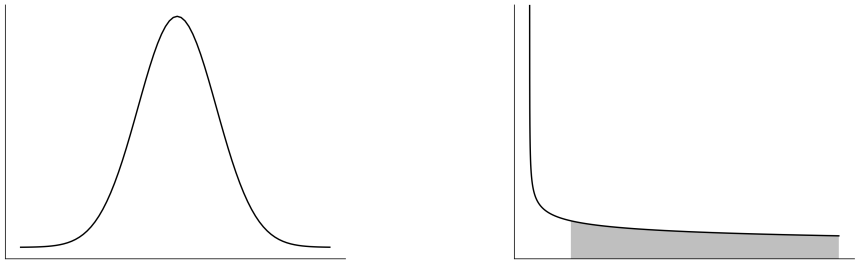


Figure 4.3 : *Comparaison des distributions de Gauss (a) et de Pareto (b). La distribution gaussienne atteint très rapidement des valeurs proches de 0 du fait de son exponentielle, alors que les distributions décroissant moins vite que l'exponentielle telles que la loi de Pareto présentent une longue queue (en gris).*

d'Irlande, de Prusse, de Saxe, du Canton de Vaud, de diverses villes italiennes... En tout, ce sont plusieurs dizaines de millions d'individus qui sont pris en compte dans ses analyses. En comptant le temps nécessaire aux contribuables pour remplir leur déclaration, aux employés des administrations fiscales pour les traiter et à Pareto pour les compiler et les analyser, ce sont, au bas mot, des milliers, sinon des dizaines de milliers d'heures qui ont été nécessaires.

Il n'aurait pas été possible de faire autrement. Si Pareto avait procédé comme Galton en échantillonnant 481 contribuables anglais (voir p.1.4), ce qui lui aurait pris quelques dizaines d'heures, il n'aurait pas pu arriver à ses conclusions car il n'aurait très probablement pas observé d'individus dans les tranches les plus élevées puisque ces individus ne représentent qu'une petite partie de la population (table 4.2). L'étalement à travers les ordres de grandeur des phénomènes sans échelle se trouve à la fois dans leurs manifestations, mais également dans les fréquences d'apparition de leurs manifestations. C'est bien parce que Pareto a pu être exhaustif, en s'appuyant sur ces données fiscales, qu'il a pu observer les cas les plus rares et mener à bout son raisonnement.

Observer des phénomènes sans échelle nécessite une telle énergie que rares sont les autres observations dans les décennies qui ont suivi le travail pionnier de Pareto.

En 1913, Auerbach, à partir des données de recensements de douze pays, a montré que la taille des villes est un phénomène sans échelle. Plus pré-

Tranche de revenu	Nombre de contribuables	Pourcentage de contribuables
de 150 £ à 200 £	400 648	37.68 %
de 200 £ à 300 £	234 185	22.03 %
de 300 £ à 400 £	121 996	11.47 %
de 400 £ à 500 £	74 041	6.96 %
de 500 £ à 600 £	54 419	5.12 %
de 600 £ à 700 £	42 072	3.96 %
de 700 £ à 800 £	34 269	3.22 %
de 800 £ à 900 £	29 311	2.76 %
de 900 £ à 1000 £	25 033	2.35 %
de 1000 £ à 2000 £	22 896	2.15 %
de 2000 £ à 3000 £	9880	0.93 %
de 3000 £ à 4000 £	6069	0.57 %
de 4000 £ à 5000 £	4161	0.39 %
de 5000 £ à 10 000 £	3081	0.29 %
10 000 £ et plus	1104	0.10 %

Table 4.2 : *Nombre de contribuables par tranche de revenu en Grande-Bretagne en 1893-1894. D'après Pareto (1897)*

cisément, après avoir trié les villes de chaque pays de la plus peuplée à la moins peuplée, il a montré que le produit de leur population par leur rang est constant.

En 1925, Yule s'est intéressé au nombre d'espèces par genre. Il a pour cela utilisé des données sur des plantes à fleurs (160 171 espèces réparties dans 12 571 genres), des chrysomèles (des coléoptères, 11 080 espèces réparties en 627 genres), les cérambycidés (également des coléoptères, 5719 espèces réparties en 1024 genres), des serpents (1475 espèces réparties en 293 genres), des lézards (1580 espèces réparties en 259 genres) et des légumineuses (9147 espèces réparties en 617 genres). Le travail des naturalistes sur lequel s'est appuyé Yule est considérable. Il a montré que le nombre d'espèces par genre est un phénomène sans échelle, une vaste majorité des genres ne comportant qu'une espèce, alors que les plus fournis peuvent en rassembler plusieurs centaines.

En 1926, Lotka a étudié la productivité scientifique des chimistes et physiciens en collectant *à la main* le nombre de fois que chaque auteur apparaissait dans deux bulletins bibliographiques : le *Decennial Index to Chemical*

Abstracts, 1907-1916, pour lequel il n'a traité que les lettres A et B, et le *Geschichtstafeln Der Physik*, un ouvrage d'Auerbach (1910) qui recense en 110 pages toute la littérature en physique de l'antiquité à 1900. En tout, ce sont 26 337 publications par 8216 auteurs qui sont passées au crible par Lotka qui montra que le nombre de publications par auteur suit une loi de puissance.

En 1935, Richter, s'appuyant sur la multiplication des installations de sismographes permettant d'enregistrer automatiquement les séismes, propose son échelle de magnitude pour caractériser les tremblements de terre. Cette échelle est logarithmique, indiquant que la puissance des séismes s'étale sur plusieurs ordres de grandeur. En 1949, Gutenberg et Richter, tout en reconnaissant que leurs données sont incomplètes (« *Listing is complete for magnitude 4 and higher; for magnitude 3.5, various estimates indicate that the count is roughly 20 per cent too small* ») démontrent que la fréquence des tremblements de terre en fonction de leur magnitude suit une loi de puissance, appelée loi de Gutenberg–Richter.

En 1942, Zipf a montré que la fréquence des mots d'un texte suit également une loi de puissance (figure 4.4). Il s'est pour cela basé sur *Word Index to James Joyce's Ulysses* (Hanley, 1937), un ouvrage qui donne pour chaque mot apparaissant dans le roman de Joyce la page et la ligne de l'ensemble de ses occurrences. Réaliser un tel index a nécessité 14 mois de travail à une équipe qui a compté jusqu'à 23 personnes. Zipf a montré que la fréquence d'un mot est inversement proportionnelle à son rang : comparé au mot le plus fréquent, le 2^{e} mot le plus fréquent apparait 2 fois moins souvent, le 3^{e} trois fois moins... Cette régularité est connue sous le nom de la loi de Zipf.

Cette liste n'est pas exhaustive, mais durant la première moitié du 20^{e} siècle l'observation de phénomènes sans échelle est restée marginale. Qu'en 1942 Zipf ait pu laisser son nom à la loi qu'il a mise au jour est symptomatique du fait que le domaine était encore alors peu défriché. La raison principale de cette lente prise de conscience de l'existence de phénomènes sans échelle au cours de la première moitié du 20^{e} siècle tient à la quantité de données nécessaires pour les observer et à la difficulté, en l'absence d'ordinateur, de les collecter et de les analyser.

Retour sur la loi de Benford

Benford, la loi est exposée au le chapitre 1, mérite sa place parmi les pionniers cités ci-dessus. S'il n'a pas directement mis en évidence l'étale-

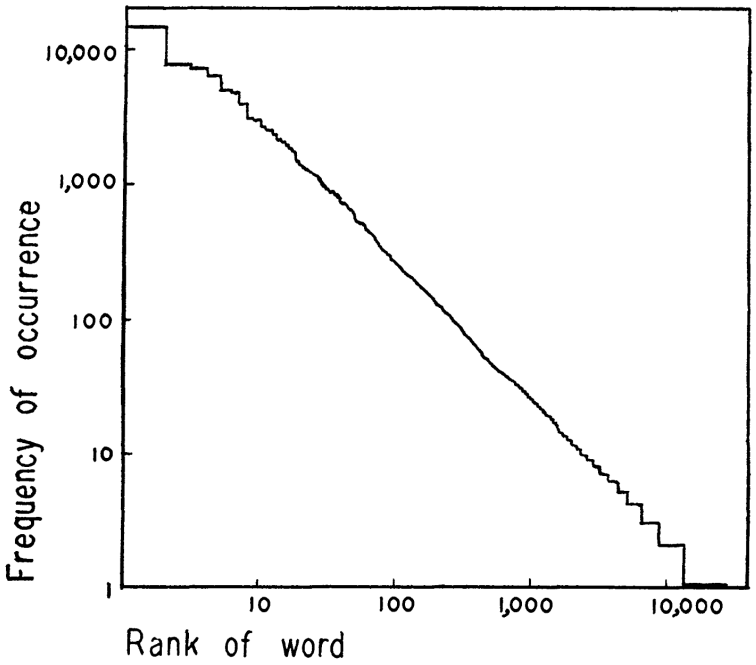


Figure 4.4 : *Fréquence des mots en fonction de leur rang dans* Ulysse *de James Joyce. D'après Zipf (1942).*

ment d'un phénomène spécifique à travers les échelles, son observation sur l'usure des pages des tables de logarithme et sur les fréquences d'apparition des chiffres en première position est directement liée aux phénomènes sans échelle : parmi les différents mécanismes pouvant conduire à la loi de Benford, on trouve en effet les phénomènes qui suivent une loi de puissance. Considérons un tel phénomène X :

$$P(X) = AX^{-\alpha}$$

avec $a \geq 1$. Si le premier chiffre de X est d, alors

$$d \cdot 10^k \leq X < (d+1) \cdot 10^k$$

Comme P(X) est strictement décroissante, on a

$$\begin{aligned} P(d \cdot 10^k) \leq P(X) < P((d+1) \cdot 10^k) \\ A(d \cdot 10^k)^{-\alpha} \leq P(X) < A((d+1) \cdot 10^k)^{-\alpha} \\ Ad^{-\alpha} \cdot 10^{k-\alpha} \leq P(X) < A(d+1)^{-\alpha} \cdot 10^{k-\alpha} \end{aligned}$$

La taille de l'intervalle dans lequel se trouve $P(X)$ est

$$\begin{aligned} &A(d+1)^{-\alpha} \cdot 10^{k-\alpha} - Ad^{-\alpha} \cdot 10^{k-\alpha} \\ =&A \cdot 10^{k-\alpha}((d+1)^{-\alpha} - d^{-\alpha}) \end{aligned}$$

Comme $\alpha \geq 1$, quel que soit k (c'est-à-dire quelle que soit l'échelle considérée) la taille de cet intervalle diminue quand d, le premier chiffre, augmente. Ainsi, lorsqu'un phénomène X suit une loi de puissance, on s'attend à observer plus souvent X avec comme premier chiffre 1 que 2, que 3..., ce qui correspond bien à l'observation de Benford.

4.3 Towards a revival of the statistical law of Pareto

Le titre de cette section est également celui d'un article publié en 1962 par Mandelbrot qui s'ouvrait par ces mots : *« Neglect and even contempt often mark the attitude of statisticians and of mathematical economists towards Pareto's well-known empirical discovery »*.

Le *revival* auquel appelle Mandelbrot a bien eu lieu dans les années qui suivirent. Il est directement lié à la généralisation à partir des années 1960 de l'usage de l'ordinateur apparu à la fin de la 2nde Guerre mondiale. On peut dégager deux domaines dans lesquels cet essor a pris racine avant de se répandre dans l'ensemble de la science : l'étude des transitions de phases en physique statistique et le développement des fractales par Mandelbrot.

Transition de phase et criticalité

Faire bouillir de l'eau est une expérience des plus banales. Alors qu'elle est liquide dans la casserole, quand la température atteint 100 °C (aux conditions standard de pression), les propriétés de l'eau changent radicalement et elle devient gazeuse : l'agitation des molécules qui a augmenté avec la température est telle que les molécules perdent la cohésion qui faisait d'elles un liquide pour devenir un gaz. C'est un exemple de transition de phase, de la phase liquide à la phase gazeuse. Ce phénomène est abrupt : en deçà de 100 °C l'eau est liquide, au-delà elle est gazeuse. Toutes les transitions de phases ne sont pas aussi abruptes. À une pression de 218 atmosphères, la

température d'ébullition de l'eau est de 374 °C. Ces valeurs correspondent au *point critique* de l'eau et à ce point critique l'ébullition n'est plus abrupte : liquide et gaz peuvent coexister et la transition se fait donc de manière continue. Les mécanismes à l'œuvre dans les transitions de phase de ce type, également appelés *phénomènes critiques* sont complexes et leur compréhension date des décennies 1960 et 1970. Ils ont valu à Kenneth Wilson le prix Nobel de physique en 1982, tant pour ses avancées théoriques que pour avoir instauré les méthodes computationnelles comme incontournables dans la démarche scientifique. Voici comment il s'exprime dans son discours lors de la remise du prix :

> *« There are a number of problems in science which have, as a common characteristic, that complex microscopic behavior underlies macroscopic effects.*
>
> *In simple cases the microscopic fluctuations average out when larger scales are considered, and the averaged quantities satisfy classical continuum equations. [...]Unfortunately, there is a much more difficult class of problems where fluctuations persist out to macroscopic wavelengths, and fluctuations on all intermediate length scales are important too. In this last category are the problems of fully developed turbulent fluid flow, critical phenomena, and elementary particle physics. [...] Theorists have difficulties with these problems because they involve very many coupled degrees of freedom. It takes many variables to characterize a turbulent flow or the state of a fluid near the critical point. [...]Computers can extend the capabilities of theorists[...] »*
>
> (Wilson, 1983)

Le modèle d'Ising est le modèle le plus utilisé pour étudier les propriétés des phénomènes critiques. Il a initialement été développé pour modéliser le ferromagnétisme (Ising, 1925). À une température suffisamment basse, un matériau ferromagnétique (comme par exemple le fer à température ambiante) possède une aimantation spontanée. Lorsque que ce matériau est chauffé au-dessus d'une certaine température critique T_C, spécifique à chaque matériau, il perd son aimantation spontanée. Il présente donc une

transition de phase. Le mécanisme sous-jacent est dû à l'orientation magnétique, ou *spin*, de chaque atome constituant le matériau. Le spin de chaque atome peut être dans deux états possibles et tend à s'aligner sur celui de ses voisins tout en pouvant, de temps à autre, changer spontanément de direction. Si les spins des atomes sont majoritairement orientés dans la même direction, l'ensemble du matériau présente une aimantation. C'est le cas si la température est en deçà de T_C, où les interactions entre spins voisins prennent le pas sur les changements spontanés. Mais lorsque la température dépasse T_C, l'agitation devient telle que les interactions entre spins voisins ne sont plus assez fortes face aux changements spontanés pour maintenir une orientation globale. Le matériau perd alors son aimantation. Le modèle d'Ising modélise ce phénomène.

Dans sa version la plus simple en deux dimensions, le modèle d'Ising représente un système défini comme une grille régulière sur laquelle chaque site a un spin σ_i pouvant être dans deux états possibles, $+1$ ou -1 (figure 4.5).

L'énergie totale d'une configuration σ de l'ensemble des spins est définie par :

$$H(\sigma) = -\sum_{\langle ij \rangle} \sigma_i \sigma_j \tag{4.2}$$

où $\langle ij \rangle$ représente une paire de voisins. La probabilité que le système se trouve dans une configuration σ dépend de l'énergie $H(\sigma)$ de la configuration et de la température T :

$$p_T(\sigma) \propto e^{\frac{H(\sigma)}{T}} \tag{4.3}$$

La figure 4.6[4] représente le magnétisme global du système en fonction de la température. On y voit apparaître la transition de phase à la température $T_C = 2.27$

La figure 4.7 présente l'état final du système à l'issue de trois simulations à trois températures différentes. Lorsque la température est inférieure à la température critique (figure 4.7a), à quelques fluctuations près, le système est complétement ordonné, tous les spins ayant convergé vers la même valeur (ici $+1$, mais dans une autre simulation ils auraient pu converger vers

[4] L'ensemble des simulations du modèle d'Ising ont été réalisées sur les serveurs du Centre Blaise Pascal de l'ENS de Lyon (Quemener & Corvellec, 2013) avec les méthodes proposées par Komura et Okabe (2016)

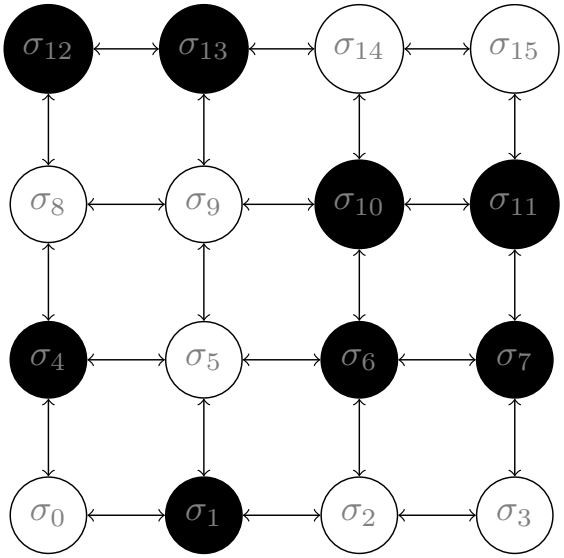


Figure 4.5 : *Modèle d'Ising sur une grille de 4 par 4. Chaque site peut être dans un état $\sigma_i = +1$ (représenté en noir) ou $\sigma_i = -1$ (représenté en blanc).*

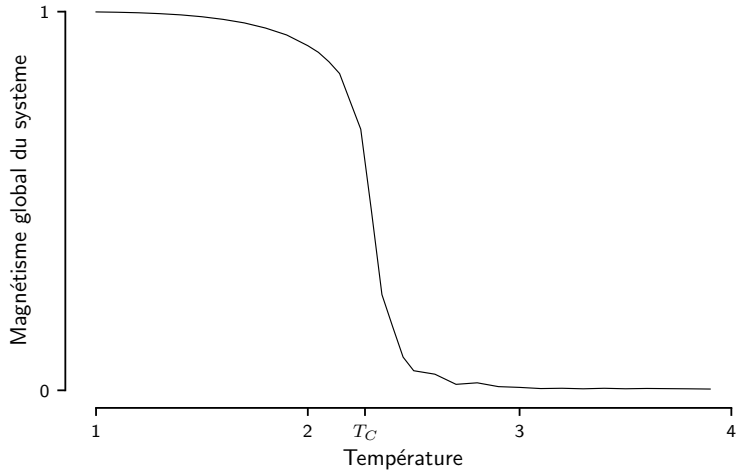
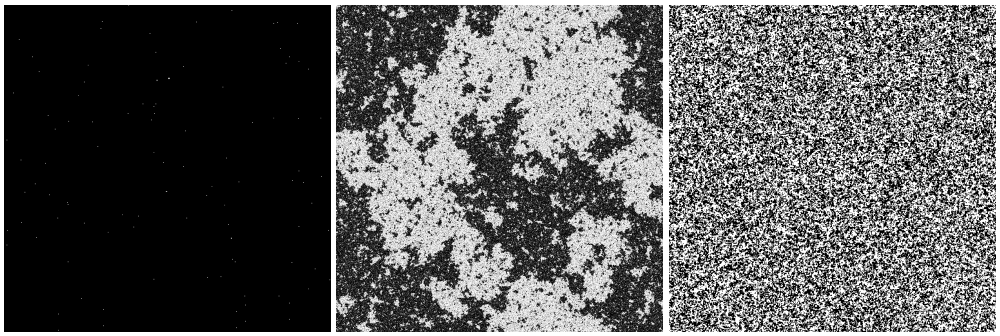


Figure 4.6 : *Transition de phase du magnétisme à la température $T_C = 2.7$ dans le modèle d'Ising.*



(a) (b) (c)

Figure 4.7 : *Résultat de trois simulations du modèle d'Ising aux températures $T = 0$ (a), $T = T_C$ (b) et $T = 4$ (c).*

-1). Lorsque la température est supérieure à la température critique (figure 4.7c), le système est complétement désordonné, les basculements spontanés ayant pris le pas sur les interactions entre voisins. À la température critique (figure 4.7b), le système est dans un état très particulier. Il est composé d'un patchwork de domaines où les spins sont dans le même état, soit -1, soit $+1$.

La figure 4.8 présente des zooms successifs sur différentes parties de la figure 4.7b. À chaque échelle, le même patchwork apparait : les domaines n'ont pas de taille caractéristique, c'est donc un phénomène sans échelle. On observe des mécanismes similaires dans de nombreux phénomènes, appelés *phénomènes critiques*. Par exemple, lorsque l'eau atteint son point critique, elle devient un entremêlâ de "bulles" liquides et de "bulles" gazeuses de toutes tailles.

Lorsque ces phénomènes critiques furent découverts, les méthodes de la physique ne permettaient pas leur analyse. Les transitions de phase étaient abordées par la théorie des champs moyens qui, pour reprendre les mots de Wilson cités page 4.3, se fonde sur l'hypothèse que *« microscopic fluctuations average out when larger scales are considered »*. Cette hypothèse est vérifiée lorsque les systèmes se trouvent loin de leur point critique. En revanche, à proximité de ce point, là où toujours avec les mots de Wilson, *« fluctuations persist out to macroscopic wavelengths, and fluctuations on all intermediate length scales are important too »*, la théorie des champs moyens ne s'applique plus. Il a donc fallu que les physiciens développent entre les années 1950 et 1970 une nouvelle approche théorique, les groupes

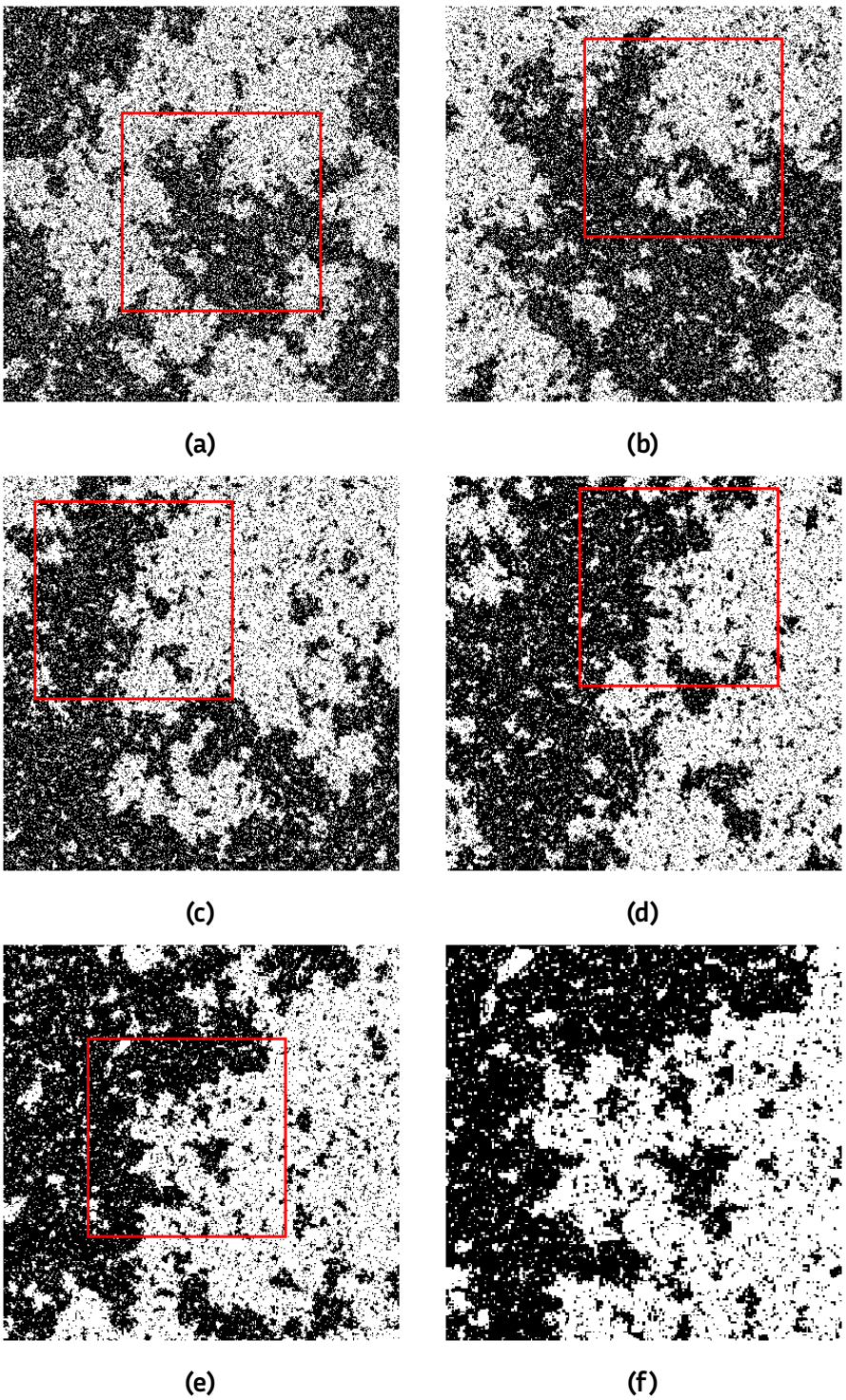


Figure 4.8 : *Résultat d'une simulation du modèle d'Ising à la température critique. Le rectangle rouge sur chaque image délimite la portion visible sur l'image suivante.*

de renormalisation, qui permette de traiter ces fluctuations à toutes les échelles. Ces développements, les plus majeurs de la physique de la 2nde partie du 20e siècle et couronnés par le prix Nobel de Wilson, ont placé les phénomènes sans échelle au cœur de la science.

Fractales

En parallèle des recherches des physiciens sur les phénomènes critiques, les phénomènes sans échelle gagnent aussi en intérêt à travers les travaux de Mandelbrot sur les fractales. Mandelbrot, surnommé *the father of long tails*(Mandelbrot, 2012), a commencé à s'intéresser aux phénomènes sans échelle à partir des années 1950 en reprenant et généralisant les travaux de Zipf (Mandelbrot, 1951, 1966).

Dans les années 1960, alors qu'il travaillait pour IBM, Mandelbrot s'intéressa aux marchés financiers et en particulier aux variations des cours du coton. À l'époque, la théorie dominante pour modéliser les cours était celle proposée par Bachelier (1900) qui supposait que les fluctuations des prix suivent une distribution gaussienne et sont indépendantes les unes des autres, c'est-à-dire que les fluctuations passées n'influencent pas les fluctuations futures. *« Thanks to the computer, [Mandelbrot] was able to note the aws in Bachelier's model »* (Mandelbrot, 2012, chapitre 22) : Mandelbrot observa empiriquement que les grandes fluctuations sont bien plus fréquentes que ne le prévoyait cette théorie et montra que les variations du cours du coton ne suivent pas une loi gaussienne mais une loi de puissance. Il montra également que les variations passées influencent les variations futures, à court terme comme à long terme : non seulement les fluctuations ne sont pas indépendantes, mais elles dépendent des fluctuations passées sur toutes les échelles de temps (Mandelbrot, 1963).

La figure 4.9 montre l'évolution du cours de l'action IBM pendant 300 mois, 300 jours et 300 minutes. Sans l'indication de l'échelle de temps sur l'axe des abscisses, il est impossible de distinguer ces trois courbes les unes des autres. Leurs propriétés statistiques sont identiques, la variation du cours est identique à toutes les échelles de temps, elles sont *autosimilaires*. Cette autosimilarité tient à leur irrégularité, à leur rugosité. Quelle que soit l'échelle à laquelle on les observe, de nouvelles anfractuosités apparaissent.

En multipliant les observations similaires, comme les lignes côtières (Mandelbrot, 1967) ou les turbulences (Mandelbrot, 1974), Mandelbrot théo-

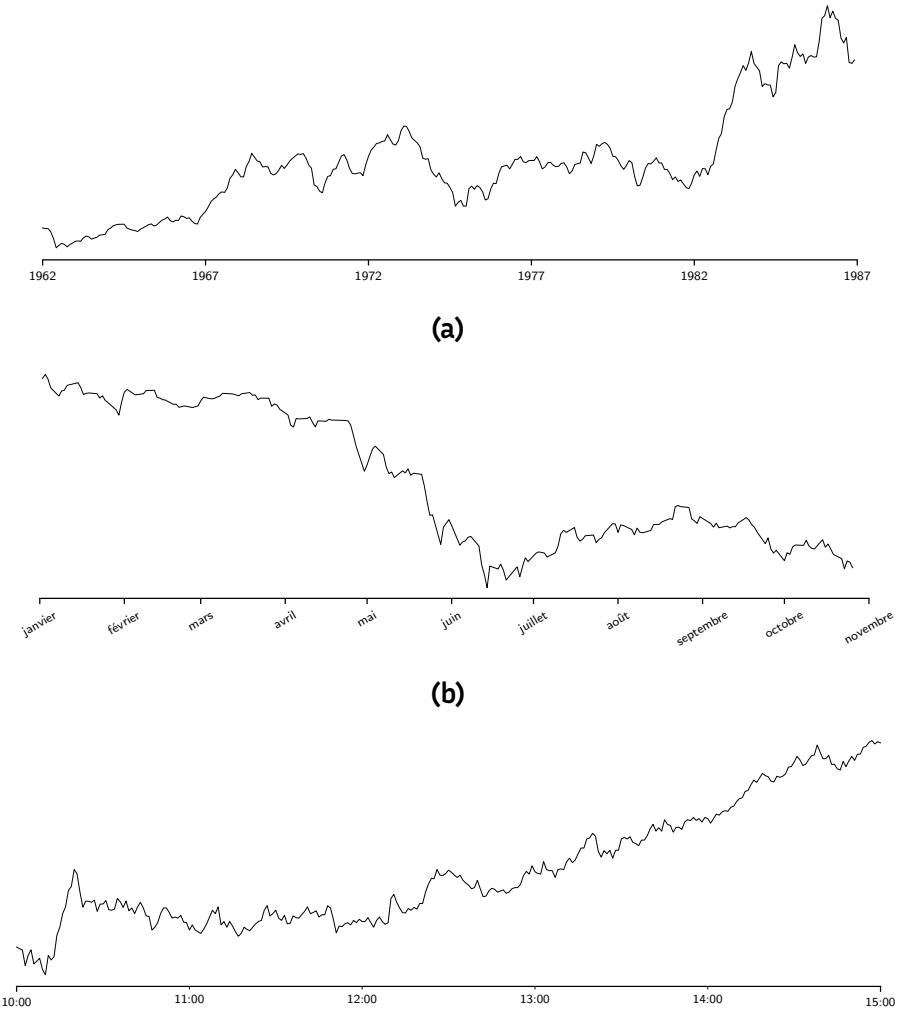


Figure 4.9 : *Cours de l'action d'IBM pendant 300 mois (a), 300 jours (b) et 300 minutes (c).*

rise le concept de fractale (Mandelbrot, 1975). Du fait de leur autosimilarité, ces objets mathématiques, à la différence par exemple des cercles ou des carrés, n'ont pas de taille spécifique. Ils permettent de fournir une description mathématique à un vaste ensemble de phénomènes naturels sans échelle (voir figure 4.10), tels que les arbres, les montagnes, les rivières, les villes, les turbulences atmosphériques, les vagues à la surface de l'océan, notre système de circulation sanguine, notre système respiratoire, le bruit d'une chute d'eau, ou encore les domaines de spins dans le modèle d'Ising à la température critique.

4.4 Origine des phénomènes sans échelle

Les phénomènes avec échelle trouvent leur origine dans le théorème central limite (page 2.1) : lorsque des erreurs, des déviations, ou toute forme de processus aléatoire interviennent dans la construction d'un phénomène[5], celui-ci suit une distribution gaussienne. Par exemple, la hauteur d'un arbre est le résultat de l'accumulation de facteurs tels que son patrimoine génétique, son exposition au soleil, la composition du sol dans lequel il pousse...Pour les phénomènes sans échelle, les mécanismes à leur origine sont plus nombreux.

Processus de Yule et attachement préférentiel

L'aspect le plus remarquable du travail de Yule (1925) sur l'analyse du nombre d'espèces par genre est qu'il a proposé un modèle mathématique simple permettant d'expliquer ses observations. Yule suppose qu'au cours d'une unité de temps donnée, mettons un siècle, chaque espèce peut muter avec une probabilité p_s pour donner naissance à une nouvelle espèce du même genre et muter avec une probabilité p_g pour donner naissance à une nouvelle espèce suffisamment différente pour créer un nouveau genre.

Si un genre a N espèces, à l'issu d'un siècle, chacune des espèces qui le composent ayant donné naissance à une nouvelle espèce avec une probabilité p_s, il aura gagné $p_s N$ espèces : plus un genre comporte d'espèces, plus il en gagne. Lorsqu'un nouveau genre apparait, il n'est initialement constitué que d'une seule espèce. Il en gagnera dans les siècles qui suivront, mais moins

[5] pour peu que leur variance soit finie

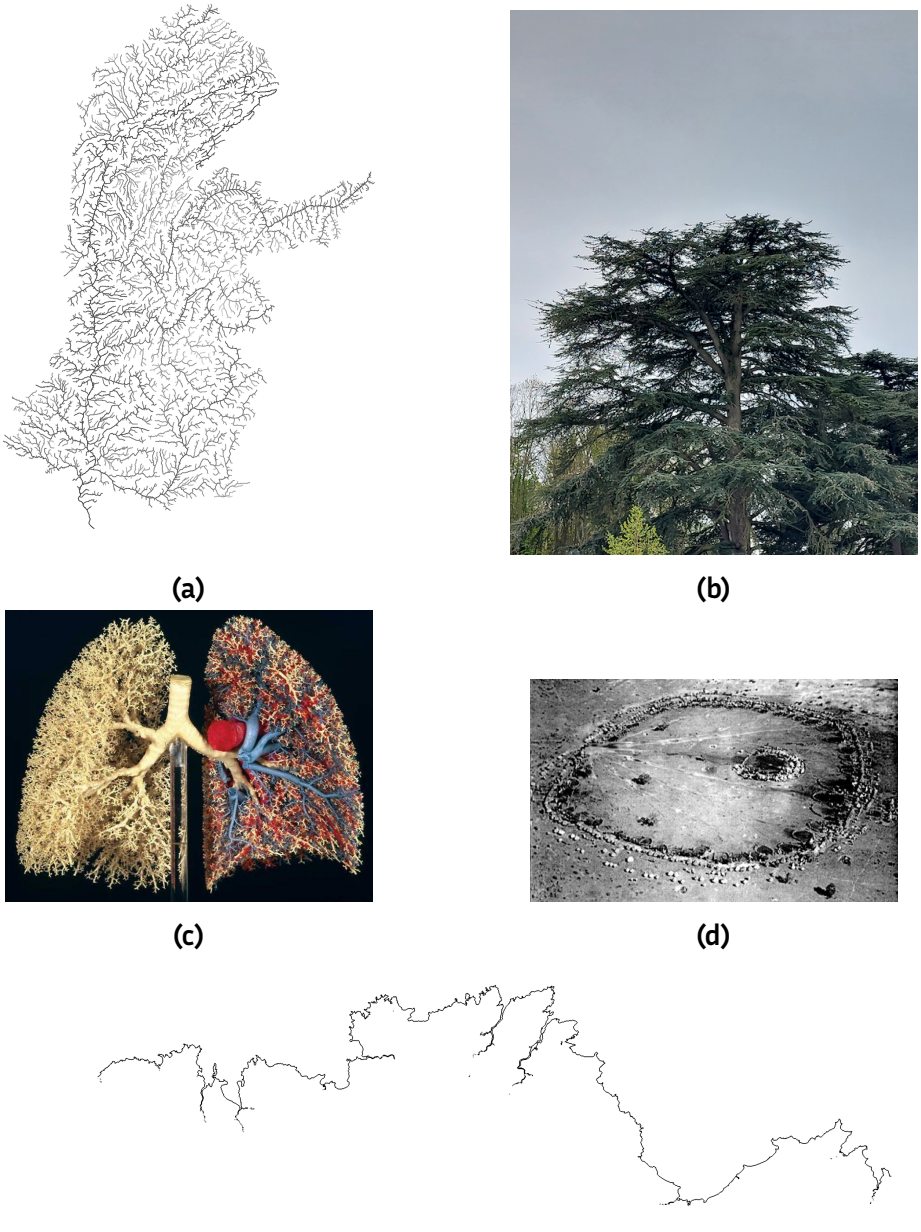


Figure 4.10 : *Exemples de fractales. a : Le Rhône et ses affluents (données d'après Lehner & Grill, 2013). b : Un cèdre. c : Structure des bronches et bronchioles et du système de circulation sanguine dans les poumons. d : Village Ba-Ila (d'après Eglash, 1999). e : Littoral des Côtes-d'Armor (données IGN).*

que les genres déjà existants, qui présentent une plus large biodiversité. Avec un tel processus, au fur et à mesure que les siècles passent, la distribution du nombre d'espèces par genre converge vers une loi de puissance.

Ce type de mécanisme, appelé processus de Yule ou attachement préférentiel, est à la source de nombreux et divers phénomènes sans échelle : plus une ville est grosse, plus elle sera le lieu de naissances et plus elle attirera de nouveaux d'habitants ; plus une personne est riche, plus son capital lui rapportera, plus elle s'enrichira ; plus une page web possède de liens entrants, plus elle est visitée et plus d'autres pages pointeront sur elle...

Ce dernier exemple est typique des phénomènes sans échelle observés dans les réseaux. En 1999, Barabási et Albert ont montré que dans de nombreux réseaux naturels, la distribution de degrés, c'est-à-dire le nombre de voisins de chaque nœud, est un phénomène sans échelle. Là encore, c'est la disponibilité de grandes masses de données qui leur a permis d'arriver à cette conclusion :

> « *Exploring several large databases describing the topology of large networks that span fields as diverse as the WWW or citation patterns in science, we show that, independent of the system and the identity of its constituents, the probability* $P(k)$ *that a vertex in the network interacts with k other vertices decays as a power law, following* $P(k) \sim k^{-\gamma}$ »
>
> (Barabási et Albert, 1999)

Pour expliquer leur observation, Barabási et Albert proposent un modèle permettant de construire des réseaux qui exhibent cette propriété. Ce modèle consiste à faire grossir un réseau de la manière suivante. Le réseau est initialement constitué d'un petit nombre de nœuds m_0. De manière itérative, de nouveaux nœuds sont ensuite ajoutés au réseau. Chaque nouveau nœud est connecté à m nœuds déjà présents dans le réseau selon une procédure d'attachement préférentiel : chacun des m nœuds est choisi avec une probabilité proportionnelle à son degré. Plus un nœud a de voisins, plus la probabilité que le nouveau nœud lui soit attaché, et donc qu'il ait un nouveau voisin, est grande. De la même manière que dans le modèle de Yule les genres ayant le plus d'espèces se diversifient plus rapidement que ceux en ayant moins, dans le modèle de Barabási et Albert les nœuds ayant le plus de voisins gagnent de nouveaux voisins plus rapidement que ceux en ayant

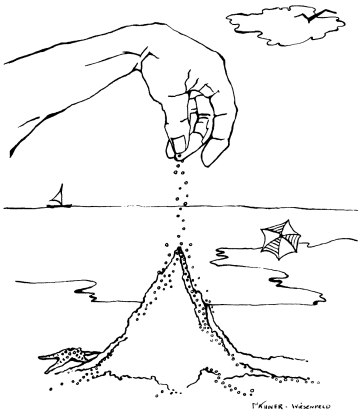


Figure 4.11 : *Illustration du modèle de Bak–Tang–Wiesenfeld. D'après (Bak, 1996)*

moins. Le nombre de voisins des nœuds, de la même manière que le nombre d'espèces par genre, finit par s'étaler sur plusieurs ordres de grandeur.

Criticalité auto-organisée

Nous avons vu à la section 4.3 qu'une manière de produire des phénomènes sans échelle est d'amener un système à son point critique. Toutefois, il est assez peu probable d'observer dans la nature des phénomènes sans échelle produits de cette manière puisque le caractère sans échelle du système dépend d'un paramètre, comme la température dans le modèle d'Ising, qui doit être réglé avec précision, par une intervention extérieure au système, à bonne valeur.

En 1987, Bak et al. ont montré qu'il existe des systèmes dont la dynamique les amène spontanément dans un état critique. Ils ont forgé le terme *criticalité auto-organisée* pour décrire ces systèmes. Ils ont également proposé un modèle simple qui présente cette criticalité spontanée.

Ce modèle consiste à jouer avec du sable en faisant tomber un à un des grains sur un tas (figure 4.11). Chaque nouveau grain tombe aléatoirement quelque part sur le tas. Si la pente à l'endroit où tombe le grain est douce, il y restera. Au contraire, si le grain tombe à un endroit trop escarpé, il roulera plus bas, entraînant éventuellement d'autres grains dans sa chute.

Plus formellement, le modèle est défini comme une grille de taille $N \times N$ sur laquelle se trouvent les grains de sables. À chacune des positions (i, j) de la grille est associé le nombre de grains de sables présents $Z(i, j)$. À chaque

itération, un grain de sable est ajouté à une position aléatoire de la grille. Si le nombre de grains de sable sur cette position dépasse 4, les grains présents à cet endroit vont se déplacer sur les positions voisines :

$$
\begin{aligned}
\text{Si} \quad & Z(i,j>3) \\
\text{alors} \quad & Z(i,j) \leftarrow 0 \\
& Z(i-1,j) \leftarrow Z(i-1,j)+1 \\
& Z(i+1,j) \leftarrow Z(i+1,j)+1 \\
& Z(i,j-1) \leftarrow Z(i,j-1)+1 \\
& Z(i,j+1) \leftarrow Z(i,j+1)+1
\end{aligned}
$$

Lorsqu'un grain de sable se déplace sur une position adjacente, il se peut qu'il y ait alors 4 grains de sables sur cette nouvelle position, déclenchant le même mécanisme, qui peut à son détour engendrer d'autres déplacements... Ainsi, la chute d'un grain de sable sur le tas peut soit ne rien faire (si $Z(i,j)$ reste inférieur à 4), soit bien provoquer une avalanche qui impliquera un nombre quelconque de grains déjà présents sur le tas. La distribution de la taille des avalanches suit une loi une puissance et les paysages dessinés présentent des structures fractales (figure 4.12)

Que le modèle établi par Bak, Tang et Wiesenfeld représente une accumulation de grains de sable n'est pas évident. On a du mal à voir apparaitre le tas tel qu'illustré par la figure 4.11 (bien que des expérimentations aient montré que des accumulations, non pas de sable, mais de riz, suivent le comportement du modèle et présentent des avalanches de toutes tailles (Frette et al., 1996)). Une autre interprétation des équations qui régissent le modèle a été proposée par Grassberger et rapportée par Bak (1996). Dans une grande salle où travaillent des bureaucrates sur des bureaux alignés en rangés et en colonnes, quelqu'un de l'extérieur apporte régulièrement un nouveau document qu'il dépose aléatoirement sur un des bureaux. Lorsque qu'un ou une bureaucrate a quatre documents sur son bureau, il ou elle s'empresse de les répartir entre ses quatre voisins, devant, derrière, à gauche et à droite, qui font de même si d'aventure leur pile dépasse quatre documents. Dans tous les cas, le point important du modèle, et de la criticalité auto-organisée de manière plus générale, est que le système n'est pas à l'équilibre, il y a un apport continu d'énergie (de nouveaux grains de sable, de nouveaux documents...). Cette énergie est emmagasinée en des points spécifiques du

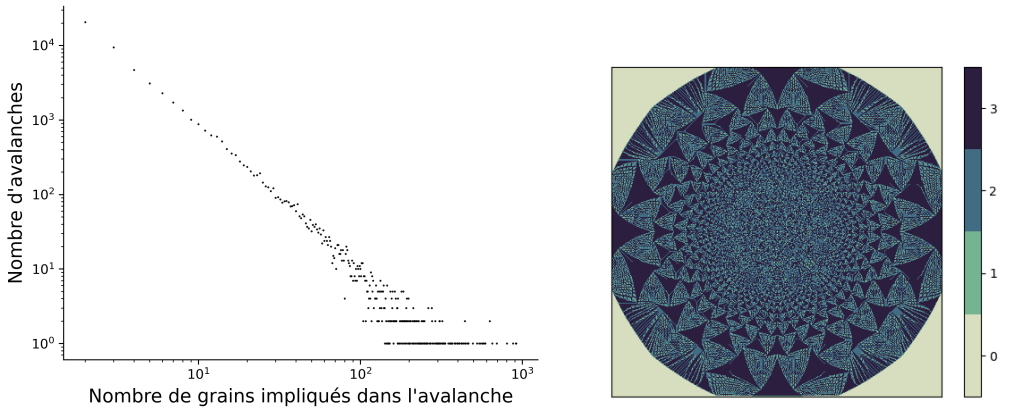


Figure 4.12 : *Résultats du modèle de Bak–Tang–Wiesenfeld. (a) Distribution de la taille des avalanches sur une grille de 512×512 positions après le dépôt de 2^{19} grains de sable. (b) Nombre de grains à chaque emplacement d'une grille de 512 positions après le dépôt de 2^{19} de grains systématiquement au centre de la grille.*

système qui, passé un certain seuil, la dissipe vers leurs voisins. Pour que la criticalité s'auto-organise, il faut que le système ait un apport continu d'énergie qu'il dissipe avec des effets de seuil.

Ce modèle, et donc plus généralement la criticalité auto-organisée, est le fruit de la révolution informatique. Malgré sa simplicité, quatre additions et une soustraction, ce modèle n'est pas solvable analytiquement, c'est-à-dire qu'on ne peut pas dériver mathématiquement ses propriétés statistiques, telles que la distribution de la taille des avalanches. Le seul moyen d'évaluer sa dynamique est de réaliser des simulations informatiques.

> *« The question of the origin of complexity from simple laws of physics — maybe the biggest puzzle of all — has only recently emerged as an active science. One reason is that high-speed computers, which are essential in this study, have not been generally available before. »*
>
> (Bak, 1996, p.6)

Dans le modèle d'Ising, et plus généralement dans le système avec des transitions de phases continues, l'état critique est instable : dès que le pa-

ramètre qui gouverne la transition de phase s'écarte du point critique, le système devient complètement ordonné ou désordonné. Ce que montre le modèle de Bak–Tang–Wiesenfeld est que certains systèmes peuvent être dans un état critique de manière stable, puisque cet état ne dépend pas d'un paramètre devant avoir une valeur précise. Depuis son introduction, la criticalité auto-organisée a été proposée comme explication pour de nombreux phénomènes naturels sans échelle (Thurner et al., 2018) : les chutes de pluie (Peters et al., 2001), les glissements de terrain (Malamud & Turcotte, 1999), les feux de forêt (Malamud et al., 1998), les entrées des salles de cinéma (D. Sornette & Zajdenweber, 1999), les tremblements de terre (A. Sornette & Sornette, 1989), les fluctuations des marchés financiers (Biondo et al., 2015), les extinctions d'espèces (Solé & Manrubia, 1996), l'activité cérébrale (Plenz et al., 2021), les inondations (Fonstad & Marcus, 2003), les guerres (Roberts & Turcotte, 1998), les interactions entre gènes (Vidiella et al., 2021), les comportements politiques (Brunk, 2001)...

4.5 La science en transition de phase

> *« Experiments must deal with length scales from as small as a grain of sand to thousands of times larger. The sand piles must be very large to test the predicted power law behavior. In nature, where landscapes extend over thousands of miles, these various length ranges are readily available, but in real life we are restricted by limited laboratory space. Also, there is a limited amount of time available; one cannot wait for hundreds of years to amass a sufficiently large amount of data. On the computer, we had the luxury of studying billions of grains of sand and millions of avalanches. The distribution of avalanches is a power law, so large events are bound to occur; however, to have just one avalanche of size 1 million, one must wait for and monitor 1 million avalanches of size 1. Experimentalists do not have that luxury. »*
>
> (Bak, 1996, p.66)

À travers ces quelques phrases, Bak résume combien il est soit facile, soit difficile, d'observer un phénomène sans échelle selon que l'on ait, ou pas, un

ordinateur. Par leur nature même, l'étalement à travers les ordres de grandeur des phénomènes sans échelle se fait simultanément dans leurs manifestations et la probabilité d'occurrence de leurs manifestations. Il est donc nécessaire pour observer un phénomène sans échelle d'accumuler beaucoup de données, suffisamment pour que les cas les plus rares, qui sont les plus importants, soient représentés.

Sans ordinateur, la tâche est titanesque. Les quelques pionniers qui y sont parvenus entre la fin du 19^{e} siècle et le milieu du 20^{e} se sont tous appuyés sur des milliers d'heures de travail de patientes collectes de données exhaustives.

À l'inverse, avec un ordinateur, la tâche se simplifie et peu même devenir triviale. Pour calculer les fréquences des mots dans un roman, plus besoin d'une équipe de 23 personnes et de 14 mois de travail, mais simplement de quelques minutes pour écrire le programme et de quelques microsecondes pour l'exécuter.

Pendant des décennies, les phénomènes sans échelle sont restés marginaux. Et soudainement, grâce à l'invention de l'ordinateur, nous avons pris conscience qu'ils sont omniprésents. Il n'y a pas de terme plus précis que *transition de phase* pour décrire la bascule qui s'est opérée. Nous sommes passés d'une vision du monde où la *normalité*, c'est le nom que l'on a donné à la distribution gaussienne, consiste à ne pas être trop éloigné de la moyenne à une vision du monde où la diversité d'un extrême à l'autre est naturelle.

Nous avons vu au chapitre 2 que la vision du monde initiale, celle de la primauté des phénomènes avec échelle, avait été un principe structurant de nos sociétés. On est en droit de s'attendre à ce que notre nouvelle vision du monde, qui abolit la primauté des phénomènes avec échelle pour s'ouvrir aux phénomènes sans échelle, ait également des conséquences sur nos vies sociales. C'est sur ce point que porte le chapitre suivant.

5 Nouvelles identités

5.1 L'individualisation des comportements

Une affaire de goûts

Aux États-Unis, pour qu'un album soit certifié Disque d'or, il faut que ses ventes atteignent 500 000 exemplaires. À 1 million d'exemplaires vendus, il est Disque de platine et même multi-platine s'il dépasse les 2 millions. La consécration commerciale ultime, Disque de diamant, est atteinte lors de la vente du 10 millionième exemplaire.

La figure 5.1, issue de Anderson (2008, p.32), montre le nombre total d'albums certifiés chaque année aux États-Unis entre 1957 et 2007. La première partie de la courbe, jusqu'en 2002, montre une croissance à peu près constante. Elle correspond à la croissance de l'industrie musicale dans la 2nde moitié du 20^{e}siècle. À partir de 2002, le nombre d'albums certifiés chute drastiquement. Une première explication à cette observation est l'apparition des applications de partage de fichiers pair-à-pair telles que Napster, Gnutella ou eDonkey à partir de 1999 et 2000. Ces applications étaient majoritairement utilisées pour partager de fichiers musicaux, le plus souvent illégalement, et ont fait chuter les ventes de disques. Il n'y a rien de surprenant que, mécaniquement, le nombre d'albums certifiés aient chuté aussi.

Mais cette explication n'est pas suffisante. Alors que de 2001 à 2007 les ventes d'albums baissaient de 25 %, le nombre d'albums certifiés s'écroulait, lui, de 60 %. L'essor du partage musical en ligne n'explique pas à lui seul pourquoi cette baisse a davantage touché les albums les plus vendus que les autres. Anderson (2008) avance une autre explication. Avant l'avènement de la musique en ligne, initialement par partage pair-à-pair et aujourd'hui par les services de streaming tels que Deezer ou Spotify, la musique s'écoutait à la radio et sur des cassettes ou des CD achetés dans des magasins physiques.

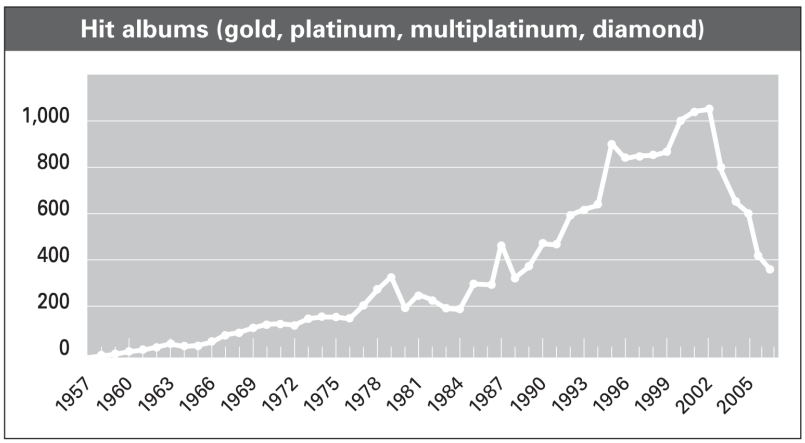


Figure 5.1 : *Nombre d'albums certifiés disque d'or, de platine ou de diamant entre 1957 et 2007. D'après (Anderson, 2008)*

Dans les deux cas, la place est limitée. À la radio, elle est limitée temporellement : les journées ne font que 24 heures ce qui contraint le nombre de chansons diffusées chaque jour; une radio qui passerait de la musique en boucle pourrait passer 480 morceaux de 3 minutes dans une journée. Dans les magasins, la place est limitée physiquement : la surface de vente disponible limite le nombre d'albums disponibles en stock à quelques centaines de références, peut-être quelques milliers dans les plus grands. Spotify propose à l'écoute 100 millions de titres. L'économie numérique est dégagée des contraintes spatio-temporelles de l'économie traditionnelle et les choix d'écoute s'en trouvent démultipliés.

La popularité musicale est un phénomène sans échelle. D'un côté, une poignée d'artistes populaires à l'extrême qui comptent leur fan par dizaines et même centaines de millions, de l'autre des myriades d'artistes à la notoriété la plus confidentielle; entre les deux, des artistes à tous les niveaux de popularité. Un magasin physique qui ne peut proposer à la vente que quelques centaines d'artistes orientera ses choix vers les populaires, les artistes qui vendent beaucoup de disques. Bien sûr, il existe des disquaires spécialisés, délaissant les têtes d'affiches pour cibler une niche musicale particulière. Mais chez ces disquaires, les contraintes sont les mêmes, à une échelle différente : à l'intérieur d'un genre musical donné, on retrouve la même structure, avec d'un côté les fers de lance, les représentants les plus importants du genre, et de l'autre une majorité d'artistes plus ou moins mineurs. De genre en sous-genre et de sous-sous-genre en sous-sous-sous-

genre, la musique est fractale.

Au 20e siècle, le choix musical était limité et fortement biaisé en faveur de la tête de la distribution. Au 21e, tout le monde a accès à tout, tout le monde peut explorer à sa guise l'ensemble de la longue queue de la distribution. L'écoute de musique se répartissant sur un choix plus vaste, mécaniquement, la tête de la distribution reçoit moins d'attention et le nombre d'albums certifiés diminue.

L'argument d'Anderson (2008) est que l'économie numérique repose en grande partie sur l'exploitation des longues queues des distributions des phénomènes sans échelle. Là où un magasin ne peut offrir qu'un choix limité, et donc la tête de la distribution, Amazon vend tout. Là où la publicité à la télé, à la radio, dans la presse ou sur les murs s'adresse au plus grand monde, et donc porte sur les produits de plus grande consommation, Google peut proposer les publicités les plus spécifiques à chacun. Là où la télévision propose les programmes les plus rassembleurs, YouTube, Instagram ou TikTok proposent des myriades de vidéos ciblées sur des préférences personnelles.

Cela ne signifie pas que chacun fait des choix musicaux, cinématographiques, littéraires radicalement différents de sa voisine ni que chacune consomme radicalement différemment que son voisin. Les vidéos les plus vues sur YouTube ont été vues plusieurs milliards de fois, touchant un nombre de personnes sans doute jamais atteint par les médias de masse. La tête de distribution, ce qui est vu ou entendu par le plus grand nombre, reste massive. Mais la queue de la distribution l'est tout autant (Cheng et al., 2008) et chacun peut y puiser de quoi construire sa différenciation.

Cela tranche avec l'audiovisuel traditionnel dont Stiegler (2008) disait qu'il *« engendre des comportements grégaires et non, contrairement à une légende, des comportements individuels. Dire que nous vivons dans une société individualiste est un mensonge patent, un leurre extraordinairement faux [...]. Nous vivons dans une société-troupeau, comme le comprit et l'anticipa Nietzsche »* (cité par Dufour, 2007)

Tout le monde debout

Nos modes de consommation ne sont pas le seul volet de nos vies où les technologies numériques promeuvent des phénomènes sans échelle. Les médias de l'ère pré-numérique sont caractérisés par des processus qui ne mettent en position de production qu'une minorité de la population : par

rapport au nombre d'auditeurs ou de téléspectateurs, peu de gens passent à la radio ou à la télévision ; les journalistes ne représentent qu'une fraction de la population ; les sociétés de production cinématographique ou musicale, les maisons d'édition ou encore les galeries d'art sont autant de processus de filtres qui sélectionnent les quelques personnes qui seront cinéastes, musiciens ou musiciennes, auteurs ou autrices, ou encore artistes. Les médias traditionnels produisent donc des catégories qui s'inscrivent dans la logique de la statistique de la loi normale, avec d'un côté la figure de l'auteur, du journaliste, du cinéaste ou de l'artiste, et de l'autre une masse peu ou prou indifférenciée de consommateurs.

À l'inverse, les médias socionumériques offrent à chacun de se trouver en position de producteur. Sur YouTube ou TikTok, tout un chacun peut proposer ses créations audiovisuelles au reste de l'humanité. Sur SoundCloud, tout musicien peut exposer sa production. Sur Facebook ou Twitter chacun peut publier ses pensées. Les plateformes de blogs sont des d'expression d'une littérature populaire. Wikipédia est une encyclopédie collective. Là encore, on voit apparaitre une désuniformisation au profit d'un phénomène sans échelle. La figure 5.2 montre la distribution du nombre de tweets produits par les utilisateurs francophones de Twitter (Magué, 2018). Point d'utilisateur moyen, mais au contraire une distribution des comportements à travers les échelles. Des observations similaires sont faites par exemple sur YouTube (Cheng et al., 2008) ou Wikipédia (Kittur et al., 2007).

5.2 Nouvelles sociabilités, nouvelles identités

En nous affranchissant de l'espace physique comme lieu quasi nécessaire de sociabilisation, les technologies numériques ont remodelé notre sociabilité à travers « *the transformation of community from solidary groups to individualized networks* » (Wellman, 2001). Tant que nos interactions sociales se déroulent dans un espace physique, deux personnes physiquement proches (du même village ou du même quartier par exemple) auront peu d'opportunités de développer des sociabilités différentes. Mais si cette contrainte physique est abrogée, ce que font les technologies numériques, ces opportunités se multiplient et chacun se révèle être en position de construire un ensemble de relations sociales qui lui est propre. Wellman (2001) décrit ce processus comme un *individualisme connecté* dont l'une des conséquences directes sont les « *reduced identity and pressures of belonging to groups* ».

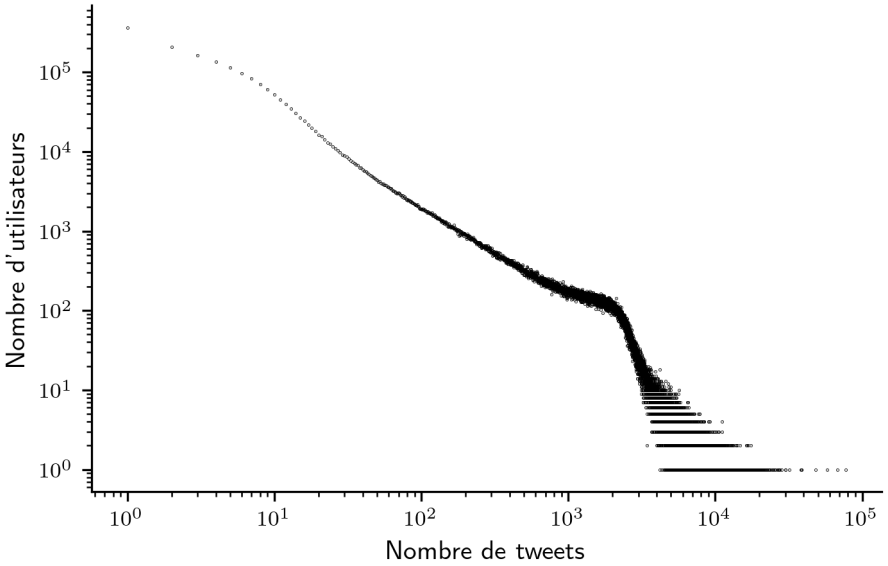


Figure 5.2 : *Distribution du nombre de tweets par utilisateur calculée à partir de données collectées dans le cadre du projet SoSweet (Magué, 2018). Elles couvrent 658 747 413 tweets produits par 2 627 890 utilisateurs francophones entre octobre 2006 et mars 2019.*

Dès 1993, dans une analyse des premières communautés en ligne nées, avant le Web, dans les années 1980, Rheingold constate que *« the latest computer-mediated communications media seem to dissolve boundaries of identity »*. Dans ce qui est sans doute le premier essai sur la manière dont nos interactions en ligne transforment nos identités, Turkle (1997) pointe le fait qu'à travers la multiplicité des manières qui nous sont offertes d'interagir en ligne, notre identité se trouve fragmentée : *« your identity on the computer is the sum of your distributed presence »* (p.13). Chaque service, chaque plateforme que nous sommes amenés à utiliser quotidiennement constitue une occasion de modeler une facette de notre identité. Cette démultiplication du soi modifie en profondeur le rapport à la norme sociale et à la diversité :

> *« When identity was defined as unitary and solid it was relatively easy to recognize and censure deviation from a norm. A more fluid sense of self allows a greater capacity for acknowl-*

> *edging diversity. It makes it easier to accept the array of our (and others') inconsistent personae –— perhaps with humor, perhaps with irony. We do not feel compelled to rank or judge the elements of our multiplicity. We do not feel compelled to exclude what does not fit.* »
>
> (p.261)

Nous ne nous définissons plus par rapport à des normes et des catégories établies, mais construisons des identités taillées sur mesure. Les premières lignes de *Emergent Identities : New Sexualities, Genders and Relationships in a Digital Era* (Cover, 2018) ne pourraient pas l'illustrer de manière plus explicite :

> « *Digital media has in recent years enabled people, including particularly younger people, to engage creatively and interactively in defining their own sense of identity. This has included the production of new, diverse 'labels' or 'categories' of sexuality and gender identity and definitions of relationships. Challenging the older languages of LGBT identity in many ways, the new labels include over a hundred new terms to describe sexuality and gender, including terms such as heteroflexible, non-binary, asexual, greysexual, sapiosexual, demisexual, ciswoman, transcurious, maverique and many more.* »
>
> (p.1)

Ces labels ou catégories sont créés par notre impératif cognitif à catégoriser et par la nécessité de les communiquer, mais ce que révèle leur prolifération est que le champ des identités n'est plus un système de catégories dont le prototype a un pouvoir normatif (hétérosexuel vs homosexuel, pour filer l'exemple des orientations sexuelles). Au contraire, ce champ des identités est un panorama complexe, fractal, à l'intérieur duquel chacune et chacun peut se positionner de manière fluide, changeante et personnelle.

6 En transition permanente

Dans l'extrait de Phèdre en exergue de l'introduction, Platon fait dire (à Socrate qui fait dire) à Thamous, roi de Thèbes, qu'il redoute les effets délétères de l'écriture que le dieu Theuth vient lui présenter après l'avoir inventée. Rétrospectivement, on peut sans risque affirmer que Thamous s'est trompé, que l'écriture a été une des inventions les plus importantes, si ce n'est la plus importante, de l'histoire de l'Humanité. Thamous avait toutefois raison en une chose, l'écriture portait en elle un pouvoir de transformation sociale radicale.

On peut généraliser : les technologies épistémiques, c'est-à-dire les technologies qui permettent la collecte de l'information, son inscription, son analyse, qui permettent la construction et la circulation de la connaissance, portent en elles un pouvoir de transformation sociale radicale (McLuhan, 2008). C'est vrai de l'écriture (Goody, 1986), de l'imprimerie (Eisenstein, 1980) ou encore des statistiques (Desrosières, 1993).

C'est également vrai d'Internet, du Web et des technologies numériques en général qui portent en elles un pouvoir de transformation sociale radicale. Toutefois, cette affirmation ne saurait être avancée sans être accompagnée de la justification de la radicalité de cette transformation. Quelle est sa nature et comment est-elle liée à ces technologies ?

L'argument qui a été développé ici est que ces technologies numériques nous ont donné accès à l'observation des phénomènes sans échelle et qu'elles promeuvent la génération de phénomènes sans échelle dans nos comportements. Ce faisant, elles ont transformé nos représentations de l'espace social et notre rapport à la norme et transformé la construction de notre sociabilité et de notre identité. Les phénomènes sans échelle, à travers les technologies numériques, transforment notre rapport à l'autre et notre rapport à soi.

Pour pouvoir parler de transformation, il est nécessaire de comparer

l'état initial à l'état final. Ici, l'état initial était un ordre social construit sur la base de représentations issues de la manière dont la statistique s'est développée entre le milieu du 18e siècle et le milieu du 20e. Dans son essai *La Rebelión de Las Masas*[1], Ortega y Gasset (1930) décrit et s'inquiète de l'avènement des sociétés de masse. Il trace une filiation directe entre les masses et la statistique du siècle précédent en faisant directement référence à Quetelet : *« la masse, c'est l'homme moyen »* (p.86). L'uniformisation figure au premier rang des critiques qu'il formule à l'égard de cette société en construction :

> *« Comme on dit en Amérique du Nord, être différent est indécent. La masse fait table rase de tout ce qui n'est pas comme elle, de tout ce qui est excellent, individuel, qualifié et choisi. Quiconque n'est pas comme tout le monde, ne pense pas comme tout le monde, court le risque d'être éliminé. »*
>
> (p.90)

Ce que les technologies numériques provoquent, c'est une désagrégation de cette uniformisation. L'effet est à double tranchant : si l'on peut se réjouir de la disparition de l'aliénation des sociétés de masse, on peut légitimement être méfiant face à la montée d'un narcissisme numérique qui se manifeste par l'exposition de soi au plus grand nombre, et on ne peut qu'être inquiet face au nouveau rapport à la vérité qui s'instaure avec la circulation d'informations erronées. Dans le travail journalistique, comme dans d'autres domaines, la normativité est une garantie.

L'Histoire de l'Humanité est marquée par des transitions de régime épistémiques provoquées par les avènements successifs de nouvelles technologies épistémiques. Mais à regarder le passé, il y a tout lieu de penser qu'une double accélération est à l'œuvre. D'une part, ces transitions se succèdent à un rythme toujours plus soutenu. Il aura fallu des millénaires entre l'écriture et l'imprimerie, des siècles entre l'imprimerie et Internet, et il n'aura fallu que quelques décennies pour que l'éclosion de l'intelligence artificielle n'amorce la prochaine transition. D'autre part, le rythme auquel ces transitions s'opèrent est également toujours plus rapide. S'il a fallu des millénaires pour que l'écriture nous fasse entrer dans l'Histoire et des siècles pour que l'imprimerie nous fasse entrer dans la Modernité, le temps de la transition

[1] *La révolte des masses*. Les citations sont tirées de l'édition en français de 2010

qui se déroule sous nos yeux sera compté en années, le temps d'une génération, tout au plus.

Cette double accélération n'est pas sans poser de questions sur les capacités des sociétés à se transformer et sur les capacités des individus qui les composent à intégrer ces transformations. Un enfant du 12e siècle n'avait pas de difficulté à imaginer son avenir, devenir paysan ou paysanne comme l'étaient ses parents. Pour un enfant du 21e siècle, comment imaginer le monde dans lequel il ou elle grandira et vieillira ? Une certitude, l'acquis social de la révolution numérique, est qu'iel aura la liberté dans la définition de soi.

Péroraison

« Les deux pôles possibles de la connaissance sont de savoir presque tout sur presque rien ou de savoir presque rien sur presque tout ».

Comme je l'expliquais dans le préambule, cette phrase m'a un jour été adressée par Kris Lund. À la lumière des arguments présentés dans ce texte, nous pouvons tenter de mieux comprendre son sens profond, ce qu'elle dit de la science et de la posture du scientifique.

On décompose la science en disciplines, qui se découpent en sous-disciplines, elles-mêmes structurées de sous-sous-disciplines… La science est fractale. À chaque échelle, le savoir se déploie dans une complexité sans cesse renouvelée, chaque détail portant en lui, pour peu qu'on l'examine d'assez près, le potentiel d'une analyse toujours plus poussée.

« Les deux pôles possibles de la connaissance sont de savoir presque tout sur presque rien ou de savoir presque rien sur presque tout ». C'est une question d'échelle. Parce que notre expérience vécue se déroule à une échelle donnée, nous sommes appelés à choisir celle à laquelle nous construisons notre savoir : les deux pôles possibles de la connaissance sont d'une part de plonger dans la fractale et de s'émerveiller de sa richesse à mesure que l'on s'y enfonce, et d'autre part de l'appréhender dans son entièreté et de s'émerveiller de la richesse de la récurrence de ses motifs. Sans juger un pôle meilleur que l'autre, j'opte pour le second.

On décompose la science en disciplines, parce que notre esprit ne peut s'empêcher de construire des catégories. On identifie des zones plus denses dans le continuum des phénomènes du monde autour de nous que l'on nomme physique, biologie, psychologie ou encore sociologie. Pour des raisons pragmatiques plus qu'ontologiques, on institutionnalise ces disciplines par des

départements dans les universités, des revues et des sociétés savantes. Ces disciplines, ces catégories, s'installent dans nos imaginaires collectifs par des représentations normatives qui conditionnent les trajectoires des scientifiques (doctorant, j'ai participé à fonder la fédération des étudiants et jeunes chercheurs en sciences cognitives. Notre action principale était de faire reconnaitre auprès des instances nationales les sciences cognitives comme discipline pour assurer la création de postes). Au vu des avancées de la science au cours des siècles passés, force est de reconnaitre que cette organisation en disciplines institutionnalisées et quasi essentialisées est efficace. Mais elle est cloisonnante.

« *Les deux pôles possibles de la connaissance sont de savoir presque tout sur presque rien ou de savoir presque rien sur presque tout* ». L'institutionnalisation de la science a opté pour le premier pôle. En choisissant le second, je cesse de me définir par des catégories normatives et j'assume une identité scientifique fluide, changeante et personnelle.

Bibliographie

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2018). Socioeconomic Dependencies of Linguistic Patterns in Twitter : A Multivariate Analysis. *Proceedings of the 2018 World Wide Web Conference*, 1125-1134.

Abitbol, J. L., Chevrot, J.-P., Karsai, M., Magué, J.-P., Léo, Y., Nardy, A., & Fleury, E. (2017). The study of optional realization of the French negative particle (ne) on Twitter : Is Sociolinguistics compatible with the Big Data ?

Abitbol, J. L., Karsai, M., Chevrot, J.-P., Magué, J.-P., & Fleury, E. (2017a). How social, economic and demographic forces shape linguistic variation on Twitter.

Abitbol, J. L., Karsai, M., Chevrot, J.-P., Magué, J.-P., & Fleury, E. (2017b). Socioeconomic and network dependencies of linguistic patterns in Twitter.

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2017). Optional realisation of the French negative particle (ne) on Twitter : Can big data reveal new sociolinguistic patterns ?

Académie Royale des sciences (Éd.). (1769). *Recueil des pieces qui ont remporte des prix de l'Académie royale des sciences*.

Ammon, O. (1895). *Die Gesellschaftsordnung und ihre natürlichen Grundlagen. Entwurf einer Sozial-Anthropologie zum Gebrauch für alle Gebildeten, die sich mit sozialen Fragen befassen.* G. Fischer.

Ammon, O. (1900). *L'ordre social et ses bases naturelles : esquisse d'une anthroposociologie* (H. Muffang, Trad.). A. Fontemoing.

Anderson, C. (2008). *The Long Tail : Why the Future of Business Is Selling Less of More* (Revised édition). Hachette Books.

Aristote. (s. d.). *De Anima*.

Armatte, M. (2006). Fréchet et la médiane : un moment dans une histoire de la robustesse. *Journal de la société française de statistique*, *147*(2), 23-37.

Auerbach, F. (1910). *Geschichtstafeln Der Physik*. Barth.

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, *59*, 74-76.

Bachelier, L. (1900). Théorie de La Spéculation. *Annales scientifiques de l'ENS*, *3*(17), 21-86.

Bak, P. (1996). *How nature works : The science of self-organized criticality* (1. softcover printing). Springer-Verlag.

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-Organized Criticality : An Explanation of the 1/f Noise. *Physical Review Letters*, *59*(4), 381-384.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509-512.

Bartko, J. (2003). *Un Instrument de Travail Dominicain Pour Les Prédicateurs Du XIII°siècle : Les Sermones Des Evangeliis Dominicalibus de Hugues de Saint-Cher (-1263) — Edition et Étude* [Thèse de doctorat]. Université Lumière Lyon 2.

Beaugiraud, V., Gedzelman, S., Ingarao, M., Magué, J.-P., & Saïdi, S. (2011). Amalia : An eSciDoc Based Solution to Manage the Production, Processing and Publishing Workflows of TEI Data.

Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551-572.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1966). Folk Taxonomies and Biological Classification. *Science*, *154*(3746), 273-275.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General Principles of Classification and Nomenclature in Folk Biology. *American Anthropologist*, *75*(1), 214-242.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1974). *Principles of Tzeltal Plant Classification : An Introduction to the Botanical Ethnography of a Mayan-speaking People of Highland Chiapas*. Academic Press.

Bernoulli, J. I. (1784). Milieu. In *Encyclopédie Méthodique. Mathématiques. Tome 2*. Chez Panckoucke.

Biondo, A. E., Pluchino, A., & Rapisarda, A. (2015). Modeling Financial Markets by Self-Organized Criticality. *Physical Review E*, *92*(4), 042814.

Boltanski, L. (1982). *Les Cadres : La Formation d'un Groupe Social*. Editions de Minuit.

Bowker, G. C., & Star, S. L. (2008). *Sorting things out : Classification and its consequences* (1. paperback ed., 8. print). MIT Press.

Bozzolo, C., & Ornato, E. (1980). *Pour une histoire du livre manuscrit au Moyen Âge : trois essais de codicologie quantitative*. Éditions du Centre National de la Recherche Scientifique.

Brunk, G. G. (2001). Self-Organized Criticality : A New Theory of Political Behaviour and Some of Its Implications. *British Journal of Political Science*, *31*(2), 427-445.

Buringh, E., & Zanden, J. (2009). Charting the "Rise of the West" : Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries. *The Journal of Economic History*, *69*, 409-445.

Campbell, W. W. (1924). *Biographical Memoir, Simon Newcomb, 1835-1909*. U.S. Government Printing Office.

Château-Dutier, E., Boschetto, S., Boulai, C., Gedzelman, S., Ingarao, M., Jallud, P.-Y., Morlok, E., Pons, P., Saïdi, S., & Magué, J.-P. (2015). SynopsX A Lightweight Xquery-Based Framework to Easily Publish and Expose XML Corpora.

Château-Dutier, E., Ingarao, M., Magué, J.-P., Pons, P., Gedzelman, S., Boschetto, S., Saïdi, S., Beaugiraud, V., Boulai, C., Morlock, E., & Jallud, P.-Y. (2016). Towards an XML Corpora Exposition as LOD with the Lightweight Xquery-Based Framework SynopsX.

Chen, M. (2014). *Big Data : Related Technologies, Challenges and Future Prospects*. Springer.

Cheng, X., Dale, C., & Liu, J. (2008). Statistics and Social Network of YouTube Videos. *2008 16th Interntional Workshop on Quality of Service*, 229-238.

Chevrot, J.-P., Levy Abitbol, J., Karsai, M., Magué, J.-P., & Fleury, E. (2019). Variations Du (Ne) Négatif Du Français Dans Twitter. Que Peut Apporter l'étude Des Données Massives Aux Questions de Sociolinguistique ? *CILPR 2019 - XXIXe Congrès International de Linguistique et de Philologie Romanes*.

Chevrot, J.-P., Nardy, A., Fleury, E., Karsai, M., & Magué, J.-P. (2015). Sociolinguistique et sciences cognitives : l'individu, le collectif et le réseau. *Journées FLORaL-PFC 2015 : la base de données Phonologie du Français Contemporain dans le champ phonologique*.

Cover, R. (2018). *Emergent Identities : New Sexualities, Genders and Relationships in a Digital Era*. Routledge.

Cryle, P., & Stephens, E. (2017). *Normality : A Critical Genealogy*. University of Chicago Press.

Dean, J., & Ghemawat, S. (2004). MapReduce : Simplified Data Processing on Large Clusters. *OSDI'04 : Sixth Symposium on Operating System Design and Implementation*, 137-150.

Desrosières, A. (1977). Élements Pour l'histoire Des Nomenclatures Socioprofessionnelles. In *Pour Une Histoire de La Statistique* (p. 155-231, T. 1). Economica, INSEE.

Desrosières, A. (1993). *La politique des grands nombres*. La Découverte.

Desrosières, A. (2013). *Pour une sociologie historique de la quantification : L'Argument statistique I*. Presses des Mines.

Desrosières, A., & Thévenot, L. (2002). *Les catégories socioprofessionnelles*. La Découverte.

Diebold, F. X. (2012). *On the Origin(s) and Development of the Term 'Big Data'* (SSRN Scholarly Paper Nº ID 2152421). Social Science Research Network. Rochester, NY.

Dufour, D.-R. (2007). Le troisième parent. *La clinique lacanienne*, *12*(1), 49-60.

Dunn, R., Higgitt, R., Rees, M. J., & Dunn, R. (2014). *Ships, Clocks, and Stars*. Harper Design, an imprint of HarperCollins Publishers; [Published] in association with Royal Museums Greenwich.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Éd.). (2022). *Ethnologue : Languages of the World.* (25e éd.). SIL International.

Eglash, R. (1999). *African Fractals : Modern Computing and Indigenous Design*. Rutgers University Press.

Eisenstein, E. L. (1980). *The Printing Press as an Agent of Change*. Cambridge University Press.

Field, J. A. (1917). Some Advantages of the Logarithmic Scale in Statistical Diagrams. *Journal of Political Economy*, *25*(8), 805-841.

Fink-Jensen, J. (2015). *Book Titles per Capita*. IISH Data Collection.

Fonstad, M., & Marcus, W. A. (2003). Self-Organized Criticality in Riverbank Systems. *Annals of the Association of American Geographers*, *93*(2), 281-296.

Foucault, M. (2004). *Sécurité, Territoire, Population : Cours Au Collège de France, 1977-1978*. Seuil : Gallimard.

Frette, V., Christensen, K., Malthe-Sørenssen, A., Feder, J., Jøssang, T., & Meakin, P. (1996). Avalanche dynamics in a pile of rice. *Nature*, *379*(6560), 49-52.

Funkhouser, H. G. (1937). Historical Development of the Graphical Representation of Statistical Data. *Osiris*, *3*, 269-404.

Galton, F. (1877). Typical Laws of Heredity. *Nature*, *15*.

Galton, F. (1883). *Inquiries Into Human Faculty and Its Development* (Macmillan). Cornell University Library.

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263.

Galton, F. (1889). *Natural inheritance.* Macmillan.

Galton, F. (1909). *Essays In Eugenics*. The eugenics education society.

Gasparri, F. (2009). Constitution et première organisation d'une bibliothèque canoniale au XIIe siècle. *Cahiers de recherches médiévales et humanistes. Journal of medieval and humanistic studies*, (17), 203-208.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solum Ambientium* (Friedrich Perthes and I.H. Besser).

Gescheider, G. A. (1997). *Psychophysics : The Fundamentals* (3e éd.). Psychology Press.

Goody, J. (1986). *The Logic of Writing and the Organization of Society*. Cambridge University Press.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360-1380.

Gutenberg, B., & Richter, C. F. (1949). *Seismicity Of The Earth And Associated Phenomena*. Princeton University Press.

Habert, B., Salaün, J.-M., & Magué, J.-P. (2012). Architecte de l'Information : Un Métier (F. Girard & M. Taillefer, Éd.). *Documentaliste-Sciences de l'Information*, *49*(1), 4.

Hacking, I. (1982). Biopower and the Avalanche of Printed Numbers. *Humanities in Society*, *5*, 279-295.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. Wiley.

Hamesse, J. (1995). Le Modèle Scolastique de La Lecture. In C. Guglielmo & R. Chartier (Éd.), *Histoire de La Lecture Dans Le Monde Occidental* (p. 125-145). Seuil.

Hanley, M. (1937). *Word Index to James Joyce's Ulysses*. University of Wisconsin Press.

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In L. G. Sergio Bolasco Isabella Chiari (Éd.), *Statistical Analysis of Textual Data - Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles* (p. 1021-1032, T. 2). Edizioni Universitarie di Lettere Economia Diritto.

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, *332*(6025), 60-65.

IDC. (2008). *The Diverse and Exploding Digital Universe : An Updated Forecast of Worldwide Information Growth Through 2011*. International Data Corporation.

IDC. (2012). *The Digital Universe in 2020 : Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. International Data Corporation.

IDC. (2014). *The Digital Universe of Opportunities : Rich Data and the Increasing Value of the Internet of Things*. International Data Corporation.

IDC. (2018). *The Digitization of the WorldFrom Edge to Core*. International Data Corporation.

IDC. (2020). *IDC's Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data*. IDC : The premier global market intelligence company. Récupérée août 17, 2020, à partir de https://www.idc.com/getdoc.jsp?containerId=prUS46286020

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253-258.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the Few vs. Wisdom of the Crowd : Wikipedia and the Rise of the Bourgeoisie. *World Wide Web-internet and Web Information Systems*, *1*(2), 19.

Komura, Y., & Okabe, Y. (2016). Improved CUDA programs for GPU computing of Swendsen–Wang multi-cluster spin flip algorithm : 2D and 3D Ising, Potts, and XY models. *Computer Physics Communications*, *200*, 400-401.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual Images Preserve Metric Spatial Information : Evidence from Studies of Image Scanning. *Journal of Experimental Psychology : Human Perception and Performance*, *4*, 47-60.

Labov, W. (1973). The Boundaries of Words and Their Meanings. In C.-J. Bailey & R. Shuy (Éd.), *New Ways of Analyzing Variation in English,* (Georgetown University Press, p. 340-373).

Lakoff, G. (1987). *Women, fire, and dangerous things what categories reveal about the mind*. The University of Chicago Press.

Lakoff, G., & Johnson, M. (1998). *Philosophy In The Flesh : The Embodied Mind And Its Challenge To Western Thought*. Basic Books.

Laney, D. (2001). *3D Data Management : Controlling Data Volume, Velocity, and Variety*. META Group.

Laplace, P. S. (1810). Mémoire Sur Les Approximations Des Formules Qui Sont Fonctions de Très-Grands Nombres, et Sur Leur Application Aux Probabilités. *Mémoires de l'Académie des sciences*, 353-415, 559-565.

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing : Baseline data and new approaches to study the world's large river systems : GLOBAL RIVER HYDROGRAPHY AND NETWORK ROUTING. *Hydrological Processes*, *27*(15), 2171-2186.

Levy Abitbol, J., Chevrot, J.-P., Karsai, M., Magué, J.-P., Léo, Y., Nardy, A., & Fleury, E. (2017). The Optional Realization of the French Negative Particle (Ne) on Twitter : Space, Status and Time. *New Ways of Analyzing Variation 46*.

Levy Abitbol, J., Karsai, M., Chevrot, J.-P., Magué, J.-P., & Fleury, E. (2018). Socioeconomic and Network Dependencies of Linguistic Patterns in Twitter. *IC2S2 2018 - 4th Annual International Conference on Computational Social Science*.

Lexis, W. (1877). *Theorie Der Massenerscheinungen in Der Menschlichen Gesellschaft* (Wagner).

Loiseau, S., Gréa, P., & Magué, J.-P. (2011). Dictionnaires, Théorie Des Graphes et Structures Lexicales. *Revue de Sémantique et de Pragmatique*, (27), 51-78.

Loiseau, S., Magué, J.-P., & Heiden, S. (2009). *The TextometrieR package : Textual data analysis for social sciences and humanities*. useR ! Récupérée mars 2, 2016, à partir de https://halshs.archives-ouvertes.fr/halshs-00984192

Lotka, A. J. (1926). The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317-323.

Lottin, J. (1911). Le libre arbitre et les lois sociologiques d'après Quetelet. *Revue Philosophique de Louvain*, *18*(72), 479-515.

Magué, J.-P. (2002). Emergence in a population of agents of a lexicon based on an individual conceptualization.

Magué, J.-P. (2005). *Changements Sémantiques et Cognition : Différentes Méthodes Pour Différentes Échelles Temporelles*.

Magué, J.-P. (2006a). From Changes in the World to Changes in the Words : Lexical Adaptation. In *Evolutionary Epistemology, Language and Culture - A nonadaptationist systems theoretical approach* (p. 169-194). Springer.

Magué, J.-P. (2006b). Semantic Changes in Apparent Time. *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society*.

Magué, J.-P. (2007). On the Importance of Population Structure in Computational Models of Language Evolution. *31st Pennsylvania Linguistic Colloquium*.

Magué, J.-P. (2011). Amalia : Integrated Access to Digital Data and Documents in the Humanities and Social Sciences.

Magué, J.-P. (2014). Les Protocoles d'Internet et Du Web. In M. Vitali-Rosati & M. E. Sinatra (Éd.), *Pratiques de l'édition Numérique* (p. 129-144). Presses de l'Université de Montréal.

Magué, J.-P. (2018). Approches Sociolinguistiques et Computationnelles Du Français Sur Twitter. *Dynamique Des Communautés Sur Twitter En Période Électorale : Analyse Par Graphes Aléatoires*.

Magué, J.-P., Fleury, E., Karsai, M., & Quignard, M. (2015). Social, geographical and linguistic structure of the French speaking Twitter community.

Magué, J.-P., & Mabillot, V. (2015). Construire Un Site : Les Niveaux de Garrett. In J.-M. Salaün & B. Habert (Éd.), *Architecture de l'information-Méthodes, Outils, Enjeux* (p. 25-49). de Boeck.

Magué, J.-P., Quignard, M., Karsai, M., & Fleury, E. (2015a). Caractérisation Dialectale de Variabilité Linguistique Sur Twitter. *Langage, Cognition et Société (AFLiCo6)*.

Magué, J.-P., Quignard, M., Karsai, M., & Fleury, E. (2015b). Dialectal Characterization of Linguistics Variability on Twitter.

Magué, J.-P., Rossi-Gensane, N., & Halté, P. (2020). De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis. *Corpus*, (20).

Malamud, B. D., Morein, G., & Turcotte, D. L. (1998). Forest fires : An example of self-organized critical behavior. *Science (New York, N.Y.)*, *281*(5384), 1840-1842.

Malamud, B. D., & Turcotte, D. L. (1999). Self-Organized Criticality Applied to Natural Hazards. *Natural Hazards*, *20*(2), 93-116.

Mandelbrot, B. (1951). Adaptation d'un Message Sur La Ligne de Transmission. *Comptes-rendu de l'Académie des sciences*, *232*.

Mandelbrot, B. (1962). Towards a Revival of the Statistical Law of Pareto. *IBM Research Report*.

Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *The Journal of Business*, *36*(4), 394-419.

Mandelbrot, B. (1966). Information Theory and Psycholinguistics : A Theory of Word Frequencies. In Paul F. Lazarsfield & Neil W. Henry (Éd.), *Readings in Mathematical Social Sciences* (p. 151-168). MIT Press.

Mandelbrot, B. (1967). How Long Is the Coast of Britain ? Statistical Self-Similarity and Fractional Dimension. *Science*, *156*(3775), 636-638.

Mandelbrot, B. (1974). Intermittent turbulence in self-similar cascades : Divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics*, *62*(2), 331-358.

Mandelbrot, B. (1975). *Les objets fractals : forme, hasard et dimension*. Flammarion.

Mandelbrot, B. (2012). *The Fractalist : Memoir of a Scientific Maverick*. Knopf Doubleday Publishing Group.

Mangold, P., Léo, Y., Chevrot, J.-P., Fleury, E., Karsai, M., Magué, J.-P., Nardy, A., & Peuvergne, J. (2017). Optional realization of the French negative particule (ne) on Twitter : Can big data reveal new sociolinguistic patterns ?

Marconi, D. (1998). *La Philosophie du langage au vingtième siècle*. Editions de l'Eclat.

Mayer, T. (1750). *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748*. Bey Ioh. Paul Krauss, Buchh ??ndler ...; Bey der Hom ??nnischen Handlung druckts Ioh. Ioseph Fleischmann.

McLuhan, M. (2008). *The Gutenberg galaxy : The making of typographic man* (Repr.). Univ. of Toronto Pr.

Mille, A., & Magué, J.-P. (2012). Le Web : La Révélation Documentaire ? In B. Stiegler (Éd.), *Confiance, Croyance, Crédit Dans Les Mondes Industriels* (FYP Éditions, p. 230-245). FYP EDITIONS.

Napier, J. (1614). *Mirifici Logarithmorum Canonis Constructio*. Bartholomaeus Vicentium.

Newcomb, S. (1881). Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, *4*(1), 39-40.

Newcomb, S. (1895). *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. U.S. Government Printing Office.

Ortega y Gasset, J. (1930). *La Rebelión de Las Masas* (Revista de Occidente).

Ortega y Gasset, J. (2010). *La révolte des masses* (L. Parrot, Éd.). les Belles lettres.

Pareto, V. (1897). *Cours d'économie politique : professé à l'Université de Lausanne* (T. 2). F. Rouge.

Paugam-Moisy, H., Puzenat, D., Reynaud, E., & Magué, J.-P. (2002). Neural Networks for Modeling Memory : Case Studies. *Proceedings of the European Symposium on Artificial Neural Networks*, 71-82.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London. (A.)*, *54*, 329-333.

Pearson, K. (1924). *The Life, Letters and Labours of Francis Galton. Volume 2 : Researches of Middle Life* (Cambridge University Press).

Pédauque, R. T. (2006). *Le Document à La Lumière Du Numérique* (J.-M. Salaün, Éd.). C&F.

Pédauque, R. T. (2007). *La redocumentarisation du monde*. Cépaduès.

Peirce, C. S. (1873). On the Theory of Errors of Observations. In *Report of the Superintendent of the U.S. Coast Survey for the Year Ending June 1870.* (p. 200-224).

Peters, O., Hertlein, C., & Christensen, K. (2001). A Complexity View of Rainfall. *Physical Review Letters*, *88*(1), 018701.

Platon. (1922). *Phèdre : ou De la beauté des âmes* (M. T. Meunier, Trad.). Payot et Cie.

Plenz, D., Ribeiro, T. L., Miller, S. R., Kells, P. A., Vakili, A., & Capek, E. L. (2021). Self-Organized Criticality in the Brain. *Frontiers in Physics*, *9*.

Porter, T. M. (1986). *The Rise of Statistical Thinking, 1820–1900* (Princeton Universty Press).

Quemener, E., & Corvellec, M. (2013). SIDUS—the Solution for Extreme Deduplication of an Operating System. *Linux Journal*, *2013*(235), 3 :3.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale. Tome 1 / par A. Quételet,...*

Quetelet, A. (1846). De l'Influence du libre arbitre de l'homme sur les faits sociaux, et particulièrement sur le nombre des mariages, par M. Quételet,... *Bulletin de la commission centrale de statistique*, *3*.

Radonjić, A., Allred, S. R., Gilchrist, A. L., & Brainard, D. H. (2011). The Dynamic Range of Human Lightness Perception. *Current Biology*, *21*(22), 1931-1936.

Rheingold, H. (1993). *The Virtual Community : Homesteading on the Electronic Frontier*. Addison-Wesley Pub. Co.

Richter, C. F. (1935). An Instrumental Earthquake Magnitude Scale*. *Bulletin of the Seismological Society of America*, *25*(1), 1-32.

Roberts, D. C., & Turcotte, D. L. (1998). Fractality and Self-Organized Criticality of Wars. *Fractals*, *06*(04), 351-357.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328-350.

Rosch, E. (1975a). Cognitive Reference Points. *Cognitive Psychology*, *7*(4), 532.

Rosch, E. (1975b). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology : General*.

Rosch, E. (1977). Human Categorization. In N. Warren (Éd.), *Advances in Cross-Cultural Psychology I,* (Academic Press).

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. B. Lloyd (Éd.), *Cognition and Categorization* (p. 27-48). Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family Resemblances : Studies in the Internal Structure of Categories. *Cognitive Psychology*, *7(4)*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural Bases of Typicality Effects. *Journal of Experimental Psychology : Human Perception and Performance*, *2*, 491-502.

Saetnan, A. R., Lomell, H. M., & Hammer, S. (Éd.). (2010). *The Mutual Construction of Statistics and Society* (0e éd.). Routledge.

Salaün, J.-M. (2012). *Vu, lu, su : les architectes de l'information face à l'oligopole du Web*. La Découverte.

Shafer, T. (2017). *The 42 V's of Big Data and Data Science*. KDnuggets. Récupérée août 13, 2020, à partir de https://www.kdnuggets.com/the-42-vs-of-big-data-and-data-science.html/

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, *171*, 701-703.

Siracusa, J. (2017). Le penchant quetelésien de Durkheim. *L'Année sociologique*, *67*(1), 255-255.

Society, A. C. (1917). *Decennial index to Chemical abstracts, 1907-1916*.

Solé, R. V., & Manrubia, S. C. (1996). Extinction and Self-Organized Criticality in a Model of Large-Scale Evolution. *Physical Review E*, *54*(1), R42-R45.

Sornette, A., & Sornette, D. (1989). Self-Organized Criticality and Earthquakes. *Europhysics Letters*, *9*(3), 197.

Sornette, D., & Zajdenweber, D. (1999). Economic returns of research : The Pareto law and its implications. *The European Physical Journal B - Condensed Matter and Complex Systems*, *8*(4), 653-664.

Stiegler, B. (2008). *Aimer, s'aimer, nous aimer du 11 septembre au 21 avril*. Galilée.

Stigler, S. M. (1986). *The History of Statistics : The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press.

Stigler, S. M. (1999). *Statistics on the Table : The History of Statistical Concepts and Methods*. Harvard University Press.

Stross, B. (1969, septembre). *Language Acquisition by Tenejapa Tzeltal Children*.

Tarrade, L., Chevrot, J.-P., & Magué, J.-P. (2022). Detecting and Categorising Lexical Innovations in a Corpus of Tweets. *Psychology of Language and Communication*.

Thibert, C., & Magué, J.-P. (2016). Twitter as Corpus for Sociolinguistic Variationist Studies : Challenges of Using Sketchy Data. *Using Twitter for Linguistic Studies : Benefits and Difficulties*.

Thibert, C., Magué, J.-P., Fleury, E., Karsai, M., & Quignard, M. (2016). Dialectal Characterization of Linguistics Variability on Twitter. *Data Driven Approach to Network and Language*.

Thibert, C., Zeynaligargari, S., Quignard, M., & Magué, J.-P. (2016). Do You Tweet Like You Write or Like You Speak ?
Published : IC2S2 :

Thurner, S., Hanel, R. A., & Klimek, P. (2018). *Introduction to the Theory of Complex Systems*. Oxford University Press.

Turkle, S. (1997). *Life on the Screen : Identity in the Age of the Internet*.

Tversky, A., & Gati, I. (1978). Studies of Similarity. In E. Rosch & B. Llyod (Éd.), *Cognition and Categorization* (Erlbaum, p. 79-98).

Tversky, A. (1977). Features of Similarity. *Psychological Review*, *84*, 327-352.

University of Portsmouth. (2004). *Great Britain Historical GIS Project*. https://www.visionofbritain.org.uk/

Vidiella, B., Guillamon, A., Sardanyés, J., Maull, V., Pla, J., Conde, N., & Solé, R. (2021). Engineering self-organized criticality in living cells. *Nature Communications*, *12*(1), 4415.

Wallis, W. A., & Roberts, H. V. (1956). *Statistics, a new approach,* Free Press.

Wellman, B. (2001). Physical Place and Cyberplace : The Rise of Personalized Networking. *International Journal of Urban and Regional Research*, *25*(2), 227-252.

Wells, J. (2017). Marx Reads Quetelet : A Preliminary Report.

White, T. (2012, mai 19). *Hadoop : The Definitive Guide*. "O'Reilly Media, Inc."

Wilson, K. G. (1983). The Renormalization Group and Critical Phenomena. *Reviews of Modern Physics*, *55*(3), 583-600.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicu* (Harcourt, Brace and Company).

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trad.). Macmillan.

Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, *213*, 21-87.

Zipf, G. K. (1942). The Unity of Nature, Least-Action, and Natural Social Science. *Sociometry*, *5*(1), 48-62.

Crédits photographiques

 Figure 2.3, page 31 : Collection Kharbine-Tapabor

 Figure 3.6a, page 52 : CC-BY-SA, Père Igor

 Figure 3.6b, page 52 : CC-BY-SA, Judicieux

 Figure 3.6c, page 52 : CC-BY, abdallahh

 Figure 3.7a, page 53 : CC-BY-NC-ND, Mathieu Bruc

 Figure 3.7b, page 53 : Domaine public

 Figure 3.7c, page 53 : CC-BY, Jeanne Menjoulet

 Figure 3.8a, page 54 : CC-NC-ND, Off. de Tourisme de ST-Jean-de-Luz

 Figure 3.8b, page 54 : CC-BY, Jean-Pierre Dalbéra

 Figure 3.8c, page 54 : copyrigth, Google

 Figure 4.10d, page 75 : CC-BY, Eglash

Annexes

Curriculum vitæ

JEAN-PHILIPPE MAGUÉ

Associate Professor in linguistics and digital humanities at ENS de Lyon

CONTACT

Address	ENS de Lyon – Site Descartes Bureau R182 15 parvis René Descartes – BP 7000 69342 Lyon Cedex 07 FRANCE
Email	jean-philippe.magueé@ens-lyon.fr
Phone	+33 (0)4 37 37 64 23
Home page	perso.ens-lyon.fr/jean-philippe.magueé

ACADEMIC POSITIONS AND AFFILIATIONS

2009-present	**Associate Professor in linguistics and digital humanities** ENS de Lyon, ICAR lab
2020-present	**Deputy Director** Rhône-Alpes Complex Systems Institute (IXXI)
2018-present	**Deputy Director** Education and Digital Humanities Department, ENS Lyon
2017-present	**Visiting scientist at Johns Hopkins University, Center for Astrophysical Sciences**
2016-present	**Member of the executive committee** Excellence Laboratory for Advanced Studies on Language Complexity (Aslan)
2017	**Visiting scholar at INRIA, Dante Team**
2016	**Visiting scholar at CNRS, ICAR Lab**
2012-2019	**Member of the executive committee** Rhône-Alpes Complex Systems Institute (IXXI)
2009	**Postdoc,** ENS de Lyon, ICAR lab.
2008	**R&D Engineer,** KSL, Charbonnière-les-bains.
2007	**Postdoc,** Projet CL^2, Lyon 2 University, DDL Lab.
2006	**Postdoc,** Department of Linguistics, University of Chicago, Chicago, IL, USA.
2005	**ATER,** IUP Documentation d'entreprise, réseaux et image, Bourgogne University, Dijon.
2003	**Visiting scholar,** Linguistics Department, University of California, Berkeley, CA, USA.

EDUCATION

2005	**PhD Cognitive Science**, Lyon 2 University
2001	**M.S Cognitive Science**, Lyon 2 University
2001	**M.S Computer Science**, Grenoble 1 University
1998	**B.S Mathematics** Paris 11 University

TEACHING

2018 – present	**Datamining**, ENS de Lyon
2018 – present	**Natural language processing**, ENS de Lyon
2012 – present	**History and current issues of the Internet,** ENS de Lyon

2012 – present	**Project management,** ENS de Lyon
2010 – present	**Introduction to Python,** ENS de Lyon
2012 – present	**Current issues in Digital Humanities,** ENS de Lyon
2018 – 2019	**Technologies for the web,** ENS de Lyon
2015 – 2017	**Mooc Information architecture.**
2016	**Digital Editing and Publishing,** ENS de Lyon
2012 – 2015	**Textual data management,** ENS de Lyon
2014	**Digital collections management,** B.A. Cultural Heritage, UGB, St Louis, Senegal
2009 – 2012	**Digital editions for humanities,** ENS de Lyon
2009 – 2010	**Textual corpus analysis,** ENS de Lyon
2009 – 2011	**Readings in linguistics,** ENS de Lyon
2004 – 2005	**Digital document management,** Bourgogne University
2004 – 2005	**Database,** Bourgogne University
2004 – 2005	**Algorithms and data structures,** Bourgogne University
2004 – 2005	**HTLM and web technologies,** Bourgogne University
2002 – 2003	**Introduction to semantics,** Lyon 2 University
2002 – 2003	**Algorithms and data structures,** Lyon 1 University

PUBLICATIONS

2022	Jean-Philippe Magué. Intelligence artificielle, représentations sociales et biais. Équité, diversité et inclusion dans un contexte numérique, Entretiens Jacques Cartier 2022, Nov 2022, Ottawa, Canada
2022	Jean-Philippe Magué, Louise Tarrade, Jean-Pierre Chevrot, Mélanie Veloso. Using AI for doing sociolinguistics? Representations of gendered social cues in neural networks Sociolinguistics Symposium 24, Jul 2022, Ghent, Belgium
2022	Louise Tarrade, Jean-Philippe Magué, Jean-Pierre Chevrot. Detecting and categorizing lexical innovations in a corpus of tweets Psychology of Language and Communication, 2022, 26, pp.313-329.
2021	Louise Tarrade, Jean-Pierre Chevrot, Jean-Philippe Magué. Buzz or Change: How the Social Network Structure Conditions the Fate of Lexical Innovations on Twitter. *8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)*, Oct 2021, Nijmegen, Radboud University, Netherlands.
2020	Jean-Philippe Magué, Nathalie Rossi-Gensane et Pierre Halté, « De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis », Corpus
2019	Jean-Pierre Chevrot, Jacob Levy Abitbol, Màrton Karsai, Jean-Philippe Magué, Eric Fleury. Variations du (ne) négatif du français dans Twitter. Que peut apporter l'étude des données massives aux questions de sociolinguistique ?. CILPR 2019 - XXIXe Congrès international de linguistique et de philologie romanes, Jul 2019, Copenhague, Danemark.
2018	Jacobo Levy Abitbol, Màrton Karsai, Jean-Pierre Chevrot, Jean-Philippe Magué, Éric Fleury. Socioeconomic and network dependencies of linguistic patterns in Twitter. IC2S2 2018 - 4th Annual International Conference on Computational Social Science, Jul 2018, Evanston, Illinois, United States
2018	Jacobo Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, Eric Fleury. Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis. The Web Conference 2018, Apr 2018, Lyon, France.
2017	Jacobo Levy Abitbol, Jean-Pierre Chevrot, Màrton Karsai, Jean-Philippe Magué, Yannick Léo et al. The optional realization of the French negative particle (ne) on Twitter: Space, status and time. New Ways of Analyzing Variation 46, Nov 2017, Madison, United States
2017	Jacobo Levy Abitbol, Márton Karsai, Jean-Pierre Chevrot, Jean-Philippe Magué, Eric Fleury. Socioeconomic and network dependencies of linguistic patterns in Twitter. COMPLEX NETWORKS 2017 - 6th International Conference on Complex Networks and Their Applications, Nov 2017, Lyon, France
2017	Jacobo Levy Abitbol, Márton Karsai, Jean-Pierre Chevrot, Jean-Philippe Magué, Eric Fleury. How social, economic and demographic forces shape linguistic variation on Twitter. POPLANG 2017 - Workshop Population

	effects on languages: Modelling population dynamics and language transmission from the perspective of language learning, contact and change, Nov 2017, Lyon, France.
2017	Jacobo Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, Eric Fleury. Optional realisation of the French negative particle (ne) on Twitter: Can big data reveal new sociolinguistic patterns? *CCS 2017 - Conference on Complex Systems* , Sep 2017, Cancun, Mexico
2017	Paul Mangold, Yannick Léo, Jean-Pierre Chevrot, Eric Fleury, Màrton Karsai, Magué JP, Nardy A, Peuvergne J, Optional realization of the French negative particule (ne) on Twitter: Can big data reveal new sociolinguistic patterns? ICLAVE 9 2017 - International Conference on Language Variation in Europe, Jun 2017, Malaga, Spain
2016	Thibert C, Zeynaligargari S, Quignard M, Magué JP. *Do You Tweet Like You Write or Like You Speak ?: Part of Speech Distribution of French Speaking Communities in Twitter*. IC2S2 : 2nd Annual International Conference on Computational Social Science, Evanston, United States. 2016,
2016	Thibert C, Magué JP, Fleury E, Karsai M, Quignard M. Dialectal Characterization of Linguistics Variability on Twitter. Data Driven Approach to Network and Language, Lyon, France. 2016,
2016	Thibert C,, Magué JP. Twitter as Corpus for Sociolinguistic Variationist Studies : Challenges of Using Sketchy Data. In Using Twitter for Linguistic Studies : Benefits and Difficulties. Canterbury, United Kingdom: Kent University.
2015	Magué JP, Fleury E, Karsai M, Quignard M. *Dialectal characterization of linguistics variability on Twitter*. 1st International Conference on Twitter for Research, Lyon, France.
2015	Chevrot, J.-P., Nardy, A., Fleury, E., Karsai, M., Magué, J.-P. Sociolinguistique et sciences cognitives: l'individu, le collectif et le réseau. Journées FLORaL-PFC 2015 : la base de données Phonologie du Français Contemporain dans le champ phonologique
2015	Château E., Beaugiraud V., Boschetto S., Boulai C., Gedzelman S., Ingarao M., Jallud, P.-Y., Morlok E., PONS P., SAÏDI, S. , Magué J.-P. *SynopsX A Lightweight Xquery-Based Framework to Easily Publish and Expose XML Corpora*. Text Encoding Initiative, Conference and members meeting, lyon, France.
2015	Magué, J.-P., Fleury F., Karsai M., Quignard, M. *Dialectal characterization of linguistics variability on Twitter.* ICCSS 2015, Helsinki, Finland.
2015	Magué, J.-P., Fleury F., Karsai M., Quignard, M. *Social, geographical and linguistic structure of the French speaking Twitter community,* NetSci2015, Zaragoza, Spain.
2015	Magué, J.-P., Mabillot V. *Construire un site*. In Jean-Michel Salaün; Benoît Habert (Eds). *Architecture de l'information : Méthodes, outils, enjeux*. De Boeck.
2015	Magué, J.-P., Fleury F., Karsai M., Quignard, M.. *Caractérisation dialectale de variabilité linguistique sur Twitter* Language, Cognition and Society (AFLiCo 6), Grenoble, France
2014	Magué, J.-P. *Les protocoles d'Internet et du web*. In Sinatra, M & Vitali-Rosati, M. Pratiques de l'édition numérique, Presses de l'Université de Montréal, pp.129-144.
2012	Mille A., Magué J.-P. : *Le Web : la révélation documentaire ?*. In Stiegler, B. Confiance, croyance, crédit dans les mondes industriels, fyp editions, 2012, Nouveau monde industriel
2011	Magué, J.-P. *Amalia : Integrated access to digital data and documents in the humanities and social sciences.* EsciDoc Days 2011, Berlin.
2011	Beaugiraud V., Gedzelman S., Ingarao M., Magué J.-P., Saïdi S. *Amalia : an eSciDoc based solution to manage the production, processing and publishing workflows of TEI data.* TEI2011, Würzburg.
2011	Loiseau S., Gréa P. & Magué J.-P. *Dictionnaires, théorie des graphes et structures lexicales*, Revue de sémantique et de pragmatique.
2010	Heiden S., Magué J.-P., Pincemin B.. *TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement.* In Bolasco S., Chiari I., Giuliano L.(eds), *Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010*, Edizioni Universitarie di Lettere Economia Diritto..
2007	Magué, J.-P. *On the importance of population structure in computational models of language evolution*. Proceedings of the 31st Pennsylvania Linguistic Colloquium.
2006	Magué, J.-P. *Semantic change in Apparent Time*. Proceedings of the 32nd meeting of the BLS. Berkeley, CA: Berkeley Linguistics Society.
2005	Magué, J.-P. *From Changes in the World to Changes in the Words: Lexical Adaptation*. In Gontier, N., Van Bendegem, J.-P., and Aerts, D. EELC – A nonadaptationist systems theoretical approach. Springer.
2004	Grataloup, C., Magué, J.-P. & Meunier, F. *Noun ending and gender predictability in French.* The Fourth International Conference on the Mental Lexicon, Windsor, Ontario, Canada. (Communication affichée).
2003	Grataloup, C., Magué, J.-P. & Meunier, F. *Noun ending predictability in French.* Conference of the European Society for Cognitive Psychology, Granada, Spain. (Communication affichée).

2002	Magué, J.-P. *Emergence in a population of agents of a lexicon based on an individual conceptualization.* Annual meeting of the European Society for Philosophy and Psychology, Lyon, France. (Communication affichée).
2002	Paugam-Moisy, H., Puzenat, D. , Reynaud, E. & Magué, J.-P. *Neural networks for modeling memory: case studies.* In Proc. of ESANN, European Symp. on Artificial Neural Networks, Bruges, Belgium.

INVITED SPEAKER

2021	Diffusion des variants (linguistiques) dans la population : le ou la covid ? Evolyon conference. Nov 2021.
2021	Approche computationnelle de la variation linguistique sur Twitter. Scidolyse meeting, 2May 021.
2018	*Approches sociolinguistiques et computationnelles du français sur Twitter.* Dynamique des communautés sur Twitter en période électorale : analyse par graphes aléatoires, Apr 2018, Grenoble, France
2016	*La notion de réseau social en sociolinguistique computationnelle.* Socionet, Rencontres interdisciplinaires sur les réseaux sociaux : description, données, modélisation, interpretation, 6-8 Juin 2016 / ENS de Lyon.
2015	*Le Master Architecture de l'information de l'ENS-Lyon Réponse à un changement de paradigme documentaire.* Journée d'étude DHNord 2015, Maison Européenne des Science de l'Homme, Lille.
2015	*Outille-toi toi-même,* Séminaire de la Cellule Corpus Complexes, Labex Aslan, Lyon.
2014	*Se construire les humanités numériques : apprendre en action,* Journée DARIAH-FR. Apprendre et enseigner quand les humanités deviennent numériques.
2014	*Un Introduction aux Digital Humanities,* Séminaire de la Haute Ecole de Gestion, Genève, Suisse.
2012	*You talkin' to me ? Complexité des langues comme objets sociaux,* Séminaire du Labex Aslan, Lyon.
2011	*Que sont les Humanités Numériques ?,* Biennale du Numérique, ENSSIB
2011	*Humanités Numériques : Histoire, enjeux et structuration institutionnelle,* Séminaire du Centre Blaise Pascal, ENSL.
2006	*A small world in the brain: Infering functional connectivity properties from automatic literature analysis,* Séminaire du Human Neuroscience Laboratory, University of Chicago, USA.
2006	*Physical limits of computation and the nature of Language*, Séminaire du Chicago Language Modelling Lab, University of Chicago, USA.
2005	*Évolution des Structures Conceptuelles.* Séminaire du laboratoire Dynamique du Langage, Lyon, France.
2004	*Changements Sémantiques et Analyse de Corpus.* Journée d'étude du projet Prox-Dilan, Toulouse, France.
2004	*Géométrisation du Sens.* Fédération de typologie - groupe rapprochements sémantiques, Villejuif, France.

SOFTWARE

2015	**SynopsX,** XML corpus publication plateform (github.com/synopsx)
2009	**Textometrie,** Textual analysis(textometrie.sourceforge.net)
2006	**Axivsorter,** sorting papers posted on the website arxiv.org (www.arxivsorter.org)

Travaux choisis

Le Web : la révélation documentaire ?

Le Web : la révélation documentaire ?

Alain Mille, Jean-Philippe Magué

Le Web : la révélation documentaire ?

Manifeste pour un fait documentaire sur le web

Alain Mille et Jean-Philippe Magué

Le web est un lieu (système ?) d'information *formidable* permettant à un nombre sans cesse croissant d'internautes de l'interroger et de le nourrir avec des méthodes de plus en plus efficaces, reposant sur des moteurs de recherche sophistiqués et sur des générateurs de contenus puissants et connectés à des masses de données en croissance continue.
Les activités humaines, en particulier intellectuelles mais pas seulement, s'exercent de plus en plus fréquemment en connexion étroite avec ce lieu d'information gigantesque et pourtant facile d'accès.
Le web fournit à l'évidence une fonction documentaire puisqu'il permet de fournir les informations utiles, nécessaires et valides à l'organisation de l'activité. En effet, c'est parce que l'information accessible est considérée comme fournissant la documentation nécessaire à l'activité en cours, qu'elle obtient un statut de validité (provisoire)car permettant les actions satisfaisant l'activité en cours.
Pourtant, le « fait documentaire » n'est pas identifiable facilement et l'internaute qui souhaite capitaliser les informations mobilisées dans ses interactions sur le web qui lui ont permis de mener à bien son activité en est bien incapable.
Dans cet article, nous nous intéressons d'abord à situer la lecture/écriture sur le web par rapport aux autres modes de lecture/écriture permettant de faire document, et nous terminons par un *manifeste* pour installer dans le lieu d'information du web les conditions d'une mise en œuvre d'un fait documentaire d'un nouveau genre, lié à la spécificité du lieu de lecture/écriture que constitue le web.
Nous commencerons par introduire quelques notions que nous utiliserons dans le document sans pour autant les considérer comme universelles naturellement. L'argumentaire du manifeste sera articulé par un point de vue sur le processus documentaire en choisissant des jalons qui nous ont semblé les plus importants à considérer pour penser le fait documentaire au sein du web. Nous terminerons par un *manifeste* dont nous espérons qu'il ouvrira le chemin à des développements théoriques et pragmatiques afin d'ouvrir le web à un fait documentaire en lui ajoutant les propriétés de mémoire et de preuve nécessaires pour y parvenir.

Introduction aux notions manipulées dans le manifeste

- **Inscription de connaissance** : Dans le contexte de ce manifeste, une inscription de connaissance est "toute inscription dans l'environnement qui fait sens pour celui qui l'observe et l'interprète, dans le cadre d'une activité cognitive". Il s'agit d'une inscription "de connaissance" car elle permet l'action à celui qui sait l'interpréter.

- **Ecriture-Lecture symbolique** : avec l'écriture et la lecture symbolique, le signe est tracé *intentionnellement* à la suite d'un processus de formation de sa signification. Le signe devient alors support possible pour le partage, l'apprentissage et la constitution de mémoire, démultipliant à chaque fois chacun de ces processus (partage, apprentissage et mémoire).
- **Fait documentaire** : cet intitulé recouvre tout à la fois le processus documentaire et ce qu'il en ressort comme "produit" (en général un document).C'est le **processus documentaire** qui permet d'établir la validité des informations qui sont concernées par la documentation en cours.
- **Processus documentaire** : Le processus documentaire se décompose en un **processus auctorial** et un **processus éditorial pouvant être, selon les types de documents et les époques, plus moins intriqués**.
    - le **processus auctorial** est le processus de construction du texte. C'est le moment ou l'auteur rassemble les connaissances qu'il va mettre en texte : soit par introspection/reflexion, soit par documentation.
    - Le **processus éditorial** est le processus au terme duquel l'éditeur introduit le document dans le système documentaire en le dotant de propriétés sociales (d'un statut social) qui influeront, à l'instar de son contenu, sur la perception du document par le lecteur. Ces propriétés renseigneront notamment le lecteur sur le contrat de lecture adopté et sur la confiance qu'il peut ou non placer dans le document.
    - **Auteur** et **Editeur** : Ce sont les rôles des responsables respectivement des processus auctorial et éditorial. Ces rôles pouvent être distribués sur une ou plusieurs personnes et/ou institutions, selon le degré d'intrication des processus auctorial et éditorial.
- Le **"produit documentaire"/ document** : ce que l'on souhaite garder à l'issue du processus documentaire.Un **"produit documentaire"/ document** peut être révisé par un nouveau processus documentaire complétant le premier et débouchant sur une nouvelle version du produit. Il s'agit de différentes versions d'un même document, car on considère une **racine commune** dans le processus documentaire. Même dans le cas d'une inscription numérique, c'est le document qui est versionné, pas l'inscription numérique (le fichier), objet bien plus général, et les informations de versions sont associées directement au document.

Jalons proposés pour explorer le fait documentaire

1. **L'ère des inscriptions basiques** (non symboliques) relevant d'un processus d'inscription non intentionnelle ou intentionnelle :
    a. L'inscription non intentionnelle de connaissance : l'inscription est considérée comme telle par l'observateur expérimenté qui la considère comme "indice" pour agir. La fumée, les empreintes sont considérées comme constituant la trace d'une activité reconnue par le chasseur par exemple. Indice pour le chasseur, il est un signe « naturel ».
    b. L'inscription peut être intentionnelle : celui qui inscrit *inscrit* à dessein pour enregistrer une connaissance. Cette inscription n'est pas nécessairement symbolique. Une trace sur une branche change de statut dès lors qu'est n'est plus laissée par accident, mais dans un but de marquer, d'inscrire une connaissance. La trace devient un *acte de communication*, et son auteur postule que le futur observateur, le lecteur de la trace, saura que celle-ci est

intentionnelle, pourvoyeuse de connaissances. Car c'est en sachant cela qu'il entreprendra de l'interpréter et que la trace pourra devenir support d'une connaissance et d'une mémoire partagée (ce qui n'était évidemment pas le cas dans l'inscription non intentionnelle).

Le passage au symbolique : Une fois qu'est institué ce mode fonctionnement, où l'inscripteur/auteur sait qu'un lecteur (au moins un) sait interpréter et que le lecteur sait que l'auteur savait qu'il saitet qu'il se met effectivement à interpréter l'inscription, des conventions d'interprétation peuvent se mettre en place. Ces conventions d'interprétation fournissent à l'inscription son caractèresymbolique. L'écriture et la lecture de traces constituées de signes symboliques (qui nous intéressent ici) nécessitent un processus "éditorial" accepté pour que les inscriptions puissent faire foi et être considérées comme fiables pour représenter l'indice « naturel » et s'imposant comme vrai par sa naturalité héritée. D'une trace « naturelle » considérée comme indicielle, on tire un "tracé" symbolique procurant les éléments nécessaires à la documentation de l'observation et nécessitant donc une compétence reconnue pour être considéré comme signifiant selon des conventions apprises et partagées.

2. **L'ère de l'écriture/lecture symbolique** : A partir de cette première révolution de l'écriture-lecture symbolique, nous considérons deux évolutions majeures.
    a. Les dispositifs de recopie permettent un partage plus grand et la constitution de mémoires externes. Pour lire-écrire et même recopier, il faut être "expert" et autorisé. La copie est "signée" d'une manière ou d'une autre pour attester de sa valeur. Le processus éditorial permet d'élaborer le "document" qui se concrétise matériellement et devient autonome, mais étroitement contrôlé par les experts de l'écriture-lecture symbolique (typiquement les clercs).Les membres de la société sont pour une grande majorité exclus du système documentaire. Les auteurs et lecteurs ne forment qu'un petit groupe d'initiés[1].
    b. Les dispositifs d'imprimerie massifient la recopie, formalisent différemment le processus éditorial. L'autonomie du document devient plus grande, sa circulation est plus facile, l'apprentissage des processus d'écriture-lecture s'étend au delà des cercles des clercs. La mémoire et le partage peuvent échapper au contrôle des tenants de l'expertise de l'écriture-lecture,même si les universités gardent un rôle de conservation des savoirs, des documents dorénavant imprimés. D'une certaine façon, les scriptoriasont alors remplacés par les industries de l'édition et del'imprimeriequi se mettent en place.Toutefois, les livres plus faciles à reproduire en grand nombre, moins coûteux deviennent accessibles à un plus grand nombre. Les inscriptions de connaissance deviennent reconnues sous une forme matérielle de plus en plus partagée. La massification et la démocratisation introduisent de nouveaux usages, facilitent l'émergence d'une nouvelle société en introduisant des processus sociaux totalement nouveaux. La révolution documentaire prépare la révolution des esprits.

    Dans ces deux étapes, le document *produit* est premier pour le lecteur, le processus éditorial est rappelé par une "signature" dont la valeur varie fortement selon

[1] [Le nombre de lecteurs est toutefois significatif] : à partir des Ier et IIème siècles après JC, dans la Rome impériale, par exemple, Le nombre de lecteurs croît : « Le nouveau lecteur, dans les premiers siècles de l'Empire, est quelqu'un qui n'est plus (ou pas seulement) « obligé » de lire par sa condition d'homme de lettres, de fonctionnaire, civil ou militaire, d'enseignants ou d'élève, ou par les besoins techniques d'une quelconque profession, c'est une lecteur « libre » qui lit par plaisir, par habitude, ou pour le prestige de la culture. » (Cavallo, 1995).

l'autorité (dans tous les sens du terme) qui garantit ce processus et le processus auctorial n'est présent que par l'auteur et encore. Le processus documentaire est explicite, important, reconnu, mais n'est pas gardé, en général comme démonstration de la valeur du document. Ce processus est pourtant considéré par les chercheurs (historiens par exemple) comme preuve nécessaire à l'explication d'un document. Seul, le "certificat" délivré par l'autorité éditoriale rappelle ce processus. Après sa fabrication, au moment de son usage dans l'activité par le lecteur**, le document est premier, le processus documentaire est second**.

3. L'ère des inscriptions numériques
    a. Il est utile de rappeler que les premiers dispositifs numériques n'étaient pas interactifs et le codage des informations à traiter était homomorphe au codage interne de l'information. Par exemple, les cartes perforées, les rubans perforés, l'entrée des instructions de démarrage « aux clés », laissaient paraître directement le codage binaire sous-jacent. Le programmeur « expert » était le médiateur obligatoire de toute écriture. Le listing de résultat, offrait une « impression » du résultat du traitement, non homomorphe au codage interne, mais impossible à considérer directement pour une ré-écriture dans le code d'entrée. Le programmeur disposait d'impressions « spéciales » lui donnant le « dump » de la mémoire dans un codage tel qu'il pouvait y retrouver le codage interne (binaire, octal, hexadecimal,.). Dans cette variante des inscriptions numériques, le modèle était très proche de celui de l'imprimerie (construction par un technicien, le typographe qui écrivait en miroir et à l'envers, pour une lecture selon une norme sociale). L'ère du numérique introduit toutefois une différence de taille, c'est un traitement qui produisait le « document » lisible à partir d'entrées codées programmant cette production à partir de données elles-mêmes codées. Dans ces documents *calculés*, l'intentionnalité de l'auteur disparait, ou du moins se dilue grandement : l'intentionnalité du programmeur se retrouve dans le codedu programme qui calcule le document et l'intentionnalité des producteurs de données se retrouve dans les codes des données. Le lecteur continue à agir en postulant l'intentionnalité d'un *auteur*, qui pourtant n'existe plus directement. Le processus documentaire se complexifie, introduisant du calcul explicite pour le programmeur. La « preuve » documentaire devrait normalement s'accompagner de la preuve du traitement, mais ce n'est vrai que rarement. En pratique, pour le « lecteur-utilisateur » des productions documentaires, le processus documentaire restait largement inconnu. Le document produit restait premier, le processus documentaire disparaissaitou du moins s'obscurcissait fortement pour le lecteur utilisateur (le processus éditorial n'avait plus de signature reconnue et l'auteur n'apparaissait plus, et pour cause).
    b. Assez rapidement, les informaticiens inventent des procédés pour faciliter leurs inscriptions numériques. De saut technologique en saut technologique, les éditeurs permettant de coder les programmes et les chaînes de développement s'automatisent. Les codes utilisés pour programmer s'éloignent des codages internes (même s'ils nécessitent de les comprendre pour coder un traitement ou des données). Le cycle de programmation/vérification se raccourcit avec des outils de plus en plus interactifs pour le codage des traitements et des outils de plus en plus intégrés pour le codage de l'information. Une nouvelle classe de logiciels, les outils « bureautique » sont proposés, donnant le rôle de « codeur » à l'utilisateur.

Les « formes » de l'écriture sont calquées sur les formes de la lecture auxquelles l'utilisateur est formé par ailleurs (en dehors d'une formation à l'informatique). C'est le « What You See Is What You Get ». Le traitement est « caché » autant que possible par les interfaces interactives et graphiques. Les processus auctoriaux et, dans une moindre mesure, les processus éditoriaux classiques sont « mimés » avec l'élaboration d'un document destiné à être « imprimé » sans l'intervention des « typos » (traitement de texte numérique) ou l'élaboration d'un calcul sans l'intervention du programmeur (feuille de calcul), ou la production de données (remplissage de feuilles de calcul). Dans tous les cas, c'est la forme de lecture qui dicte la forme d'écriture. Pour **faire document**, le processus documentaireest explicite, le nommage du fichier-document est lié à l'évolution du processus éditorial qui se déploie sous la forme du « versionning », les informations de validité sont intégrées (signatures, certificats, propriété, etc.) dans une description du document qui existe virtuellement indépendamment du fichier informatique sous-jacent (le fait documentaire est codé dans le fichier, l'outil de présentation/édition ne présente ces descriptions documentaires que si on les demande et si on sait les demander). On parle de virtualisation documentaire car ce n'est pas le document lui-même qui est inscrit, mais une version codée de celui-ci, même si naturellement cette version codée sous-jacente est parfaitement concrete mais le plus souvent inconnu de celui qui écrit-lit. La concrétisation du document, en tant que document présenté pour sa lecture selon les normes sociales et apprises se fait sur un support visuel éphémère (les écrans) ou consommable (écriture matérielle sur papier). Lorsque le document est « prêt », il peut être soit imprimé soit fourni sous un format numérique plus ou moins « stable » (PDF, e-book, …), Le document (numérique) reste premier mais son caractère éphémère apparaît plus clairement à l'utilisateur s'il le consulte par l'intermédiaire d'un outil interactif de visualisation. Il est alors observable que ce que l'on est amené à lire est le résultat d'une construction immédiate par l'outil utilisé, ce que l'on peut d'ailleurs constaterpar des modifications liées au contexte de lecture, comme par exemple le choix sous-jacent de l'imprimante associée au document. Le processus documentaire (essentiellement le processus auctorial) devient accessible, il est « vécu » par un nombre croissant d'utilisateurs qui peuvent donc se le « figurer ».Les outils numériques rendent plus accessibles le rôle d'auteur. On passe donc d'une société ou il y avait très peu d'auteurs en proportion des lecteurs à une société qui comprend beaucoup d'auteurs. Le processus auctorial, qui était jusqu'alors l'affaire d'une petite minorité et méconnu de la majorité se généralise et devient banal. Le processus éditorial est lui aussi outillé (notamment par des fonctionnalités offertes par les logiciels de bureautique, peu utilisées sinon par des experts), mais il mime les mécanismes mis en place pour les documents papier. Dans les entreprises et les institutions par exemple, le processus éditorial, c'est-à-dire, l'insertion d'un document dans le système documentaire par l'adjonction de propriétés socialement reconnues, se matérialise par la diffusion, le classement et l'éventuel archivage du document. Si les systèmes de gestion électronique de document (GED) permettent d'automatiser ces opérations pour des documents électroniques, celles-ci restent les mêmes que dans les modes d'organisation établis pour les documents papier.

Même s'il est maintenant pour une bonne part familier pour tous, le processus documentaire reste second pour un lecteur (non auteur du document qu'il lit) car simplement représenté, en partie et dans le meilleur des cas sous la forme d'une signature plus ou moins formelle (propriétés du document, dates, auteur, révisons, incluant des informations de validation, etc.). Le document présenté à la lecture est le plus souvent fidèle au document tel que l'auteur l'avait pensé pour la lecture.

c.L'invention du Web introduit une évolution majeure dans le processus documentaire. D'emblée, il est proposé que le code de description d'un fragment documentaire devienne explicite (html) et que l'outil de lecture/écriture se normalise avec un protocole unique (http). La « page » qui s'affiche est éphémère par construction, construite à la volée par le navigateur qui en reçoit une description comprenant à la fois ce que l'on peut donner à lire et ce que l'on peut donner à écrire. La notion de « ressource » est inventée pour représenter des choses très variées (fragments documentaires, données, objets, personnes, etc.) mais qui doivent pouvoir être « adressées » de manière unique (URI). Cette notion de ressource n'est pas fermée, elle implique simplement l'idée qu'une ressource sera mobilisable (selon une méthode qu'il faut connaître pour y parvenir) pour produire une **interaction** (une potentielle lecture-écriture). Pour manipuler ces ressources et produire les interactions utiles dans des contextes d'activité, le web s'augmente de nombreux dispositifs d'indexation facilitant des démarches de découverte active. Cette évolution n'est pas sans rappelé l'émergence de la lecture scholastique. Hamesse(1995) décrit celle-ci comme étant une conséquence de croissance rapide de la production documentaire à partir du XIIe siècle, lorsque les livres quittent les monastères pour les universités. Cette croissance nécessite le développement de nouveaux modes d'organisation des documents permettant un accès plus direct à l'information que ne le permet la lecture linéaire. Les documents se structurent alors avec des paragraphes, des sections avec des titres et sont équipés de tables, index et concordances permettant un repérage rapide de l'information : « La lecture continue et chronologique d'une œuvre qui se faisant lentement, permettant d'assimiler sinon l'ensemble, au moins la substance d'un ouvrage, va céder le pas à une lecture fragmentaire et morcellée []. » (ibid., p129)

Sur le web cette fragmentation s'accentue. La production de ressources est l'affaire d'un spectre très large d'utilisateurs : les experts naturellement, de plus en plus les utilisateurs eux-mêmes et des « robots » de fabrication de contenus à partir de ressources alimentées par des processus dynamiques. Ces ressources sont autant de fragments rassemblés dynamiquement dans l'espace de lecture-écriture (un navigateur le plus souvent encore) par des processus dynamiques associant « calcul » et « choix de l'utilisateur » dans une co- à considérer comme mobilisées pour l'action (et pourtant c'est souvent ce qui est mis dans les « favoris » en guise de fait documentaire).

Le web est un dispositif hautement interactionnel qui permet l'action sans qu'un document ne soit jamais considéré comme présent en tant que tel, mais le processus documentaire (« je me documente pour agir ») est très conscient car mobilisant l'utilisateur concrètement. **C'est sur le « chemin documentaire » que le lecteur-scribe a construit que l'action s'articule. Le processus documentaire devient plus important, le document se dilue.**

Le lecteur-scribe manipule les fragments documentaires de la même manière qu'un auteur utilise des sources, des proto-documents (Pédauque, 2006), lors du processus auctorial, si ce n'est que le document final n'est jamais produit. Le lecteur est co-auteur des proto-documents qu'il consulte et reste proto-auteur d'un document final jamais finalisé.

La prédominance du processus documentaire est double, elle apparaît à 2 niveaux : d'une part l'utilisateur est impliqué dans la (co-)construction de chaque fragment donc dans le processus de documentaire de chacun de ces fragment ; d'autre part, ces fragments sont les proto-documents d'un processus documentaire (inachevé) de plus haut niveau.

Y-a-t-il "fait documentaire" ? La question mérite d'être posée. En effet, la plupart du temps l'action se déroule dans le même temps que le processus documentaire et n'a pas nécessité un **temps** documentaire particulier pour récapituler tout ce qui a été intéressant par rapport à l'activité. C'est parfois une "information" présentée dans un fragment qui fait sens pour l'activité (un horaire de train, une adresse, le numéro de téléphone...) sans que le fragment n'apparaisse particulièrement fabriqué pour cette information. Et pourtant, au contraire, c'est la démonstration que l'utilisateur est co-auteur des proto-documents, que son intentionnalité est à l'origine du fragment, et que c'est précisément pour cette raison que l'information nécessaire pour l'action y est intentionnellement présente. Si l'utilisateur est co-auteur, le concepteur du traitement en est l'un des multiples *éditeurs* et le propriétaire des données sur lesquelles se font les traitements est un des multiples co-auteurs. C'est en effet le concepteur du traitement (qui permet la présentation du fragment documentaire à l'utilisateur) qui s'est chargé de la *fonction éditoriale* de ce fragment qui permet de donner une forme mais aussi une *confiance* au contenu documentaire du fragment proposé. C'est lui qui est chargé de *publier* le contenu mis à disposition par un propriétaire des données (co-auteur). Le processus éditorial est donc également *explosé* mais nécessairement présent. Qui plus est, il a lieu en amont du processus auctorial. Le processus documentaire est donc complet et premier tandis que le document est second et *inabouti*.

Toutefois, il existe de nombreuses situations où l'envie de "faire document" (pas seulement proto-document) survient, pour capitaliser, mémoriser, partager, comprendre, justifier son action… Un temps est alors nécessaire pour "faire document", afin de revenir sur ce qui a été fait, revenir sur le processus documentaire tel qu'il est encore accessible : garder dans les favoris telle ou telle ressource, organiser les favoris, les annoter, les partager, "imprimer" tel ou tel fragment documentaire tel qu'il est présenté (copie d'écran) ou tel qu'il est codé pour être imprimé, enregistrer un fragment documentaire (le code et les données associées)… Les éléments enregistrés sont alors organisables (dans un document numérique, dans une collection, dans une ressource web nouvelle…). On tenteainsi d'achever le processus auctorial.

Il est par contre difficile d'exploiter les traces du processus documentaire (logs, historiques de pages, mémoires de saisie, etc.). Par exemple, sur un« Google Docs », on peut voir l'historique des modifications (pour revenir à un état précédent par exemple), mais exploiter les informations disponibles dans l'historique comme preuve éditoriale du document final n'est pas possible immédiatement, il est impossible de « figer » (par exemple télécharger dans un document numérique) une version AVEC son historique associé.

Si c'est pendant le processus documentaire que l'action se réalise, alors partager les connaissances permettant l'action consiste à partager d'une manière ou d'une autre le processus documentaire accompagné de "l'exemple" des fragments documentaires produits et dont les contenus ont permis l'action.

Le manifeste : faire document sur le web

Faire document sur le web consiste à construire un produit documentaire rassemblant les deux bonnes propriétés de preuve et de mémoire, à des fins de réutilisation et de partage.

Faire document sur le web ?

Pour faire document, l'utilisateur doit construire un objet nouveau achevant le processus auctorial à partir des fragments mobilisés et « agis »pendant le processus documentaire qui est instancié par l'ensemble de l'activité web ayant permis l'action. En pratique, l'utilisateur est l'animateur principal du processus documentaire en le construisant interactivement à partir de ses saisies, choix de liens, navigations, choix de présentations, choix de filtrage, des réponses obtenues, des flux disponibles, et bien entendu ses propres productions (annotations, enrichissements, …). A tout moment, il peut utiliser une production documentaire intermédiaire pour l'action et, en particulier, pour continuer le processus documentaire lui-même.

S'il décide de mémoriser et partager ce qu'il vient de produire, que peut-il faire actuellement ?

- Garder l'URI des ressources dans les "marque-pages", les organiser éventuellement et même les partager (Delicious), ce qui n'est pas sans posé problème car si la ressource est un flux, on perd l'état qui était pertinent.
- S'il est intéressé par le résultat "instantané" d'une production documentaire, il peut éventuellement "imprimer" sur un support permanent l'état d'une URI, mais sans résultat garanti si la ressource n'est pas prévue pour être "imprimable" (par exemple, un son ou une vidéo).
- Une copie d'écran (forme d'impression du rendu graphique)
- Télécharger les éléments qui l'intéressent (videos, animations, documents, …)
- Visualiser l'historique de navigation, le sauvegarder..
- Assembler tous ces éléments dans un document et tenter de synthétiser ce qui sera utile pour réutiliser et partager les informations... (ce qui peut être fait sous la forme d'une ressource du web).

Manifestement, ce n'est pas facile et il sera vraiment difficile de "rejouer" tout ou partie du processus documentaire pour permettre de reconstituer les conditions documentaires permettant l'action (pour un lecteur ayant les capacités d'interprétation associées).
Les fragments documentaires produits lors du processus documentaires sont, à la différence de documents classiques, instables voire éphémères. Ils peuvent être l'état d'un flux à un instant donné ou le résultat d'un calcul déclenché par l'utilisateur. Ils existent donc au moment du processus documentaire, mais rien ne garanti qu'ils seront à nouveau mobilisables. S'il veut pouvoir justifier son action (faire preuve), il lui faut donc produire un document issu du processus documentaire.

Nous pensons qu'il faut introduire ce "faire document" comme une activité identifiée, introduisant un temps documentaire explicite articulé souplement au processus

documentaire usuel et imprévisible. Il s'agirait alors de construire un produit documentaire d'un type nouveau, issu d'un processus documentaire considéré explicitement pour faciliter le partage et la réutilisation de tout ce qui a permis l'action. Nous recommandons de prendre comme source initiale de ce produit, la trace annotée (car il convient de repérer au sein de chaque fragment documentaire l'élément qui a permis l'action) du processus documentaire qu'il conviendra alors de considérer comme un objet potentiellement documentaire et pas uniquement comme (dans le meilleur des cas) un historique de ressources mobilisées lors d'une activité exploitant le web.
Dans un monde cadencé par des actions rapprochées, intégrons le "temps du recul" permettant de démultiplier l'efficacité du flux d'inter-actions documentaires par un temps de la réflexion, du partage et de la réutilisation.
Ce nouveau produit documentaire pourrait être composé :

- de l'histoire interactive reformulable et rejouable dans différents registres d'usage (un registre d'usage est une façon de parler de l'usage : vocabulaire, niveau d'abstraction, signes, etc.) en "complicité" avec l'environnement web mobilisé pour résoudre les "trous" qui peuvent apparaître au moment du rejouage. Documenter le processus de façon à le rendre apte à faciliter l'action interactive dans l'environnement web.
- des modèles de description de ce processus pour les rendre plastiques dans leur réutilisation dans des contextes différents
- des outils collaboratifs pour échanger sur ces expériences explicitées et les rendre disponibles en dehors du contexte de l'utilisateur initial -> assistance à la construction opportuniste de communautés.

Conclusion

Ce manifeste pourrait être illustré par un certain nombre de travaux tendant à donner à l'utilisateur (lecteur-scribe) une conscience plus claire des processus documentaires qu'il anime par ses interactions et même à rendre compte de cette activité tracée par un *fait documentaire*. Nous pensons en particulier aux travaux tendant à fournir des éléments de réflexivité au processus documentaire (trace modélisée des interactions), et donc à instancier le processus documentaire sous la forme d'un objet informatique présenté de manière proto-documentaire à l'utilisateur [Alain Mille, 2012], ainsi que le principe de la (re)-documentation d'une activité médiée par un environnement informatique [Yahiaoui et al., 2012]. L'idée de documenter l'activité en cours n'est pas nouvelle et [Vanevar Bush, 1945], avait déjà imaginé d'exploiter les traces de recherche documentaire des savants pour en permettre d'autres (voir Figure 1).

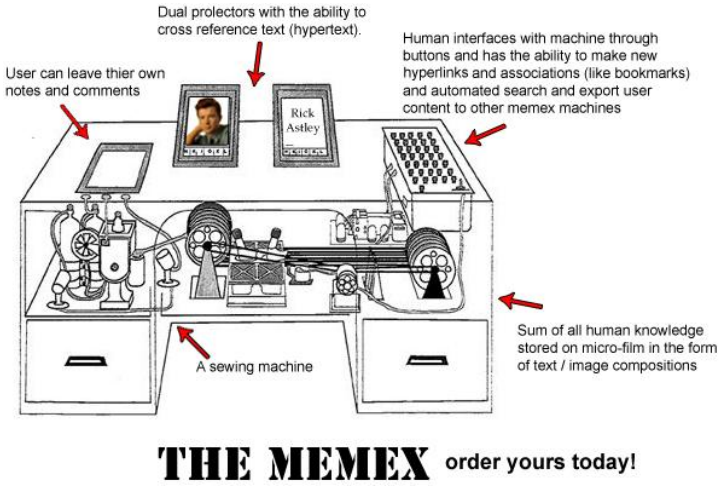


Figure 1 The MEMEX machine (Vaneshar Bush, 1945)

Pour que le processus documentaire aboutisse sur le web, il faut donc fournir les services nécessaires aux lecteurs-scribes, et leur permettre ainsi de devenir auteurs-éditeurs. Les conditions pour y parvenir sont de faciliter la prise de conscience du processus documentaire, l'accompagnement de sa finalisation par un aménagement des services du web conduisant à l'émergence d'un *temps documentaire* propre, un nouveau temps d'appropriation démultipliant encore les capacités d'agir par un partage et la construction de sens personnel et collectif.

Si le processus auctorial est relativement accessible à l'expérience individuelle, le processus éditorial est beaucoup moins évident à repérer dans les pratiques mais les initiatives comme celles de Wikipédia en montrent toute la puissance et l'importance sociale.

Cavallo, Guglielmo. 1995. "La lecture dans le monde romain." in *Histoire de la lecture dans le monde occidental*, edited by CavalloGuglielmo and Chartier, Roger. Paris : Seuil.

Hamesse, Jacqueline. 1995. "Le modèle scolastique de la lecture." in *Histoire de la lecture dans le monde occidental*, edited by CavalloGuglielmo and Chartier, Roger. Paris: Seuil.

Bush, Vanevar. 1945. « As wemaythink », in *Atlantic Monthly*, vol 1, issue 176, pages 101-108.

Mille, Alain. 2012.« Expérience tracée. Système de gestion de base de traces », in *Résaux sociaux*

Culture politique et ingénierie des réseaux sociaux, edited by Bernard Stiegler et al., FYP Editions, Collection du Nouveau Monde Industriel, ISBN 978-2-916571-35-5, janvier 2012

Yahiaoui, Leila, Prié, Yannick &Boufaïda, Zizette. "Du traçage de l'activité informatique à sa redocumentation en texte", in *Technique et science informatiques*, Lavoisier, à paraître.

Socioeconomic Dependencies of Linguistic Patterns in Twitter : A Multivariate Analysis

Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis*

Jacob Levy Abitbol
Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668, IXXI
Lyon, France
jacob.levy-abitbol@ens-lyon.fr

Márton Karsai
Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668, IXXI
Lyon, France
marton.karsai@ens-lyon.fr

Jean-Philippe Magué
ENS de Lyon, ICAR UMR 5191, CNRS
Lyon, France
jean-philippe.mague@ens-lyon.fr

Jean-Pierre Chevrot
Lidilem, University of Grenoble Alpes
Grenoble, France
jean-pierre.chevrot@u-grenoble3.fr

Eric Fleury
Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668, IXXI
Lyon, France
eric.fleury@ens-lyon.fr

ABSTRACT

Our usage of language is not solely reliant on cognition but is arguably determined by myriad external factors leading to a global variability of linguistic patterns. This issue, which lies at the core of sociolinguistics and is backed by many small-scale studies on face-to-face communication, is addressed here by constructing a dataset combining the largest French Twitter corpus to date with detailed socioeconomic maps obtained from national census in France. We show how key linguistic variables measured in individual Twitter streams depend on factors like socioeconomic status, location, time, and the social network of individuals. We found that *(i)* people of higher socioeconomic status, active to a greater degree during the daytime, use a more standard language; *(ii)* the southern part of the country is more prone to use more standard language than the northern one, while locally the used variety or dialect is determined by the spatial distribution of socioeconomic status; and *(iii)* individuals connected in the social network are closer linguistically than disconnected ones, even after the effects of status homophily have been removed. Our results inform sociolinguistic theory and may inspire novel learning methods for the inference of socioeconomic status of people from the way they tweet.

KEYWORDS

computational sociolinguistics, Twitter data, socioeconomic status inference, social network analysis, spatiotemporal data

ACM Reference Format:
Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186011

*supported by the SoSweet ANR project (ANR-15-CE38-0011-03).

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186011

1 INTRODUCTION

Communication is highly variable and this variability contributes to language change and fulfills social functions. Analyzing and modeling data from social media allows the high-resolution and long-term follow-up of large samples of speakers, whose social links and utterances are automatically collected. This empirical basis and long-standing collaboration between computer and social scientists could dramatically extend our understanding of the links between language variation, language change, and society.

Languages and communication systems of several animal species vary in time, geographical space, and along social dimensions. Varieties are shared by individuals frequenting the same space or belonging to the same group. The use of vocal variants is flexible. It changes with the context and the communication partner and functions as "social passwords" indicating which individual is a member of the local group [15]. Similar patterns can be found in human languages if one considers them as evolving and dynamical systems that are made of several social or regional varieties, overlapping or nested into each other. Their emergence and evolution result from their internal dynamics, contact with each other, and link formation within the social organization, which itself is evolving, composite and multi-layered [25, 32].

The strong tendency of communication systems to vary, diversify and evolve seems to contradict their basic function: allowing mutual intelligibility within large communities over time. Language variation is not counter adaptive. Rather, subtle differences in the way others speak provide critical cues helping children and adults to organize the social world [24]. Linguistic variability contributes to the construction of social identity, definition of boundaries between social groups and the production of social norms and hierarchies.

Sociolinguistics has traditionally carried out research on the quantitative analysis of the so-called linguistic variables, i.e. points of the linguistic system which enable speakers to say the same thing in different ways, with these variants being *"identical in reference or truth value, but opposed in their social [...] significance"* [31]. Such variables have been described in many languages: variable pronunciation of -ing as [in] instead of [iŋ] in English (*playing* pronounced *playin'*); optional realization of the first part of the

French negation (*je (ne) fume pas*, "I do not smoke"); optional realization of the plural ending of verb in Brazilian Portuguese (eles disse(ram), "they said"). For decades, sociolinguistic studies have showed that hearing certain variants triggers social stereotypes [4]. The so-called standard variants (*e.g.* [iŋ], realization of negative ne and plural -ram) are associated with social prestige, high education, professional ambition and effectiveness. They are more often produced in more formal situation. Non-standard variants are linked to social skills, solidarity and loyalty towards the local group, and they are produced more frequently in less formal situation.

It is therefore reasonable to say that the sociolinguistic task can benefit from the rapid development of computational social science [34]: the similarity of the online communication and face-to-face interaction [16] ensures the validity of the comparison with previous works. In this context, the nascent field of computational sociolinguistics found the digital counterparts of the sociolinguistic patterns already observed in spoken interaction. However a closer collaboration between computer scientists and sociolinguists is needed to meet the challenges facing the field [40]:

- Going beyond lexical variation (standard or non-standard usage of words) and English language
- Extending the focus to factors unexplored in digital communication such as social class
- Using the social sciences as a source of methodological inspiration for controlling for multiple factors instead of focusing on one factor as in the field of computational sociolinguistics
- Emphasizing the interpretability of the models and the insights for sociolinguistic theory.

The present work meets most of these challenges. It constructs the largest dataset of French tweets enriched with census sociodemographic information existent to date to the best of our knowledge. From this dataset, we observed variation of two grammatical cues and an index of vocabulary size in users located in France. We study how the linguistic cues correlated with three features reflective of the socioeconomic status of the users, their most representative location and their daily periods of activity on Twitter. We also observed whether connected people are more linguistically alike than disconnected ones. Multivariate analysis shows strong correlations between linguistic cues and socioeconomic status as well as a broad spatial pattern never observed before, with more standard language variants and lexical diversity in the southern part of the country. Moreover, we found an unexpected daily cyclic evolution of the frequency of standard variants. Further analysis revealed that the observed cycle arose from the ever changing average economic status of the population of users present in Twitter through the day. Finally, we were able to establish that linguistic similarity between connected people does arises partially but not uniquely due to status homophily (users with similar socioeconomic status are linguistically similar and tend to connect). Its emergence is also due to other effects potentially including other types of homophilic correlations or influence disseminated over links of the social network. Beyond we verify the presence of status homophily in the Twitter social network our results may inform novel methods to infer socioeconomic status of people from the way they use language. Furthermore, our work, rooted within the web content analysis line of research [19], extends the usual focus on aggregated textual features (like document frequency metrics or embedding methods) to specific linguistic markers, thus enabling sociolinguistics knowledge to inform the data collection process.

2 RELATED WORK

For decades, sociolinguistic studies have repeatedly shown that speakers vary the way they talk depending on several factors. These studies have usually been limited to the analysis of small scale datasets, often obtained by surveying a set of individuals, or by direct observation after placing them in a controlled experimental setting. In spite of the volume of data collected generally, these studies have consistently shown the link between linguistic variation and social factors [5, 30].

Recently, the advent of social media and publicly available communication platforms has opened up a new gate to access individual information at a massive scale. Among all available social platforms, Twitter has been regarded as the choice by default, namely thanks to the intrinsic nature of communications taking place through it and the existence of data providers that are able to supply researchers with the volume of data they require. Work previously done on demographic variation is now relying increasingly on corpora from this social media platform as evidenced by the myriad of results showing that this resource reflects not only morpholexical variation of spoken language but also geographical [9, 41].

Although the value of this kind of platform for linguistic analysis has been more than proven, the question remains on how previous sociolinguistic results scale up to the sheer amount of data within reach and how can the latter enrich the former. To do so, numerous studies have focused on enhancing the data emanating from Twitter itself. Indeed, one of the core limitations of Twitter is the lack of reliable sociodemographic information about the sampled users as usually data fields such as user-entered profile locations, gender or age differ from reality. This in turn implies that user-generated profile content cannot be used as a useful proxy for the sociodemographic information [11].

Many studies have overcome this limitation by taking advantage of the geolocation feature allowing Twitter users to include in their posts the location from which they were tweeted. Based on this metadata, studies have been able to assign home location to geolocated users with varying degrees of accuracy [1]. Subsequent work has also been devoted to assigning to each user some indicator that might characterize their socioeconomic status based on their estimated home location. These indicators are generally extracted from other datasets used to complete the Twitter one, namely census data [8, 9, 36] or real estate online services as Zillow.com [43]. Other approaches have also relied on sources of socioeconomic information such as the UK Standard Occupation Classification (SOC) hierarchy, to assign socioeconomic status to users with occupation mentions [42]. Despite the relative success of these methods, their common limitation is to provide observations and predictions based on a carefully hand-picked small set of users, letting alone the problem of socioeconomic status inference on larger and more heterogeneous populations. Our work stands out from this well-established line of research by expanding the definition of socioeconomic status to include several demographic features as well as by pinpointing potential home location to individual users

with an unprecedented accuracy. Identifying socioeconomic status and the network effects of homophily[44] is an open question [10]. However, recent results already showed that status homophily, i.e. the tendency of people of similar socioeconomic status are better connected among themselves, induce structural correlations which are pivotal to understand the stratified structure of society [35]. While we verify the presence of status homophily in the Twitter social network, we detect further sociolinguistic correlations between language, location, socioeconomic status, and time, which may inform novel methods to infer socioeconomic status for a broader set of people using common information available on Twitter.

3 DATA DESCRIPTION

One of the main achievements of our study was the construction of a combined dataset for the analysis of sociolinguistic variables as a function of socioeconomic status, geographic location, time, and the social network. As follows, we introduce the two aforementioned independent datasets and how they were combined. We also present a brief cross-correlation analysis to ground the validity of our combined dataset for the rest of the study. In what follows, it should also be noted that regression analysis was performed via linear regression as implemented in the Scikit Learn Toolkit while data preprocessing and network study were performed using respectively pandas [37] and NetworkX [12] Python libraries.

3.1 Twitter dataset: sociolinguistic features

Our first dataset consists of a large data corpus collected from the online news and social networking service, Twitter. On it, users can post and interact with messages, "tweets", restricted to 140 characters. Tweets may come with several types of metadata including information about the author's profile, the detected language, where and when the tweet was posted, etc. Specifically, we recorded 170 million tweets written in French, posted by 2.5 million users in the timezones GMT and GMT+1 over three years (between July 2014 to May 2017). These tweets were obtained via the Twitter powertrack API feeds provided by Datasift and Gnip with an access rate varying between $15 - 25\%$[1].

Linguistic data: To obtain meaningful linguistic data we preprocessed the incoming tweet stream in several ways. As our central question here deals with the variability of the language, repeated tweets do not bring any additional information to our study. Therefore, as an initial filtering step, we decided to remove retweets. Next, in order to facilitate the detection of the selected linguistic markers we removed any URLs, emoticons, mentions of other users (denoted by the `@` symbol) and hashtags (denoted by the # symbol) from each tweet. These expressions were not considered to be semantically meaningful and their filtering allowed to further increase the speed and accuracy of our linguistic detection methods when run across the data. In addition we completed a last step of textual preprocessing by down-casing and stripping the punctuation out of the tweets body. POS-taggers such as MElt [7] were also tested but they provided no significant improvement in the detection of the linguistic markers.

Network data: We used the collected tweets in another way to infer social relationships between users. Tweet messages may be direct interactions between users, who mention each other in the text by using the `@` symbol (`@username`). When one user u, mentions another user v, user v will see the tweet posted by user u directly in his / her feed and may tweet back. In our work we took direct mentions as proxies of social interactions and used them to identify social ties between pairs of users. Opposite to the follower network, reflecting passive information exposure and less social involvement, the mutual mention network has been shown [20] to capture better the underlying social structure between users. We thus use this network definition in our work as links are a greater proxy for social interactions.

In our definition we assumed a tie between users if they mutually mentioned each other at least once during the observation period. People who reciprocally mentioned each other express some mutual interest, which may be a stronger reflection of real social relationships as compared to the non-mutual cases [18]. This constraint reduced the egocentric social network considerably leading to a directed structure of $508,975$ users and $4,029,862$ links that we considered being undirected in what follows.

Geolocated data: About 2% of tweets included in our dataset contained some location information regarding either the tweet author's self-provided position or the place from which the tweet was posted. These pieces of information appeared as the combination of self reported locations or usual places tagged with GPS coordinates at different geographic resolution. We considered only tweets which contained the exact GPS coordinates with resolution of ~ 3 meters of the location where the actual tweet was posted. This actually means that we excluded tweets where the user assigned a place name such as "Paris" or "France" to the location field, which are by default associated to the geographical center of the tagged areas. Practically, we discarded coordinates that appeared more than 500 times throughout the whole GPS-tagged data, assuming that there is no such 3×3 meter rectangle in the country where 500 users could appear and tweet by chance. After this selection procedure we rounded up each tweet location to a 100 meter precision.

To obtain a unique representative location of each user, we extracted the sequence of all declared locations from their geolocated tweets. Using this set of locations we selected the most frequent to be the representative one, and we took it as a proxy for the user's home location. Further we limited our users to ones located throughout the French territory thus not considering others tweeting from places outside the country. This selection method provided us with $110,369$ geolocated users who are either detected as French speakers or assigned to be such by Twitter and all associated to specific 'home' GPS coordinates in France. To verify the spatial distribution of the selected population, we further assessed the correlations between the true population distributions (obtained from census data [22]) at different administrative level and the geolocated user distribution aggregated correspondingly. More precisely, we computed the R^2 coefficient of variation between the inferred and official population distributions (a) at the level of 22 regions[2].

[1] In order to uphold the strict privacy laws in France as well as the agreement signed with our data provider GNIP, full disclosure of the original dataset is not possible. Data collection and preprocessing pipelines could however be released upon request.

[2] Note that since 2016 France law determines 13 metropolitan regions, however the available data shared by INSEE [22] contained information about the earlier administrative structure containing 22 regions.

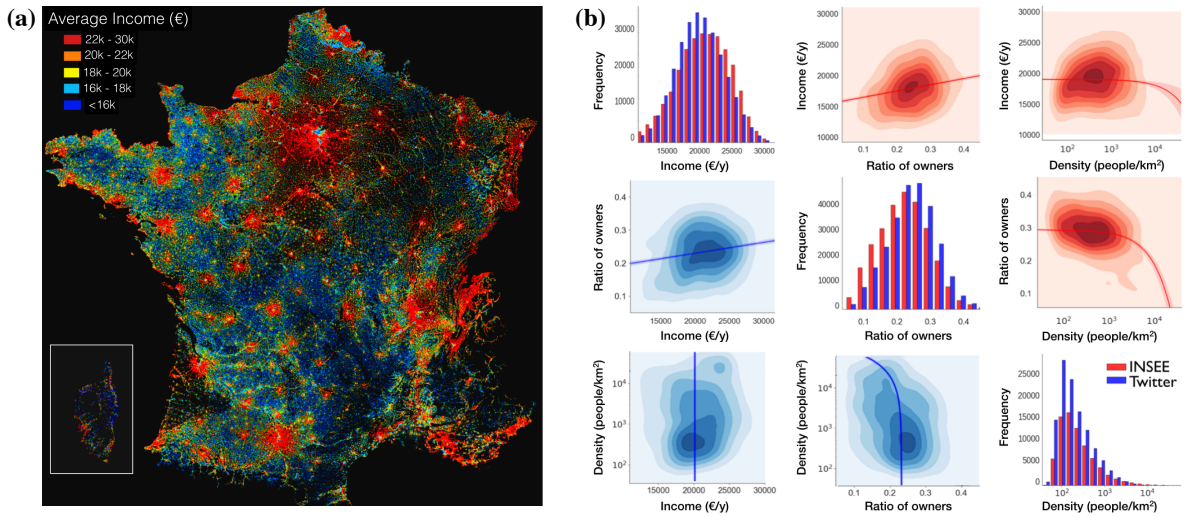


Figure 1: Distributions and correlations of socioeconomic indicators. (a) Spatial distribution of average income in France with $200m \times 200m$ resolution. (b) Distribution of socioeconomic indicators (in the diag.) and their pairwise correlations measured in the INSEE (upper diag. panels) and Twitter geotagged (lower diag. panels) datasets. Contour plots assign the equidensity lines of the scatter plots, while solid lines are the corresponding linear regression values. Population density in log.

Correlations at this level induced a high coefficient of $R^2 \simeq 0.89$ ($p < 10^{-2}$); (b) At the arrondissement level with 322 administrative units and coefficient $R^2 \simeq 0.87$ ($p < 10^{-2}$); and (c) at the canton level with 4055 units with a coefficient $R \simeq 0.16$ ($p < 10^{-2}$). Note that the relatively small coefficient at this level is due to the interplay of the sparsity of the inferred data and the fine grained spatial resolution of cantons. All in all, we can conclude that our sample is highly representative in terms of spatial population distribution, which at the same time validate our selection method despite the potential inherent biases induced by the method taking the most frequented GPS coordinates as the user's home location.

3.2 INSEE dataset: socioeconomic features

The second dataset we used was released in December 2016 by the National Institute of Statistics and Economic Studies (INSEE) of France. This data corpus [23] contains a set of sociodemographic aggregated indicators, estimated from the 2010 tax return in France, for each 4 hectare ($200m \times 200m$) square patch across the whole French territory. Using these indicators, one can estimate the distribution of the average socioeconomic status (SES) of people with high spatial resolution. In this study, we concentrated on three indicators for each patch i , which we took to be good proxies of the socioeconomic status of the people living within them. These were the S^i_{inc} average yearly income per capita (in euros), the S^i_{own} fraction of owners (not renters) of real estate, and the S^i_{den} density of population defined respectively as

$$: S^i_{\text{inc}} = \frac{S^i_{hh}}{N^i_{hh}}, \quad S^i_{\text{own}} = \frac{N^i_{\text{own}}}{N^i}, \quad \text{and} \quad S^i_{\text{den}} = \frac{N^i}{(200m)^2}. \quad (1)$$

Here S^i_{hh} and N^i_{hh} assign respectively the cumulative income and total number of inhabitants of patch i, while N^i_{own} and N^i are respectively the number of real estate owners and the number of individuals living in patch i. As an illustration we show the spatial distribution of S^i_{inc} average income over the country in Fig.1a.

In order to uphold current privacy laws and due to the highly sensitive nature of the disclosed data, some statistical pretreatments were applied to the data by INSEE before its public release. More precisely, neighboring patches with less than 11 households were merged together, while some of the sociodemographic indicators were winsorized. This set of treatments induced an inherent bias responsible for the deviation of the distribution of some of the socioeconomic indicators. These quantities were expected to be determined by the Pareto principle, thus reflecting the high level of socioeconomic imbalances present within the population. Instead, as shown in Fig.1b [diagonal panels], distributions of the derived socioeconomic indicators (in blue) appeared somewhat more symmetric than expected. This doesn't hold though for $P(S^i_{\text{den}})$ (shown on a log-log scale in the lowest right panel of Fig.1b), which emerged with a broad tail similar to an expected power-law Pareto distribution. In addition, although the patches are relatively small ($200m \times 200m$), the socioeconomic status of people living may have some local variance, what we cannot consider here. Nevertheless, all things considered, this dataset and the derived socioeconomic indicators yield the most fine-grained description, allowed by national law, about the population of France over its whole territory.

Despite the inherent biases of the selected socioeconomic indicators, in general we found weak but significant pairwise correlations between these three variables as shown in the upper diagonal panels in Fig.1b (in red), with values in Table 1. We observed that while S^i_{inc} income and S^i_{own} owner ratio are positively correlated ($R = 0.24$,

$p < 10^{-2}$), and the S^i_{own} and S^i_{den} population density are negatively correlated ($R = -0.23$, $p < 10^{-2}$), S^i_{inc} and S^i_{den} appeared to be very weakly correlated ($R = -0.07$, $p < 10^{-2}$). This nevertheless suggested that high average income, high owner ratio, and low population density are consistently indicative of high socioeconomic status in the dataset.

Table 1: Pearson correlations and p-values measured between SES indicators in the INSEE and Twitter datasets.

	$S^i_{\text{inc}} \sim S^i_{\text{own}}$	$S^i_{\text{inc}} \sim S^i_{\text{den}}$	$S^i_{\text{own}} \sim S^i_{\text{den}}$
INSEE	0.24 ($p < 10^{-2}$)	-0.07 ($p < 10^{-2}$)	-0.23 ($p < 10^{-2}$)
Twitter	0.19 ($p < 10^{-2}$)	0.00 ($p > 10^{-2}$)	-0.22 ($p < 10^{-2}$)

3.3 Combined dataset: individual socioeconomic features

Data collected from Twitter provides a large variety of information about several users including their tweets, which disclose their interests, vocabulary, and linguistic patterns; their direct mentions from which their social interactions can be inferred; and the sequence of their locations, which can be used to infer their representative location. However, no information is directly available regarding their socioeconomic status, which can be pivotal to understand the dynamics and structure of their personal linguistic patterns.

To overcome this limitation we combined our Twitter data with the socioeconomic maps of INSEE by assigning each geolocated Twitter user to a patch closest to their estimated home location (within 1 km). This way we obtained for all 110, 369 geolocated users their dynamical linguistic data, their egocentric social network as well as a set of SES indicators.

Such a dataset associating language with socioeconomic status and social network throughout the French metropolitan territory is unique to our knowledge and provides unrivaled opportunities to verify sociolinguistic patterns observed over a long period on a small-scale, but never established in such a large population.

To verify whether the geolocated Twitter users yet provide a representative sample of the whole population we compared the distribution and correlations of the their SES indicators to the population measures. Results are shown in Fig.1b diagonal (red distributions) and lower diagonal panels (in blue) with correlation coefficients and p-values summarized in Table.1. Even if we observed some discrepancy between the corresponding distributions and somewhat weaker correlations between the SES indicators, we found the same significant correlation trends (with the exception of the pair density / income) as the ones seen when studying the whole population, assuring us that each indicator correctly reflected the SES of individuals.

4 LINGUISTIC VARIABLES

We identified the following three linguistic markers to study across users from different socioeconomic backgrounds: Correlation with SES has been evidenced for all of them. The optional deletion of negation is typical of spoken French, whereas the omission of the mute letters marking the plural in the nominal phrase is a variable cue of French writing. The third linguistic variable is a global measure of the lexical diversity of the Twitter users. We present them here in greater detail.

4.1 Standard usage of negation

The basic form of negation in French includes two negative particles: *ne* (no) before the verb and another particle after the verb that conveys more accurate meaning: *pas* (not), *jamais* (never), *personne* (no one), *rien* (nothing), etc. Due to this double construction, the first part of the negation (*ne*) is optional in spoken French, but it is obligatory in standard writing. Sociolinguistic studies have previously observed the realization of *ne* in corpora of recorded everyday spoken interactions. Although all the studies do not converge, a general trend is that *ne* realization is more frequent in speakers with higher socioeconomic status than in speakers with lower status [2, 14]. We built upon this research to set out to detect both negation variants in the tweets using regular expressions.[3] We are namely interested in the rate of usage of the standard negation (featuring both negative particles) across users:

$$L^u_{\text{cn}} = \frac{n^u_{\text{cn}}}{n^u_{\text{cn}} + n^u_{\text{incn}}} \quad \text{and} \quad \overline{L}^i_{\text{cn}} = \frac{\sum_{u \in i} L^u_{\text{cn}}}{N_i}, \tag{2}$$

where n^u_{cn} and n^u_{incn} assign the number of correct negation and incorrect number of negation of user u, thus L^u_{cn} defines the rate of correct negation of a users and $\overline{L}^i_{\text{cn}}$ its average over a selected i group (like people living in a given place) of N_i users.

4.2 Standard usage of plural ending of written words

In written French, adjectives and nouns are marked as being plural by generally adding the letters *s* or *x* at the end of the word. Because these endings are mute (without counterpart in spoken French), their omission is the most frequent spelling error in adults [6]. Moreover, studies showed correlations between standard spelling and social status of the writers, in preteens, teens and adults [3, 6, 45]. We then set to estimate the use of standard plural across users:

$$L^u_{\text{cp}} = \frac{n^u_{\text{cp}}}{n^u_{\text{cp}} + n^u_{\text{incp}}} \quad \text{and} \quad \overline{L}^i_{\text{cp}} = \frac{\sum_{u \in i} L^u_{\text{cp}}}{N_i} \tag{3}$$

where the notation follows as before (cp stands for correct plural and incp stands for incorrect plural).

4.3 Normalized vocabulary set size

A positive relationship between an adult's lexical diversity level and his or her socioeconomic status has been evidenced in the field of language acquisition. Specifically, converging results showed that the growth of child lexicon depends on the lexical diversity in the speech of the caretakers, which in turn is related to their socioeconomic status and their educational level [17, 21]. We thus proceeded to study the following metric:

$$L^u_{\text{vs}} = \frac{N^u_{\text{vs}}}{N^u_{tw}} \quad \text{and} \quad \overline{L}^i_{\text{vs}} = \frac{\sum_{u \in i} N^u_{\text{vs}}}{N_i}, \tag{4}$$

[3] Negation:\\b(pas|pa|aps|jamais|ni|personne|rien|ri1|r1|aucun|aucune)\\b
Standard Negation:. * \\b(ne|n')\\b. * \$

where $N_v s^u$ assigns the total number of unique words used by user u who tweeted N^u_{tw} times during the observation period. As such L^u_{vs} gives the normalized vocabulary set size of a user u, while $\overline{L}^i_{\text{vs}}$ defines its average for a population i.

5 RESULTS

By measuring the defined linguistic variables in the Twitter timeline of users we were finally set to address the core questions of our study, which dealt with linguistic variation. More precisely, we asked whether the language variants used online depend on the socioeconomic status of the users, on the location or time of usage, and on ones social network. To answer these questions we present here a multidimensional correlation study on a large set of Twitter geolocated users, to which we assigned a representative location, three SES indicators, and a set of meaningful social ties based on the collection of their tweets.

5.1 Socioeconomic variation

The socioeconomic status of a person is arguably correlated with education level, income, habitual location, or even with ethnicity and political orientation and may strongly determine to some extent patterns of individual language usage. Such dependencies have been theoretically proposed before [30], but have rarely been inspected at this scale yet. The use of our previously described datasets enabled us to do so via the measuring of correlations between the inferred SES indicators of Twitter users and the use of the previously described linguistic markers.

To compute and visualize these correlations we defined linear bins (in numbers varying from 20 to 50) for the socioeconomic indicators and computed the average of the given linguistic variables for people falling within the given bin. These binned values (shown as symbols in Fig.2) were used to compute linear regression curves and the corresponding confidence intervals (see Fig.2). An additional transformation was applied to the SES indicator describing population density, which was broadly distributed (as discussed in Section 3.2 and Fig.1b), thus, for the regression process, the logarithm of its values were considered. To quantify pairwise correlations we computed the R^2 coefficient of determination values in each case.

Table 2: The R^2 coefficient of determination and the corresponding p-values computed for the pairwise correlations of SES indicators and linguistic variables.

	S^i_{inc}	S^i_{own}	S^i_{den}
$\overline{L}_{\text{cn}}$	0.19 ($p < 10^{-2}$)	0.59 ($p < 10^{-2}$)	0.74 ($p < 10^{-2}$)
$\overline{L}_{\text{cp}}$	0.59 ($p < 10^{-2}$)	0.66 ($p < 10^{-2}$)	0.76 ($p < 10^{-2}$)
$\overline{L}_{\text{vs}}$	0.70 ($p < 10^{-2}$)	0.32 ($p < 10^{-2}$)	0.41 ($p < 10^{-2}$)

In Fig.2 we show the correlation plots of all nine pairs of SES indicators and linguistic variables together with the linear regression curves, the corresponding R^2 values and the 95 percentile confidence intervals (note that all values are also in Table 2). These results show that correlations between socioeconomic indicators and linguistic variables actually exist. Furthermore, these correlation trends suggest that people with lower SES may use more

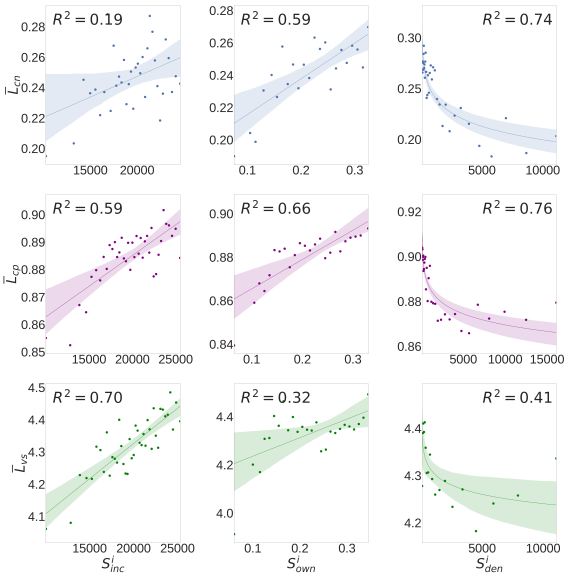


Figure 2: Pairwise correlations between three SES indicators and three linguistic markers. Columns correspond to SES indicators (resp. S^i_{inc}, S^i_{own}, S^i_{den}), while rows correspond to linguistic variables (resp. $\overline{L}_{\text{cn}}$, $\overline{L}_{\text{cp}}$ and $\overline{L}_{\text{vs}}$). On each plot colored symbols are binned data values and a linear regression curve are shown together with the 95 percentile confidence interval and R^2 values.

non-standard expressions (higher rates of incorrect negation and plural forms) have a smaller vocabulary set size than people with higher SES. Note that, although the observed variation of linguistic variables were limited, all the correlations were statistically significant ($p < 10^{-2}$) with considerably high R^2 values ranging from 0.19 (between $\overline{L}_{\text{cn}} \sim S_{\text{inc}}$) to 0.76 (between $\overline{L}_{\text{cp}} \sim S_{\text{den}}$). For the rates of standard negation and plural terms the population density appeared to be the most determinant indicator with $R^2 = 0.74$ (and 0.76 respectively), while for the vocabulary set size the average income provided the highest correlation (with $R^2 = 0.7$).
One must also acknowledge that while these correlations exhibit high values consistently across linguistic and socioeconomic indicators, they only hold meaning at the population level at which the binning was performed. When the data is considered at the user level, the variability of individual language usage hinders the observation of the aforementioned correlation values (as demonstrated by the raw scatter plots (grey symbols) in Fig. 2).

5.2 Spatial variation

Next we chose to focus on the spatial variation of linguistic variables. Although officially a standard language is used over the whole country, geographic variations of the former may exist due to several reasons [27, 46]. For instance, regional variability resulting from remnants of local languages that have disappeared, uneven spatial distribution of socioeconomic potentials, or influence spreading from neighboring countries might play a part in this process. For

the observation of such variability, by using their representative locations, we assigned each user to a department of France. We then computed the $\overline{L}^i_{\text{cn}}$ (resp. $\overline{L}^i_{\text{cp}}$) average rates of standard negation (resp. plural agreement) and the $\overline{L}^i_{\text{vs}}$ average vocabulary set size for each "département" i in the country (administrative division of France – There are 97 départements).

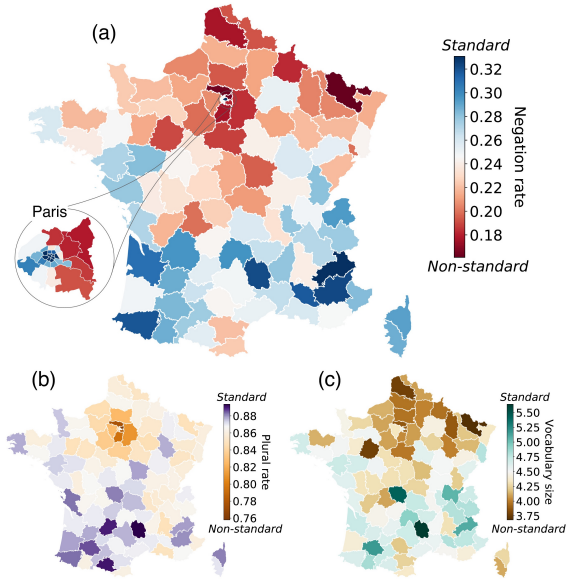


Figure 3: Geographical variability of linguistic markers in France. (a) Variability of the rate of correct negation. Inset focuses on larger Paris. (b) Variability of the rate of correct plural terms. (c) Variability of the average vocabulary size set. Each plot depicts variability on the department level except the inset of (a) which is on the "arrondissements" level.

Results shown in Fig.3a-c revealed some surprising patterns, which appeared to be consistent for each linguistic variable. By considering latitudinal variability it appeared that, overall, people living in the northern part of the country used a less standard language, i.e., negated and pluralized less standardly, and used a smaller number of words. On the other hand, people from the South used a language which is somewhat closer to the standard (in terms of the aforementioned linguistic markers) and a more diverse vocabulary. The most notable exception is Paris, where in the city center people used more standard language, while the contrary is true for the suburbs. This observation, better shown in Fig.3a inset, can be explained by the large differences in average socioeconomic status between districts. Such segregation is known to divide the Eastern and Western sides of suburban Paris, and in turn to induce apparent geographic patterns of standard language usage. We found less evident longitudinal dependencies of the observed variables. Although each variable shows a somewhat diagonal trend, the most evident longitudinal dependency appeared for the average rate of standard pluralization (see Fig.3b), where users from the Eastern side of the country used the language in less standard ways. Note that we

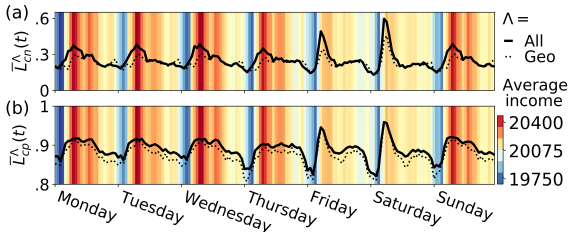


Figure 4: Temporal variability of (a) $\overline{L}^{\Lambda}_{\text{cn}}(t)$ (resp. (b) $\overline{L}^{\Lambda}_{\text{cp}}(t)$) average rate of correct negation (resp. plural terms) over a week with one hour resolution. Rates were computed for $\Lambda = all$ (solid line) and $\Lambda = geo$located Twitter users. Colors indicates the temporal variability of the average income of geolocated population active in a given hour.

also performed a multivariate regression analysis (not shown here), using the linguistic markers as target and considering as factors both location (in terms of latitude and longitude) as and income as proxy of socioeconomic status. It showed that while location is a strong global determinant of language variability, socioeconomic variability may still be significant locally to determine standard language usage (just as we demonstrated in the case of Paris).

5.3 Temporal variation

Another potentially important factor determining language variability is the time of day when users are active in Twitter [13, 26]. The temporal variability of standard language usage can be measured for a dynamical quantity like the $L_{\text{cn}}(t)$ rate of correct negation. To observe its periodic variability (with a ΔT period of one week) over an observation period of T (in our case 734 days), we computed

$$\overline{L}^{\Lambda}_{\text{cn}}(t) = \frac{\Delta T}{|\Lambda| T} \sum_{u \in \Lambda} \sum_{k=0}^{\lfloor T/\Delta T \rfloor} L^u_{\text{cn}}(t + k\Delta T), \tag{5}$$

in a population Λ of size $|\Lambda|$ with a time resolution of one hour. This quantity reflects the average standard negation rate in an hour over the week in the population Λ. Note that an equivalent $\overline{L}^{\Lambda}_{\text{cp}}(t)$ measure can be defined for the rate of standard plural terms, but not for the vocabulary set size as it is a static variable.

In Fig. 4a and b we show the temporal variability of $\overline{L}^{\Lambda}_{\text{cn}}(t)$ and $\overline{L}^{\Lambda}_{\text{cp}}(t)$ (respectively) computed for the whole Twitter user set ($\Gamma = all$, solid line) and for geolocated users ($\Gamma = geo$, dashed lines). Not surprisingly, these two curves were strongly correlated as indicated by the high Pearson correlation coefficients summarized in the last column of Table 3 which, again, assured us that our geolocated sample of Twitter users was representative of the whole set of users. At the same time, the temporal variability of these curves suggested that people tweeting during the day used a more standard language than those users who are more active during the night. However, after measuring the average income of active users in a given hour over a week, we obtained an even more sophisticated picture. It turned out that people active during the day have higher average income (warmer colors in Fig. 4) than people active during the night (colder colors in Fig. 4). Thus the variability of standard language patterns was largely explained by the changing overall

composition of active Twitter users during different times of day and the positive correlation between socioeconomic status and the usage of higher linguistic standards (that we have seen earlier). This explanation was supported by the high coefficients (summarized in Table 3), which were indicative of strong and significant correlations between the temporal variability of average linguistic variables and average income of the active population on Twitter.

Table 3: Pearson correlations and p-values of pairwise correlations of time varying $S_{\text{inc}}(t)$ average income with $\overline{L}^{\Lambda}_{\text{cn}}(t)$ and $\overline{L}^{\Lambda}_{\text{cp}}(t)$ average linguistic variables; and between average linguistic variables of $\Lambda = all$ and Λ = geo-localized users.

	$\overline{L}^{all}_{*}(t) \sim S_{\text{inc}}(t)$	$\overline{L}^{geo}_{*}(t) \sim S_{\text{inc}}(t)$	$\overline{L}^{geo}_{*}(t) \sim \overline{L}^{all}_{*}(t)$
$*$ = cn	0.5915 ($p < 10^{-2}$)	0.622 ($p < 10^{-2}$)	0.805 ($p < 10^{-2}$)
$*$ = cp	0.7027 ($p < 10^{-2}$)	0.665 ($p < 10^{-2}$)	0.98021 ($p < 10^{-2}$)

5.4 Network variation

Finally we sought to understand the effect of the social network on the variability of linguistic patterns. People in a social structure can be connected due to several reasons. Link creation mechanisms like focal or cyclic closure [28, 33], or preferential attachment [29] together with the effects of homophily [38] are all potentially driving the creation of social ties and communities, and the emergence of community rich complex structure within social networks. In terms of homophily, one can identify several individual characteristics like age, gender, common interest or political opinion, etc., that might increase the likelihood of creating relationships between disconnected but similar people, who in turn influence each other and become even more similar. Status homophily between people of similar socioeconomic status has been shown to be important [35] in determining the creation of social ties and to explain the stratified structure of society. By using our combined datasets, we aim here to identify the effects of status homophily and to distinguish them from other homophilic correlations and the effects of social influence inducing similarities among already connected people.

To do so, first we took the geolocated Twitter users in France and partitioned them into nine socioeconomic classes using their inferred income S^u_{inc}. Partitioning was done first by sorting users by their S^u_{inc} income to calculate their $C(S^u_{\text{inc}})$ cumulative income distribution function. We defined socioeconomic classes by segmenting $C(S^u_{\text{inc}})$ such that the sum of income is the same for each classes (for an illustration of our method see Fig.6a in the Appendix). We constructed a social network by considering mutual mention links between these users (as introduced in Section 3). Taking the assigned socioeconomic classes of connected individuals, we confirmed the effects of status homophily in the Twitter mention network by computing the connection matrix of socioeconomic groups normalized by the equivalent matrix of corresponding configuration model networks, which conserved all network properties except structural correlations (as explained in the Appendix). The diagonal component in Fig.6 matrix indicated that users of similar socioeconomic classes were better connected, while people from classes far apart were less connected than one would expect by chance from the reference model with users connected randomly.

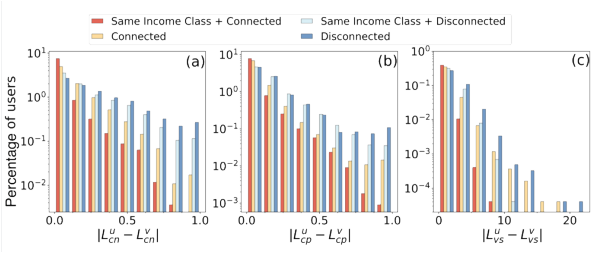


Figure 5: Distribution of the $|L^u_* - L^v_*|$ absolute difference of linguistic variables $* \in \{\text{cn}, \text{cp}, vs\}$ (resp. panels (a), (b), and (c)) of user pairs who were connected and from the same socioeconomic group (red), connected (yellow), disconnected and from the same socioeconomic group (light blue), disconnected pairs of randomly selected users (blue).

In order to measure linguistic similarities between a pair of users u and v, we simply computed the $|L^u_* - L^v_*|$ absolute difference of their corresponding individual linguistic variable $* \in \{\text{cn}, \text{cp}, vs\}$. This measure appeared with a minimum of 0 and associated smaller values to more similar pairs of users. To identify the effects of status homophily and the social network, we proceeded by computing the similarity distribution in four cases: for connected users from the same socioeconomic class; for disconnected randomly selected pairs of users from the same socioeconomic class; for connected users in the network; and randomly selected pairs of disconnected users in the network. Note that in each case the same number of user pairs were sampled from the network to obtain comparable averages. This number was naturally limited by the number of connected users in the smallest socioeconomic class, and were chosen to be 10, 000 in each cases. By comparing the distributions shown in Fig.5 we concluded that (a) connected users (red and yellow bars) were the most similar in terms of any linguistic marker. This similarity was even greater when the considered tie was connecting people from the same socioeconomic group; (b) network effects can be quantified by comparing the most similar connected (red bar) and disconnected (light blue bar) users from the same socioeconomic group. Since the similarity between disconnected users here is purely induced by status homophily, the difference of these two bars indicates additional effects that cannot be explained solely by status homophily. These additional similarities may rather be induced by other factors such as social influence, the physical proximity of users within a geographical area or other homophilic effects that were not accounted for. (c) Randomly selected pairs of users were more dissimilar than connected ones as they dominated the distributions for larger absolute difference values. We therefore concluded that both the effects of network and status homophily mattered in terms of linguistic similarity between users of this social media platform.

6 CONCLUSIONS

The overall goal of our study was to explore the dependencies of linguistic variables on the socioeconomic status, location, time varying activity, and social network of users. To do so we constructed a combined dataset from a large Twitter data corpus, including geotagged posts and proxy social interactions of millions of users,

as well as a detailed socioeconomic map describing average socioeconomic indicators with a high spatial resolution in France. The combination of these datasets provided us with a large set of Twitter users all assigned to their Twitter timeline over three years, their location, three individual socioeconomic indicators, and a set of meaningful social ties. Three linguistic variables extracted from individual Twitter timelines were then studied as a function of the former, namely, the rate of standard negation, the rate of plural agreement and the size of vocabulary set.

Via a detailed multidimensional correlation study we concluded that (a) socioeconomic indicators and linguistic variables are significantly correlated. i.e. people with higher socioeconomic status are more prone to use more standard variants of language and a larger vocabulary set, while people on the other end of the socioeconomic spectrum tend to use more non-standard terms and, on average, a smaller vocabulary set; (b) Spatial position was also found to be a key feature of standard language use as, overall, people from the North tended to use more non-standard terms and a smaller vocabulary set compared to people from the South; a more fine-grained analysis reveals that the spatial variability of language is determined to a greater extent locally by the socioeconomic status; (c) In terms of temporal activity, standard language was more likely to be used during the daytime while non-standard variants were predominant during the night. We explained this temporal variability by the turnover of population with different socioeconomic status active during night and day; Finally (d) we showed that the social network and status homophily mattered in terms of linguistic similarity between peers, as connected users with the same socioeconomic status appeared to be the most similar, while disconnected people were found to be the most dissimilar in terms of their individual use of the aforementioned linguistic markers.

Despite these findings, one has to acknowledge the multiple limitations affecting this work: First of all, although Twitter is a broadly adopted service in most technologically enabled societies, it commonly provides a biased sample in terms of age and socioeconomic status as older or poorer people may not have access to this technology. In addition, home locations inferred for lower activity users may induced some noise in our inference method. Nevertheless, we demonstrated that our selected Twitter users are quite representative in terms of spatial, temporal, and socioeconomic distributions once compared to census data. Other sources of bias include the "homogenization" performed by INSEE to ensure privacy rights are upheld as well as the proxies we devised to approximate users' home location and social network. Currently, a sample survey of our set of geolocated users is being conducted so as to bootstrap socioeconomic data to users and definitely validate our inference results. Nonetheless, this INSEE dataset provides still the most comprehensive available information on socioeconomic status over the whole country. For limiting such risk of bias, we analyzed the potential effect of the confounding variables on distribution and cross-correlations of SES indicators. Acknowledging possible limitations of this study, we consider it as a necessary first step in analyzing income through social media using datasets orders of magnitude larger than in previous research efforts.

Finally we would like to emphasize two scientific merits of the paper. On one side, based on a very large sample, we confirm and clarify results from the field of sociolinguistics and we highlight new findings. We thus confirm clear correlations between the variable realization of the negative particle in French and three indices of socioeconomic status. This result challenges those among the sociolinguistic studies that do not find such correlation. Our data also suggested that the language used in the southern part of France is more standard. Understanding this pattern fosters further investigations within sociolinguistics. We finally established that the linguistic similarity of socially connected people is partially explained by status homophily but could be potentially induced by social influences passing through the network of links or other terms of homophilic correlations. Beyond scientific merit, we can identify various straightforward applications of our results. The precise inference of socioeconomic status of individuals from online activities is for instance still an open question, which carries a huge potential in marketing design and other areas. Our results may be useful moving forward in this direction by using linguistic information, available on Twitter and other online platforms, to infer socioeconomic status of individuals from their position in the network as well as the way they use their language.

A APPENDIX: Status homophily

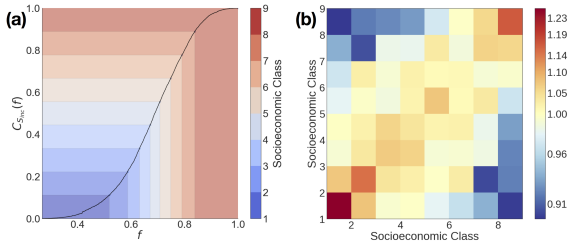


Figure 6: (a) Definition of socioeconomic classes by partitioning users into nine groups with the same cumulative annual income. (b) Structural correlations between SES groups depicted as matrix of the ratio $|E(s_i, s_j)|/|E_{rand}(s_i, s_j)|$ between the original and the average randomized mention network

Status homophily in social networks appears as an increased tendency for people from similar socioeconomic classes to be connected. This correlation can be identified by comparing likelihood of connectedness in the empirical network to a random network, which conserves all network properties except structural correlations. To do so, we took each (s_i, s_j) pair of the nine SES class in the Twitter network and counted the number of links $|E(s_i, s_j)|$ connecting people in classes s_i and s_j. As a reference system, we computed averages over 100 corresponding configuration model network structures [39]. To signalize the effects of status homophily, we took the ratio $|E(s_i, s_j)|/|E_{rand}(s_i, s_j)|$ of the two matrices (shown in Fig.6b). The diagonal component in Fig.6b with values larger than 1 showed that users of the same or similar socioeconomic class were better connected in the original structure than by chance, while the contrary was true for users from classes far apart (see blue off-diagonal components). To verify the statistical significance of this finding, we performed a χ^2-test, which showed that the distribution of links in the original matrix was significantly different from the one of the average randomized matrix ($p < 10^{-5}$). This observation verified status homophily present in the Twitter mention network.

REFERENCES

[1] Oluwaseun Ajao. 2015. A survey of location inference techniques on Twitter. *Journal of Information Science, 1-10* (2015). https://doi.org/10.1177/0165551510000000

[2] William J Ashby. 2017. Un nouveau regard sur la chute du ne en tourangeau : s'agit-il d'un francais parle changement en cours? *Journal of French Language Studies* 11, 2001 (2017).

[3] Catherine Brissaud. 1999. La realisation de l'accord du participe passe employe avec avoir. De l'influence de quelques variables linguistiques et sociales. *Langage et societe* 88, 1 (1999), 5–24. https://doi.org/10.3406/lsoc.1999.2866

[4] Kathryn Campbell-Kibler. 2010. New directions in sociolinguistic cognition. *University of Pennsylvania Working Papers in Linguistics* 15, 2 (2010), 31–39. http://repository.upenn.edu/pwpl/vol15/iss2/5/

[5] J. K Chambers. 1995. *Sociolinguistic theory : linguistic variation and its social significance.* Wiley-Blackwell; Cambridge, Mass. Paperback.

[6] Collectif, Vincent Lucci, and Agnès Millet. 1994. *L'orthographe de tous les jours. Enquête sur les pratiques orthographiques des Français.* Honoré Champion, Paris.

[7] Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation* 46, 4 (2012), 721–736. https://doi.org/10.1007/s10579-012-9193-0

[8] Nathan Eagle, Rob Claxton, and Michael W Macy. 2010. Network Diversity and Economic Development. *Science* 328 (2010), 1029–1031.

[9] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of Lexical Change in Social Media. *PLOS ONE* 9, 11 (11 2014), 1–13. https://doi.org/10.1371/journal.pone.0113114

[10] Martin Fixman, Ariel Berenstein, Jorge Brea, Martin Minnoni, and Carlos Sarraute. 2016. Inference of Socioeconomic Status in a Communication Graph. *Argentine Symposium on Big Data (AGRANDA)* (2016), 95–106.

[11] Mark Graham, Scott A Hale, and Devin Gaffney. 2017. Where in the World Are You ? Geolocation and Language Identification in Twitter Identification in Twitter. *The Professional Geographer* 66, April (2017), 568–578. https://doi.org/10.1080/00330124.2014.907699

[12] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008).* Pasadena, CA USA, 11–15.

[13] William L. Hamilton, Jure Leskovec, and Daniel Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *CoRR* abs/1605.09096 (2016).

[14] Anita Berit Hansen and Isabelle Malderez. 2004. une étude en temps réel. *Langage & Société* (2004), 5–30. https://doi.org/10.3917/ls.107.0005

[15] L. Henry, S. Barbu, A. Lemasson, and M. Hausberger. 2015. Dialects in animals: Evidence, development and potential functions. *Animal Behavior and Cognition* 2, 2 (2015), 132–155. http://abc.sciknow.org/archive_files/201502/03.Henry_FINAL.pdf

[16] Philippe Hert. 1999. Quasi-oralite de l'ecriture electronique et sentiment de communaute dans les debats scientifiques en ligne. *Reseaux* 17, 97 (1999), 211–259. https://doi.org/10.3406/reso.1999.2171

[17] Erika Hoff. 2003. The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development* 74, 5 (2003), 1368–1378. https://doi.org/10.1111/1467-8624.00612

[18] Hadrien Hours, Eric Fleury, and Márton Karsai. [n. d.]. Link prediction in the Twitter mention network: impacts of local structure and similarity of interest. *ICDMW'16* ([n. d.]), 95–106.

[19] Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User Review Sites As a Resource for Large-Scale Sociolinguistic Studies. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 452–461. https://doi.org/10.1145/2736277.2741141

[20] Bernardo Huberman, Daniel Romero, and Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *First Monday* 14, 1 (2008). https://doi.org/10.5210/fm.v14i1.2317

[21] Janellen Huttenlocher, Marina Vasilyeva, Heidi R. Waterfall, Jack L. Vevea, and Larry V. Hedges. 2007. The Varieties of Speech to Young Children. *Developmental Psychology* 43, 5 (9 2007), 1062–1083. https://doi.org/10.1037/0012-1649.43.5.1062

[22] INSEE. 2016. (2016). https://www.insee.fr/fr/statistiques/2119431?sommaire=2119504

[23] INSEE. 2016. (2016). https://www.insee.fr/fr/statistiques/2520034

[24] Katherine D. Kinzler, Emmanuel Dupoux, and Elizabeth S. Spelke. 2007. The native language of social cognition. *Proceedings of the National Academy of Sciences* 104, 30 (2007), 12577–12580. http://www.pnas.org/content/104/30/12577.short

[25] William A. Kretzschmar. 2010. Language Variation and Complex Systems. *American Speech* 85, 3 (2010), 263–286. https://doi.org/10.1215/00031283-2010-016

[26] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 625–635. https://doi.org/10.1145/2736277.2741627

[27] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. In *ICWSM.*

[28] Jussi M. Kumpula, Jukka-Pekka Onnela, Jari Saramäki, Kimmo Kaski, and János Kertész. 2007. Emergence of Communities in Weighted Networks. *Phys. Rev. Lett.* 99 (Nov 2007), 228701. Issue 22. https://doi.org/10.1103/PhysRevLett.99.228701

[29] Blattner Marcel Kunegis, Jerome and Christine Moser. 2013. Birds of a feather: Homophily in social networks. *Proceedings of the 5th Annual ACM Web Science Conference WebSci '13 Paris, France,ACM, New York, NY, USA.* (2013), 205–214.

[30] William Labov. 1966. *The Social Stratification of English in New York City.* Center for Applied Linguistics, Washington.

[31] William Labov. 1972. *Sociolinguistic Patterns* (blackwell ed.). University of Pennsylvania Press.

[32] Bernard Laks. 2013. Why is there variation rather than nothing? *Language Sciences* 39 (2013), 31–53. https://doi.org/10.1016/j.langsci.2013.02.009

[33] Guillaume Laurent, Jari Saramäki, and Márton Karsai. [n. d.]. From calls to communities: a model for time-varying social networks. *Eur. Phys. J. B* 88 ([n. d.]).

[34] David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. [n. d.]. Life in the network: the coming age of computational social science. *Science* 323, 5915 ([n. d.]), 721–723. https://doi.org/10.1126/science.1167742

[35] Yannick Leo, Eric Fleury, Carlos Sarraute, Ignacio Alvarez-hamelin, and Márton Karsai. 2016. Socioeconomic correlations in communication networks. *J. R. Soc. Interface* 13 (2016).

[36] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. Social Media Fingerprints of Unemployment. *PLOS ONE* 10, 5 (05 2015), 1–13. https://doi.org/10.1371/journal.pone.0128692

[37] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 51 – 56.

[38] Miller McPherson, Lovin Lynn S., and Cook James M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* (2001), 415–444.

[39] Mark Newman. 2010. *Networks: an introduction.* Oxford university press.

[40] Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Comput. Linguist.* 42, 3 (Sept. 2016), 537–593. https://doi.org/10.1162/COLI_a_00258

[41] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data. *EMNLP 2015* (2015).

[42] Daniel Preot, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (2015), 1754–1764.

[43] Patrick S Park, Minsu Park, and Michael W Macy. 2017. Economic Opportunity and Network Position Patrick. *Encyclopedia of African American Popular Culture, Vol. 1* NetSci 2017 (2017).

[44] Sanja Šćepanović, Igor Mishkovski, Bruno Gonçalves, Trung Hieu Nguyen, and Pan Hui. 2017. Semantic homophily in online communication: evidence from twitter. *Online Social Networks and Media* 2 (2017), 1–18.

[45] Corinne Totereau, Catherine Brissaud, Caroline Reilhac, and Marie-line Bosse. 2013. L'orthographe grammaticale au college : une approche sociodifferencielle. *Approche Neuropsychologique de Apprentissages de l'Enfant* 123 (2013), 164–171.

[46] Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2011. Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE* 6 (09 2011), 1–14. https://doi.org/10.1371/journal.pone.0023613

De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis

CORPUS

Corpus

20 | 2020
Corpus complexes

De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis

Segmentation devices in tweets: punctuation marks, connectives, emoticons and emojis

Jean-Philippe Magué, Nathalie Rossi-Gensane and Pierre Halté

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/corpus/4619
DOI: 10.4000/corpus.4619
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Jean-Philippe Magué, Nathalie Rossi-Gensane and Pierre Halté, « De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis », *Corpus* [Online], 20 | 2020, Online since 24 January 2020, connection on 08 September 2020. URL : http://journals.openedition.org/corpus/4619 ; DOI : https://doi.org/10.4000/corpus.4619

De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis[1]

Segmentation devices in tweets: punctuation marks, connectives, emoticons and emojis

Jean-Philippe MAGUÉ*, Nathalie ROSSI-GENSANE**
et Pierre HALTÉ***
* ICAR (UMR 5191) et ENS de Lyon ; ** ICAR (UMR 5191) et Université Lumière Lyon 2 ; *** EDA (EA 4071) et Université de Paris

Résumé
Dans cet article, nous appuyant sur un corpus de 3 444 075 tweets correspondant à 44 107 210 tokens (mots, signes de ponctuation, émojis, émoticônes, etc.) recueillis en décembre 2016, nous nous intéressons aux procédés de segmentation à l'œuvre dans les tweets. Après avoir évoqué certaines caractéristiques de ces écrits particuliers, nous rappelons les procédés généraux de segmentation à l'écrit : les signes de ponctuation et les connecteurs. Nous nous penchons ensuite sur la segmentation opérée dans les tweets par ces deux procédés généraux. Enfin, nous montrons que les émoticônes et les émojis constituent des procédés spécifiques permettant de diversifier les stratégies de segmentation des utilisateurs de tweets (et d'autres écrits numériques, tels les SMS et les courriels).

Mots-clés
segmentation, tweets, signes de ponctuation, connecteurs, émoticônes, émojis

1 Étude réalisée dans le cadre du projet SoSweet soutenu par l'ANR (ANR-15-CE38-0011-03).

Abstract
In this paper, relying on a corpus of 3,444,075 tweets corresponding to 44 107 210 tokens (words, signs of punctuation, emojis, emoticons, etc.) collected in December 2016, we focus on segmentation processes at work in tweets. After mentioning some characteristics of these particular writings, we review the general segmentation processes in writing, punctuation and connectors. We then look at how these processes operate in tweets. Finally, we show that emoticons and emojis are specific processes allowing users to diversify their segmentation strategies (and other digital writings, such as SMS and email).

Keywords
segmentation processes, tweets, punctuation marks, connectives, emoticons, emojis

Introduction

Les tweets font partie de ces genres liés à l'apparition de nouvelles technologies (par exemple, les SMS, les tchats, les courriels), qui, « manifestation[s] de l'immédiat dans le graphique », « viennent introduire [du] bougé dans les paramètres fondamentaux de la communication par oral et par écrit » alors que, « jusqu'à il y a un siècle et demi environ, l'oral et l'écrit se présentaient comme des entités stabilisées dans quelques propriétés fondamentales » (Gadet, 2007-2008 : 135 et 136).

Dans cet article, nous appuyant sur un corpus de 3 444 075 tweets correspondant à 44 107 210 tokens (mots, signes de ponctuation, émojis, émoticônes, etc.) recueillis en décembre 2016, nous nous intéressons aux procédés de segmentation à l'œuvre dans les tweets. Après avoir évoqué certaines caractéristiques de ces écrits particuliers, nous rappelons les procédés généraux de segmentation à l'écrit : les signes de ponctuation et les connecteurs. Nous nous penchons ensuite sur la segmentation opérée dans les tweets par ces deux procédés généraux. Enfin, nous nous interrogeons sur le rôle des émoticônes et des émojis en tant que procédés de segmentation

spécifiques aux tweets (à côté d'autres écrits numériques, tels les SMS et les courriels).

1. Les tweets : considérations générales

À la suite de Söll (1974), Koch & Oesterreicher (2001) différencient nettement, dans tout énoncé, l'aspect médial, lié au canal, généralement oral ou écrit, et l'aspect conceptionnel, pour sa part lié à la proximité ou la distance communicatives. Dans les tweets, écrits sur le plan médial, ressort particulièrement « le caractère scalaire » (Koch & Oesterreicher 2001 : 586) du plan conceptionnel. Certains « paramètres de la communication » (Koch & Oesterreicher 2001 : 586) sont en effet clairement déterminés, tels le « détachement actionnel et situationnel », le « détachement référentiel de la situation », la « séparation spatio-temporelle » (tout au moins la séparation spatiale, une coprésence temporelle étant possible) et, sans doute, une « coopération communicative minime » et une « liberté thématique ». En revanche, d'autres paramètres sont soumis à une forte gradation interne : les tweets livrent des messages susceptibles de relever de la « communication privée » ou, s'il s'agit de tweets collectifs, « publique », avec « interlocuteur intime » ou « inconnu », empreints d'une « émotionnalité forte » ou « faible », s'inscrivant dans un « monologue » ou dans un « dialogue » (mais, dans ce dernier cas, généralement, en différé), la communication pouvant être « spontanée » comme « préparée ». Les tweets sont donc dotés d'un « relief conceptionnel » (Koch & Oesterreicher 2001 : 586) à géométrie variable. De ce fait, si l'on convoque la notion de genre textuel, définie comme une « corrélation entre fonctionnements linguistiques et situation extralinguistique » (Condamines 2006 : 633), sans doute convient-il de parler de genres, au pluriel, des tweets.

Les tweets peuvent de plus être considérés comme relevant de ce que l'on appelle parfois les genres brefs (voir notamment Roukhomovsky (2001)). Jusqu'en novembre 2017 (date postérieure au recueil de notre corpus qui a eu lieu en décembre 2016), leur longueur était limitée à 140 caractères (elle peut aujourd'hui aller jusqu'à 280). Les formes brèves que

constituent les tweets sont à elles seules un texte séparé (ce qui ne serait pas nécessairement le cas pour des genres brefs tels que les maximes, susceptibles de se trouver au sein d'un autre texte, par exemple en tant que morale d'une histoire), matériellement contraint par la limite du nombre de caractères, d'une façon, par certains côtés, comparable à l'épigramme qui « ne devait contenir plus de vers qu'il s'en pouvait écrire dessus un portail » (Sébillet 1548).

Enfin, dans cette réflexion sur la segmentation dans les tweets, il est important de souligner qu'en pays de littératie, que Gadet (2007 : 45) décrit comme « la pratique généralisée de la lecture-écriture, les effets d'une culture de l'écrit sur les énoncés, les pratiques, attitudes et représentations, pour un locuteur ou une communauté de locuteurs », la standardisation s'applique d'abord aux genres de l'écrit et, surtout, à ceux relevant de l'écrit littéraire, dans la mesure où ce sont « les meilleurs écrivains qui [sont considérés comme] fourniss[a]nt les modèles les plus sûrs de la Norme » (Lodge 2011 : 87). On peut ainsi s'attendre à ce que les tweets tendent à être moins normés que d'autres écrits, en particulier sur le plan de la segmentation qui nous intéresse ici.

2. Les tweets du corpus

Notre corpus a été acquis auprès de l'entreprise Gnip, et correspond à 10% des tweets produits en décembre 2016 dans les fuseaux horaires GMT et GMT+1 et identifiés par Twitter comme étant en français. Afin d'éliminer les tweets créés automatiquement par des robots, nous avons filtré notre corpus en fonction du client (le logiciel) utilisé pour produire les tweets. Nous avons écarté les clients contenant le mot *bot* (pour robot) ou comptant moins de 1000 tweets. Les tweets automatiques étant très standardisés, nous avons calculé l'écart-type de la longueur des tweets pour chaque client. Ceux pour lesquels cet écart-type est inférieur à 6, c'est-à-dire présentant une faible diversité, ont été également éliminés. Notre corpus final compte 34 clients différents. Les cinq plus fréquents (Twitter for iPhone, Twitter for Android, Twitter Web Client, TweetDeck et Twitter for iPad), qui représentent 97.7% des tweets de notre corpus, sont clairement des logiciels conçus pour être utilisés par des humains

et non des robots. Le corpus initial se montait à 4 243 789 tweets ; le corpus filtré en comprend 3 444 075. Ce corpus constitue donc un échantillon substantiel des énoncés produits par les utilisateurs francophones de Twitter en Europe (et, dans une moindre mesure, en Afrique). Sans nier la diversité des productions qu'il contient (analysée par ailleurs, voir par exemple Abitbol *et al.* (2015)), notre étude se concentre sur les procédés de segmentation à l'échelle de l'ensemble de ce corpus (Bolander & Locher 2014).

Dans le corpus, nous avons identifié et décompté les séquences (éventuellement constituées d'un seul élément) de signes de ponctuation, d'émoticônes ou d'émojis avec des expressions régulières. Dans l'exemple ci-dessous, :((((compte pour une séquence d'émoticônes, ?? pour une séquence de signes de ponctuation, et 😂😂😂😂 pour une séquence d'émojis :

(1) @X je sais jamais les anniversaires en vrai :((((mais c'est le geste qui compte n'est ce pas?? MDRRRR😂😂😂😂

Dans le cadre de cet article, nous nous sommes limités aux signes de ponctuation suivants : la virgule, le point, le point d'interrogation, le point d'exclamation et les points de suspension. Les émojis et les émoticônes retenus sont, quant à eux, donnés en annexe. Par ailleurs, étant donnée la diversité des modes de présentation des tweets selon le logiciel utilisé, nous avons délibérément opté pour une mise en forme de nos exemples qui ne préserve que leur contenu linguistique, sans essayer de mimer l'interface d'un logiciel arbitraire. La graphie d'origine a été conservée. De plus, comme il apparaît dans l'exemple ci-dessus, nous avons anonymisé les mentions aux utilisateurs en les remplaçant systématiquement par @X.

3. Les procédés généraux de segmentation à l'écrit : les signes de ponctuation et les connecteurs

Tout texte, pour être appréhendé, doit être préalablement découpé. Ce rôle, tout au moins dans les écrits normés, est

généralement dévolu à la ponctuation, définie de manière très large comme l'« ensemble des signes graphiques non alphabétiques utilisés dans un texte pour noter les rapports syntaxiques entre les divers éléments de la phrase ou de la proposition, les rapports avec le sens, les idées du texte, les variations d'ordre affectif (intonation, rythme, mélodie de la phrase) » (TLFi, CNRTL). En particulier, Dürrenmatt (2015 : 23) envisage, parmi les « opérations proposées au lecteur par la ponctuation », celle de l'« agence[ment] », grâce à la virgule, au point-virgule, au deux-points, au point et, dans une moindre mesure, au point d'interrogation, au point d'exclamation, aux points de suspension : « le ponctuant rend visibles des unités et permet de les interpréter pour elles-mêmes et dans leurs rapports les unes avec les autres ». Néanmoins, cette opération d'agencement n'est pas réservée aux seuls signes de ponctuation, comme le souligne également Dürrenmatt (2015 : 23), car, dans la mesure où les unités susmentionnées « sont de même niveau : la ponctuation participe […] de la jonction ». Dürrenmatt (2015 : 23) signale ainsi que « les ponctuants peuvent […] accompagner des connecteurs ». Outre le possible cumul entre signes de ponctuation et connecteurs, il est important de préciser que les connecteurs sont dotés de la capacité d'assurer seuls cette mission d'agencement et que, par conséquent, signes de ponctuation et connecteurs doivent être vus comme agissant à l'écrit de manière complémentaire. De plus, dans les tweets, comme dans des écrits d'autres genres liés aux technologies numériques, tels les SMS ou les courriels, le scripteur a la possibilité de recourir à des émoticônes et à des émojis, susceptibles d'apparaître aux mêmes endroits de la chaîne écrite (voir *infra*, partie 5). De par la complémentarité entre signes de ponctuation et connecteurs, que l'on peut aussi envisager d'étendre ici aux émoticônes et aux émojis, il semble pertinent d'utiliser la notion de segmentation, plus large que celle de ponctuation.

La segmentation, « division en segments » d'après le TLFi (CNRTL), sera entendue comme la démarcation visible, matérialisée à l'aide de bornes de diverses sortes (signes de ponctuation, connecteurs, etc., mais aussi, comme il sera discuté

infra, émoticônes et émojis), d'un texte en unités. Toutefois, selon l'unité adoptée, cette démarcation peut se situer à plusieurs niveaux. Dans cet article, nous privilégierons l'unité maximale syntaxique, que nous dénommons phrase « syntaxique » (voir par exemple Paolacci & Rossi-Gensane (2012), à partir de la notion de phrase chez Feuillard (1989)) et qui est loin de toujours correspondre à la phrase telle qu'elle est conçue traditionnellement, c'est-à-dire comme étant délimitée à l'initiale par une majuscule et en finale par un point (ou tout autre signe de ponctuation dite forte, comme le point d'exclamation, le point d'interrogation, les points de suspension). La phrase « syntaxique » est définie comme constituée de l'ensemble des éléments reliés par des rapports de dépendance à un même élément central, qui n'est pas nécessairement verbal (comme, par exemple, dans *Quelle belle journée !*, où l'élément central est *journée*). Lorsque l'élément central est verbal, la phrase « syntaxique » équivaut généralement à une proposition indépendante ou à une proposition principale accompagnée de sa (ou ses) proposition(s) subordonnée(s). Ainsi, dans le cas des propositions traditionnellement dites coordonnées, *Marie déteste la soupe et elle adore le chocolat.*, là où l'on voit habituellement une seule phrase déterminée à partir d'un critère typographique (majuscule-point), nous dégagerons, à partir d'un critère relationnel, plusieurs phrases « syntaxiques », en l'occurrence deux (d'une part, *Marie déteste la soupe*, d'autre part, *et elle adore le chocolat*, les éléments du premier ensemble n'entretenant aucune relation de dépendance syntaxique avec ceux du second ensemble, et inversement). Dans cet exemple normé, on note qu'il ne manque aucune borne : la première, entre les deux phrases « syntaxiques » (plus précisément, au début de la seconde phrase « syntaxique »), est matérialisée par un connecteur (*et*) ; la seconde, à la fin de la seconde phrase « syntaxique », est matérialisée par un point. Alors que la phrase traditionnelle, « typographique » dans les faits, est certes immédiatement identifiable en réception mais instable en production (si l'on pense aux innombrables façons dont un texte est susceptible d'être ponctué), la phrase « syntaxique », qui s'abstrait de la ponctuation et s'appuie sur un critère syntaxique,

garantit généralement sur ce dernier plan un même « calibrage » tout au long du texte.

La segmentation peut également être prise en compte en deçà de la phrase « syntaxique ». C'est ainsi que, par exemple, à côté d'une virgule séparant deux phrases « syntaxiques », appelée virgule « interphrastique » (et alors dotée d'un rôle de ponctuation forte), nous distinguerons une virgule « intraphrastique » séparant des mots ou des syntagmes[2]. De même, un mot en apparence identique, tel *et*, peut jouer le rôle de connecteur quand il sépare deux phrases « syntaxiques » ou, à la manière de la virgule « intraphrastique », le rôle de coordonnant quand il sépare (tout en les reliant) deux éléments de même fonction.

Enfin, on signalera une différence d'orientation entre signes de ponctuation et connecteurs, particulièrement nette lorsqu'ils ont un rôle de démarcation interphrastique. Alors que les signes de ponctuation clôturent ce qu'ils segmentent, les connecteurs, à l'inverse, l'ouvrent, ce qui rend notamment compte du fait que, sauf exception, un connecteur n'apparaîtra pas en finale de texte et que, généralement, dans les cas où signe(s) de ponctuation et connecteur(s) sont contigus, ceux-là précèdent ceux-ci.

4. La segmentation dans les tweets par la ponctuation et les connecteurs

4.1. La segmentation dans les tweets par la ponctuation

Nous avons identifié 472 séquences de signes de ponctuation différentes, avec 0.84 séquence par tweet en moyenne. Les cinq séquences les plus fréquentes sont constituées d'un seul signe :

, 0.226 par tweet
. 0.181 par tweet
! 0.091 par tweet
… 0.090 par tweet

2 Pour la tradition, qui n'envisage qu'une phrase « graphique », la virgule est par définition nécessairement intraphrastique.

? 0.083 par tweet

Pour ce qui concerne les signes de ponctuation, au premier abord, on note que sont employées dans les tweets davantage de virgules que de points, ce qui peut notamment s'expliquer par le fait que, du point de vue de la segmentation, la virgule est un signe bivalent, potentiellement interphrastique ou intraphrastique, contrairement au point, presque toujours interphrastique. Sans doute en relation à cette bivalence, la virgule est considérée comme un signe particulièrement complexe, voire « le plus difficile de tous les signes de ponctuation » (Popin 1998 : 38 ; voir aussi Catach (1994 : 64-71) ou Drillon (1991 : 145)).

Cette supériorité numérique de la virgule sur le point apparaît tout particulièrement dans des tweets où la virgule est utilisée dans un rôle interphrastique (où elle se substitue, donc, au point) et où, possiblement, le point final est absent, phénomène relativement répandu qui semble indiquer qu'à cet emplacement, le point est jugé redondant. En effet, seulement 6.33% des tweets ont un point final et, plus généralement, seulement 18,6% se terminent par un signe de ponctuation, soit, par ordre décroissant, le point, le point d'interrogation, le point d'exclamation.

(2) Emma elle connaît trop de mots/expressions françaises, jla soupçonne d'avoir passé les 3/4 de sa vie ici

L'exemple suivant, outre des virgules interphrastiques, comporte également une virgule intraphrastique située entre *mois* et *voir* [*sic*] :

(3) @X @X Non, je pense qu'il est même sorti y'a moins d'un mois, voir une semaine, il est même pas référencé sur Amazon US

Si l'exemple (4) nécessite une borne interphrastique matérialisée après *sérieux*, il contient une virgule intraphrastique après *hier* :

(4) Sérieux depuis hier, j'ai des messages d'alerte qu'il n'est pas en état de focntionner

Les exemples de cette sorte ne sont pas sans évoquer l'usage de la virgule « par défaut » que Bessonnat (1991 : 24) souligne dans une perspective didactique pour les élèves de collège : tout se passe « comme si l'élève, emporté par le flux de l'écriture, craignait de prendre la décision coûteuse du point qui casserait l'inspiration ». Il convient tout de même de signaler, à l'inverse, des cas totalement normés sur le plan de la segmentation, où les tweets sont composés de phrases, à la fois « syntaxiques » et « graphiques », démarquées par une majuscule à l'initiale et un point en finale :

(5) J'ai pas envie de retourner au collège demain.

(6) Les pions sont aller chercher les gens de ma classe au foyer. Ils leur ont donné des exercices. J'étais pas là.

Dans d'autres cas, moins normés, le point final est présent, toutefois sans qu'il y ait de majuscule à l'initiale, ou alors c'est la majuscule qui est présente sans le point :

(7) @X j'ai inventé un mot je crois, c'était de rigueur.

(8) Mon cousin va rester tout seul en Allemagne

Il est intéressant de remarquer que les écarts par rapport à la norme impliquant les signes de ponctuation et les majuscules relèvent pour les uns et les autres des erreurs à « dominante idéogrammique » (Catach 2003 : 282).

Peut aussi être constatée la supériorité numérique du point d'exclamation sur le point d'interrogation (qui se reproduit pour les séquences à plusieurs éléments, le point d'exclamation dupliqué (0.01050 par tweet) étant par exemple plus fréquent que le point d'interrogation dupliqué (0.00430 par tweet)). Cette

constatation est conforme à la remarque suivante de Dürrenmatt (2015 : 15) :

« S'il [le point d'interrogation] reste obligatoire dans le "bon usage", c'est par volonté d'uniformisation. L'ajout par le scripteur du signe à sa phrase interrogative relève donc du respect de la norme et non d'une nécessité communicationnelle. Il apparaît dès lors comme facilement volatile alors que le point d'exclamation, indice d'une modalité affective moins explicitement visible à travers des modifications morphosyntaxiques, résiste beaucoup mieux. »

La présence d'un point d'interrogation dans l'exemple (9) n'est en effet pas liée à une « nécessité communicationnelle », la modalité interrogative étant déjà indiquée par un élément interrogatif, ici *in situ* (c'est-à-dire en même place que l'élément correspondant dans une phrase assertive) :

(9) @X donc pour vous ça prouve quoi?[3] Certainement pas l'absence de conflits d'intérêts! #Filimbi #lucha

En revanche, le point d'interrogation a une valeur distinctive (et significative) pour les interrogatives totales dans les exemples suivants, qui, conformément à une tendance importante de l'oral, ne bénéficient d'aucun marquage morphologique :

(10) @X j'essaye de m'améliorer... T'es forte en classe toi?

(11) @X comme d'habitude , tu passes à la cafetria ?

Dans l'exemple suivant contenant un point d'exclamation, la « modalité affective » qu'il indique est néanmoins « visible à travers des modifications

3 La présence ou l'absence d'un espace avant un signe de ponctuation n'ont pas été commentées.

morphosyntaxiques » (Dürrenmatt 2015 : 15), c'est-à-dire l'adverbe exclamatif *tellement* :

(12) @X un vrai délice, et tellement d'actualité!

Dans de nombreux cas, les points d'exclamation et d'interrogation sont utilisés en séquences de plusieurs éléments, généralement homogènes (s'il s'agit de la répétition du même signe), néanmoins quelquefois hétérogènes[4] :

(13) @X non je trouve que ton Wolf est bien fait !!!!! J'aime surtt les yeux !!!!

(14) Je sors du kine, y a une voiture allumé je tourne la tête et je vois une chèvre dans la voiture... C'est quoi ce Putain de délire ??

(15) Qui c'est qui taff dans 4h ?!??? C bibi

Si, pour Popin (1998 : 34 et 36), qui s'est surtout penché sur les écrits littéraires, il est « possible de faire des séries de points d'interrogation » et « il n'est pas impossible de répéter les points d'exclamation, voire de combiner un point d'exclamation et un point d'interrogation », pour Dürrenmatt (2015 : 26), en revanche, « il est courant de multiplier les points d'exclamation ou d'interrogation pour signaler l'intensité de la modalisation ». Dürrenmatt (2015 : 100) envisage les emplois de cette sorte comme relevant de la visilibilité définie par Anis (1983 : 89), selon laquelle « les formes graphiques [sont] un corps signifiant intégré aux isotopies textuelles ». Dürrenmatt (2015 : 101) évoque également cet usage de la ponctuation déploré par le poète italien Leopardi, où les « idées [sont] représent[ées] par des agencements de signes de ponctuation au lieu [d'être] exprim[ées] de façon complexe par des mots ».

4 Par exemple, les séquences homogènes !!!, !!!! et !!!!! sont respectivement utilisées à hauteur de 0.00794, 0.00255 et 0.00095 par tweet ; les séquences homogènes ??? et ???? sont respectivement utilisées à hauteur de 0.00317 et 0.00126 par tweet ; les séquences hétérogènes ?! et !? sont respectivement utilisées à hauteur de 0.00191 et 0.00036 par tweet.

Enfin, pour ce qui concerne les points de suspension, Dürrenmatt (2015 : 94) en souligne le rôle « affectif », « dans la mesure où ils marquent la perte de contrôle, le refus de dire, la difficulté à maîtriser son discours et, par là même, rejoignent l'exclamation » :

(16) Ça me manque le temps ou je préparais mes poissons, je les mettais dans mon sac avec du scotch et je les collais sur les gens à la récré...

(17) @X Pendant ce temps à Taïwan, on peut faire balayer la cours aux fouteurs de merde...

Apparaît aussi un emploi minoritaire du « point multiple » (Drillon 1991 : 136) parfois constitué de deux points (0.01815 par tweet, soit un emploi cinq fois moindre que celui normé des trois points) :

(18) @X même mes parents préfère ma soeur..

Drillon (1991 : 137) évoque l'usage que Françoise Sagan avait souhaité faire de ce « point multiple » à deux points, dans le titre de son roman *Aimez-vous Brahms..*

4.2. La segmentation dans les tweets par les connecteurs

Pour ce qui concerne les connecteurs, nous nous pencherons, dans le cadre de cet article, surtout sur l'élément *mais*, qui, avec *et*, fait partie des connecteurs et coordonnants les plus fréquents (respectivement, 207 270 occurrences et 339 515 occurrences). À la différence de *et* dont les emplois de coordonnant sont moins rares, *mais* est presque toujours connecteur dans les tweets, où il se comporte donc comme une sorte d'hyperconnecteur.

Dans l'exemple (19), conformément à un usage écrit répandu, *mais* exprime une contradiction logique (accompagné, dans la phrase « syntaxique » suivante, par un autre connecteur, *donc*, exprimant, quant à lui, une déduction logique) :

(19) Je suis motivée mais je vais être traumatisée donc vous avez intérêt à me soutenir

Souvent, on note un cumul signe de ponctuation (surtout la virgule)/connecteur dans cet ordre, conformément à une tendance extrêmement répandue (dans les cas de contiguïté entre signes de ponctuation et connecteurs, la ponctuation précède le connecteur dans 74478 cas et le suit dans 6988 cas) :

(20) @X Bonjour, j'ai eu écho comme quoi, on pouvait acheter des boosters avec des ogrines,mais je ne trouve pas l'option dans la boutique

(21) @X S'insurger pour quoi? C'est internet. Je trouve le post que t'as mis révoltant, mais je vais pas partir en croisade pour autant.

Toutefois, dans de nombreux cas, *mais* semble indiquer, plutôt que l'« inversion argumentative [formulée dans les travaux d'Oswald Ducrot] », un « [changement] de point de vue sur [l']objet [de discours] » selon les termes de Morel & Danon-Boileau (1998 : 118) caractérisant cet élément à l'oral :

(22) Visite sur TF1.fr (pas ma faute j'ai cliqué). Vous entendez pas mais derrière y a en + une vidéo de pub en autoplay. http://t.co/ZetwU4VRGv

Cette valeur est particulièrement nette en début de question, *mais* fonctionnant plus comme une « particule énonciative » (Bouchard 2001) que comme un connecteur :

(23) Mais c'est quoi la différence entre un pansexuel et un bisexuel?

(24) @X mais t'es sérieux à faire des histoires comme ça !? Ouah mais le gars

Dans l'exemple précédent, on notera en outre une seconde occurrence de *mais* dans un usage exclamatif (malgré l'absence d'un point d'exclamation), qui, d'après Bouchard (2001 : 66), « d'une part manifeste une émotion ressentie individuellement et d'autre part la verbalise, la donne à entendre et à partager à autrui ».

Bouchard (2001 : 71) note des usages variés pour des éléments tels que *mais, alors* ou *donc*, entre « purs régulateurs de l'action » dans les oraux polygérés et « instructions d'inférence » dans les écrits monogérés. Cette diversité semble se déployer tout particulièrement dans les tweets, écrits souvent spontanés, de par leur nature hybride.

Outre les connecteurs, d'autres éléments apparaissent comme jouant sous forme de mot(s) un rôle démarcatif dans les tweets. Ainsi, Morel & Danon-Boileau (1998 : 94) évoquent, pour l'oral, des éléments à l'initiale, tels *bon* et, tout au moins dans certains cas, *oui* ou *ouais*, appelés « ligateurs énonciatifs », qui, avec d'autres éléments en finale, appelés « ponctuants », tels *hein* et *quoi*, assurent « la régulation de la coénonciation, qui permet d'expliciter la position de l'énonciateur ». Dans l'exemple (25), de même qu'à l'oral, des ligateurs énonciatifs constituent une borne à gauche (en cela, à la manière des connecteurs) :

(25) Ouai bon je viens de voir un tweet disant que l'histoire de la S2 de tokyo ghoul était bien on va aller ce recoucher en faite

Dans l'exemple suivant, on note, à la frontière entre les deux premières phrases « syntaxiques », un cumul virgule/connecteur/ligateur énonciatif/virgule :

(26) @X Et c'etait le cas, mais bon, je ne pense pas que Oda puisse nous faire une chose pareille, pendant ce temps, apprécions

Dans les exemples suivants, les ponctuants constituent une borne à droite (en cela, à la manière des signes de

ponctuation, en particulier de ponctuation forte). Selon Morel & Danon-Boileau (1998 : 102), *hein* « participe à la construction d'une convergence de points de vue » :

(27) @X vaux mieux tard que jamais hein

En revanche, *quoi*, toujours selon Morel & Danon-Boileau (1998 : 102), « signifie à autrui qu'on énonce sa position à soi et qu'elle n'est pas soumise à discussion » :

(28) @X ta du manger trop vite lol
ne rien faire le soir du 31 c chaud quoi

On remarquera, en finale de la première phrase « syntaxique », outre la présence d'une interjection acronymique (voir *infra*, partie 5), celle d'un saut de ligne qui fait parfois office de démarcation dans les tweets (comme dans d'autres écrits non normés, tels les textes d'élèves).

4.3. De la sous-segmentation (et de la sur-segmentation) dans les tweets

Rappelons que, dans les écrits non normés, tels les textes d'élèves, la notion de ponctuation appelle en creux celle de sous-ponctuation (Béguelin 2000). Cette sous-ponctuation opère à un plus ou moins grand degré selon que les signes de ponctuation sont totalement absents ou bien seulement manquants à certains endroits où une borne devrait être matérialisée (notamment en fin de phrase « syntaxique »). De même, la notion de segmentation (qui permet de prendre en compte des bornes de diverses natures) peut être envisagée comme appelant en creux celle de sous-segmentation, dans des tweets « écrits au kilomètre » où des signes de ponctuation et/ou des connecteurs sont absents à certains emplacements où une borne matérialisée est obligatoire.

L'exemple (29) se caractérise par une sous-segmentation à un très grand degré de par l'absence totale de ponctuation et de connecteurs (*et* jouant un rôle de coordonnant) :

(29) excellent le projet final de maena et tout j'adore

On note dans l'exemple suivant, par ailleurs totalement dépourvu de connecteurs et de ponctuation (et notamment d'un point d'interrogation final), une majuscule à l'initiale :

(30) J'ai mixer une grippe avec une bronchite et une crise d'asthme vous voyez ca souvent vous

La segmentation n'est pas non plus totalement absente dans l'exemple suivant, un connecteur (*et*) ouvrant la troisième et dernière phrase « syntaxique » :

(31) trop heureuse pr miss Guyane elle est magnifique et elle a niqué en tte beauté languedoc la tchoin

En outre, il convient sans doute de relativiser l'absence fréquente de ponctuation finale, qui pourrait être rapportée à un « effet de genre », comme dans l'exemple suivant où aucune (autre) borne interphrastique ne manque :

(32) Je suis toujours en avance sur mon temps. Par exemple, j'ai pas d'enfant et je regrette déjà d'en avoir fait. ça evite les conneries parfois

Une réflexion sur ce que recouvre, ou non, la sous-segmentation dans les tweets paraît tout particulièrement à mener à la lumière d'exemples comme celui qui suit, sans ponctuation et sans connecteur(s), néanmoins borné à gauche et à droite par respectivement trois ligateurs énonciatifs et un ponctuant (de même qu'à la lumière d'exemples comportant des émoticônes et/ou des émojis ; voir *infra*, partie 5) :

(33) @X beh oui beh c pas lui qui va choisir pour moi hein

Par ailleurs, à l'inverse, certains tweets peuvent être considérés comme sur-segmentés. L'exemple suivant comporte

ce qui est parfois appelé un ajout après le point, ou encore un « complément différé », si l'on analyse *les arts martiaux* comme un complément d'objet direct « après coup » de l'élément central verbal *ont* (notons cependant que ce phénomène se rencontre dans les écrits littéraires) :

> (34) @X Ils ont un coté décadent qu'j'aime pas trop mais également une morale et des us et coutumes remarquables. Et les arts martiaux !

5. Les émoticônes et les émojis : des procédés de segmentation spécifiques dans les tweets (et autres écrits liés aux technologies numériques) ?

5.1. Quelques observations

Dans les tweets, comme dans des écrits d'autres genres liés aux technologies numériques, tels les SMS ou les courriels, le scripteur a la possibilité de recourir à des émoticônes et à des émojis. Ces signes iconiques sont des pictogrammes (au sens de Vaillant (2013)), visant la plupart du temps à indiquer l'émotion ou l'attitude énonciative du locuteur[5]. Un pictogramme se définit selon trois critères. D'abord, c'est un signe dont la saisie est iconique. Ensuite, c'est un signe qui appartient à un système d'écriture : il est compositionnel, constitué d'une combinaison d'unités iconiques minimales (Halté 2019), et il peut être placé (auprès de signes similaires, ou appartenant à d'autres systèmes sémiotiques) sur la chaîne syntagmatique et sur l'axe paradigmatique. Enfin, il faut que les pictogrammes, dans un système donné, aient la même taille, et soient dotés des mêmes couleurs, des mêmes formes, etc.

Nous appellerons « émoticône » tout pictogramme, constitué de signes du code ASCII[6], qui indique (au sens

5 La terminologie utilisée pour désigner ces pictogrammes n'est pas fixée, plusieurs termes étant souvent employés : émoticônes, emojis, smileys, binettes, etc. La terminologie que nous proposons ici relève de choix méthodologiques organisant notre recherche, mais n'a pas de valeur dogmatique. Par ailleurs, nous choisissons de franciser le mot japonais *emoji* en « émoji », que nous accorderons au masculin.

6 *American Standardized Code for Information Interchange*, le tout premier système d'encodage de signes pour un usage informatique : lettres, chiffres, signes de ponctuation, etc.

sémantico-pragmatique du terme, celui de la *deixis*) l'émotion du locuteur ou, plus généralement, son attitude énonciative. On peut lire les pictogrammes de cette sorte en penchant la tête vers la gauche (émoticônes « occidentales » : :-) :-(:-P pour, respectivement, un sourire, une mimique triste et un tirage de langue) ou de face (émoticônes « orientales » : ^_^ O_o pour, respectivement, un sourire et une mimique de surprise)[7].

Les cinq séquences d'émoticônes les plus fréquentes de notre corpus sont :

:) 0.009 par tweet
^^ 0.006 par tweet
;) 0.005 par tweet
:(0.004 par tweet
xD 0.004 par tweet

Nous appellerons « émoji » tout pictogramme dessiné accompagnant du texte dans les communications numériques, interagissant avec lui ou le remplaçant. Les émojis peuvent indiquer une émotion du locuteur, représenter un objet, une partie du corps ou un geste : 👊😑😟😖😏🚬...

Outre la différence formelle existant entre les deux signes, les émoticônes étant moins figuratives que les émojis, une autre différence est que ces derniers ne se réduisent pas à l'indication des émotions ou attitudes énonciatives du locuteur, contrairement aux émoticônes.

Les cinq séquences d'émojis les plus fréquentes sont :

😂 0.015 par tweet
😂😂 0.005 par tweet
😓 0.005 par tweet
😍 0.004 par tweet
😂😂😂 0.004 par tweet

7 Pour plus d'informations sur le sens des émoticônes et des émojis, nous renvoyons tout simplement à leurs pages *wikipedia* respectives. Nous n'avons pas la place, ici, d'aborder cette question.

Dans notre corpus, 11.42% des tweets se terminent par un émoji et 3.36% par une émoticône.

Les pictogrammes ont plusieurs fonctions dans les écrits numériques. Il est possible de les rapporter aux six fonctions discursives de Jakobson, comme le suggère par exemple Danesi (2013 : 100). Nous proposons de réduire ces fonctions à trois grandes catégories. Tout d'abord, les émojis, puisque certains d'entre eux servent simplement à représenter des objets, peuvent être utilisés pour remplacer un syntagme et ont alors des caractéristiques sémantiques et syntaxiques proches de celles d'un nom commun, d'un verbe, voire d'une proposition. Dans ce cas, ils ne servent pas à segmenter, puisqu'ils prennent tout simplement la place d'un élément syntaxique de la phrase, qu'il s'agisse d'un nom commun (ici pour le premier émoji) :

(35) Un grand 👋 @X vous avez grave assuré toute vos chorégraphies vous êtes génial 👈. 👐👋 pour votre super parcours #DALS ! 💋😳☺

Ou d'un verbe :

(36) (Okay je me lève malade, un mal de crâne pas possible, je bosse jusqu'à 18h (au lieu de 20h mdr) et j'organise une soirée chez moi. Je ☹.

Dans l'exemple (37), l'émoji se substitue même à une proposition :

(37) @X je t'aime bien mais 🖕

Nous laisserons ce cas de côté puisqu'il n'entre pas dans le cadre de la segmentation.

Une seconde fonction est « illustrative ». Dans ce cas, le pictogramme a aussi une fonction référentielle, mais il ne remplace pas un lexème. Il sert plutôt à illustrer, dans une relation de redondance sémantique (bien décrite par Klinkenberg (2009 : 26) concernant les rapports texte/image), une dénotation ou une connotation signifiée par la proposition :

(38) joue les pères noël et vous accompagne jusqu'au réveillon ! 🎄

Troisième et dernière fonction : la modalisation (Halté 2018 : 59). Le pictogramme, dans ce cas, indique l'émotion ou l'attitude énonciative du locuteur. Il porte alors une modalité (au sens de Gosselin (2010)) servant à valider ou invalider une proposition. Les modalités signifiées par ces pictogrammes sont le plus souvent appréciatives et/ou épistémiques. C'est la fonction la plus utilisée : comme mentionné *supra*, les cinq émojis et émoticônes les plus fréquents sont des icônes de mimiques faciales et indiquent donc une attitude ou une émotion du locuteur. Dans ce cas, le pictogramme employé vient modifier l'interprétation littérale d'une proposition, sur laquelle il fait porter l'attitude ou l'émotion du locuteur :

(39) **J'ai trop mangé chez ma pote, encore plus qu'hier** 😩😩...et elle m'a fait deux tuperware blindés pour chez moi (nous soulignons pour indiquer la portée)

Notons, enfin, que ces deux dernières fonctions peuvent occasionnellement s'hybrider. Ainsi, dans l'exemple suivant, le pictogramme représentant une tête d'extra-terrestre est à la fois illustratif (il est sémantiquement redondant avec le lexème *alien*) et modal (il indique l'émotion positive du locuteur, portant sur la proposition qui précède) :

(40) **Coupe vent Alien bien reçu** 👽 mais pq jai ps de stickers moi ? 😐 (nous soulignons)

Ces différentes fonctions conditionnent les lieux d'apparition des pictogrammes dans la chaîne syntaxique.

5.2. Où apparaissent les émoticônes et les émojis dans la chaîne syntaxique ?

Danesi (2013 : 84) note que « [...] les émojis sont placés à des endroits qui sont habituellement occupés par des catégories de mots spécifiques, des marques de ponctuation, ou des particules discursives » (notre traduction). Danesi (2013 : 105) va même plus loin en rapprochant le fonctionnement des émojis « joyeux » de celui des virgules et des points : « La mimique joyeuse fonctionne habituellement comme une virgule ou un point dans les messages hybrides, ajoutant une valeur émotive aux pauses » (notre traduction).

Les mêmes remarques sont formulées, concernant les émoticônes, chez Dresner & Herring (2010). En effet, les émoticônes et les émojis, qu'ils soient employés comme illustrations ou comme modalisateurs, sont susceptibles d'apparaître aux mêmes endroits de la chaîne syntaxique que la ponctuation et les connecteurs. C'est notamment parce que leur fonction de modalisateur, alliée à leur dimension iconique, les rapproche de la ponctuation modale. Dürrenmatt (2015 : 23-24) évoque en effet, parmi les autres « opérations proposées au lecteur par la ponctuation », celle de la « modalisation », « la modalité [étant définie comme] l'expression de l'attitude du locuteur par rapport au contenu propositionnel de son énoncé ». Ainsi, à plusieurs égards, émoticônes et émojis peuvent être envisagés comme complémentaires, à l'écrit, des signes de ponctuation et des connecteurs, ce qui peut expliquer en partie leurs possibilités de positionnement dans la chaîne syntaxique.

Ils sont susceptibles (par ordre croissant de fréquence d'usage) d'être :

antéposés (3% des occurrences) :

(41) 😍 Vacances à Bangkok à prix choc, pour une Saint Valentin loin d'être en toc !

en position interphrastique (apparition entre deux phrases « syntaxiques » au sein d'un même tour de parole, 19% des occurrences) :

(42) J'ai trop mangé chez ma pote, encore plus qu'hier 😩😩...et elle m'a fait deux tuperware blindés pour chez moi

postposés (78% des occurrences) :

(43) Restons modestes, juste de gros connards 😁😁 !!!

Notons aussi que les émoticônes et les émojis apparaissant en tout début de tour de parole constituent souvent, lorsqu'ils sont modaux, une réaction à un contenu énoncé précédemment, et sont alors susceptibles de relever d'une forme de modalisation dialogique (Halté 2018 : 61).

De par ces caractéristiques positionnelles, ces pictogrammes « ont le plus souvent par ailleurs un rôle de démarcation des tours de parole mais peuvent intervenir à tout moment à la manière des points d'exclamation » (Dürrenmatt 2015 : 105).

Le rapprochement avec la ponctuation permet peut-être d'expliquer l'une des caractéristiques des pictogrammes étudiés ici : leur portée sémantique est quasi systématiquement dirigée vers le co-texte gauche (voir Halté (2017 : 9)), comme c'est le cas pour la ponctuation modale. Cette portée explique aussi la grande fréquence d'apparition des pictogrammes en toute fin des tours de parole. Ils peuvent donc marquer une segmentation en y ajoutant une valeur modale s'exerçant (le plus souvent) sur leur co-texte gauche, permettant ainsi de circonscrire un segment de la chaîne syntaxique et de faire porter sur lui une modalité particulière. Comme dans le cas de la ponctuation, il est possible pour les scripteurs d'agencer les pictogrammes en séquences plus ou moins répétitives, dont la fonction peut être d'intensifier la modalité indiquée ou de la complexifier. Ces séquences occupent les mêmes positions dans la chaîne syntaxique que les occurrences « simples ».

Il est fréquent que les émoticônes ou les émojis soient utilisés en remplacement pur et simple de la ponctuation, que ce soit au sein d'un tour de parole ou à sa fin, mais toujours de manière interphrastique :

(44) @X Je retire ce que j'ai dis alors ;) Je vous ai répondu sur Facebook !

Ils viennent alors souvent en place de signes de ponctuation (le locuteur, dans cet exemple, met une majuscule après l'émoticône comme s'il s'agissait d'un signe de ponctuation). Leur présence permet de relativiser une sous-segmentation souvent envisagée, traditionnellement, uniquement en fonction de l'emploi de signes de ponctuation et de connecteurs.

Par ailleurs, les pictogrammes, les connecteurs et la ponctuation, apparaissant nécessairement aux mêmes endroits de la chaîne syntaxique, peuvent aussi se combiner pour former des ensembles constituant des systèmes modaux plus ou moins complexes. Lorsqu'un pictogramme et un signe de ponctuation sont adjacents, tous les agencements sont possibles, que ce soit au milieu d'un tour de parole ou à sa fin. Ainsi, le pictogramme peut être placé avant le signe de ponctuation, ou après, et ce, qu'il s'agisse d'émoticônes (ici, avant le signe de ponctuation) :

(45) Pas le temps d'allumer les bougies que c'est déjà revenu :) .

Ou d'émojis (ici, après les signes de ponctuation) :

(46) Donc c'est ça la culture « geek» fin 2016?? 😂😂😂😂 mais c'est désolant

Ou même de combinaisons entre signes de ponctuation, émoticônes et émojis :

(47) Mon allemand étant extrêmement limité tu me résumé cet article ^^ ?😮

S'il est possible, notamment avec le point d'interrogation, que le pictogramme soit situé après le signe de ponctuation (comme dans l'exemple précédent), la combinaison

entre pictogrammes, ponctuation et connecteurs apparaît habituellement selon l'agencement suivant (nous soulignons) :

(48) @X je sais 😟**, mais** je vais essayer promis, juré, crach.. Non pas cracher !

Enfin, évoquons la proximité (illustrée et argumentée dans Halté (2018 : 133)) sémiotique, sémantique et syntaxique de ces pictogrammes avec les interjections, notamment avec les interjections acronymiques spécifiques des corpus numériques comme *lol* ou *mdr*. La segmentation à droite peut aussi être assurée par ces interjections, comme *ptdrrrrr* dans l'exemple suivant :

(49) @X Au four ????? azy jme casse ptdrrrrr

Là encore, les pictogrammes, la ponctuation, les connecteurs et les interjections peuvent se combiner dans des configurations déterminément agencées.

Conclusion

Nous espérons avoir précisé les procédés de segmentation à l'œuvre dans les tweets, écrits souvent spontanés et plus ou moins normés, et où se déploie donc une palette très riche d'outils : les signes de ponctuation de l'écrit, les connecteurs utilisés parfois comme dans les écrits monogérés et parfois comme dans les oraux polygérés, les ligateurs énonciatifs et les ponctuants de l'oral, les émoticônes, les émojis et les interjections des écrits numériques. Face à cette diversité, il pourrait être proposé la notion de segmenteur (corollaire de celle de segmentation), certains segmenteurs, tels les connecteurs, les ligateurs énonciatifs, voire les majuscules de début, étant orientés à droite et d'autres, tels les signes de ponctuation (notamment forte), les émoticônes, les émojis et les interjections, étant tendanciellement orientés à gauche.

Par la suite, dans le prolongement de ce travail sur la segmentation dans les tweets, il serait intéressant de se pencher plus particulièrement sur les connecteurs, ou encore sur les

combinaisons (contiguës) de segmenteurs. Une autre perspective pourrait aussi consister à dégager des profils de scripteurs, par exemple en établissant, avec un éclairage quantitatif, des corrélations entre l'utilisation de la segmentation et le niveau orthographique.

Références bibliographiques

Abitbol J., Karsai M., Magué J.-P., Chevrot J.-P. & Fleury E. (2015). « Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis », *WWW '18 - World Wide Web Conference*, 1125-1134.

Anis J. (1983). « Vilisibilité du texte poétique », *Langue française* 59 : 88-102.

Béguelin M.-J. (dir.) (2000). *De la phrase aux énoncés : grammaire scolaire et descriptions linguistiques*. Bruxelles : De Boeck Duculot.

Bessonnat D. (1991). « Enseigner… la ponctuation ?(!) », *Pratiques* 70 : 9-45.

Bolander B. & Locher M. A. (2014). « Doing sociolinguistic research on computer-mediated data: a review of four methodological issues », *Discourse, Context and Media* 3(1) : 14-26.

Bouchard R. (2001). « *Alors, donc, mais*… "particules énonciatives" et/ou "connecteurs" ? Quelques considérations sur leur emploi et leur acquisition », *in* G. Ledegen & N. Rossi-Gensane (éds) *Les grammaires du français et les « mots-outils ». Syntaxe & Sémantique* 3 : 63-73.

Catach N. (1994). *La ponctuation*. Paris : Presses Universitaires de France.

Catach N. (2003). *L'orthographe française. L'orthographe en leçons : un traité théorique et pratique.* Paris : Nathan (1ère édition : 1995).

Condamines A. (2006). « *Avec* et l'expression de la méronymie : l'importance du genre textuel », *in* G. Kleiber, C.

Schnedecker & A. Theissen (éds) *La relation partie-tout*. Louvain-Paris : Peeters, 633-650.

Danesi M. (2016). *The Semiotics of Emoji*. London-New York : Bloomsbury Academic.

Dresner E. & Herring S. C. (2010). « Functions of the nonverbal in CMC: emoticons and illocutionary force », *Communication Theory 20*(3) : 249-268.

Drillon J. (1991). *Traité de la ponctuation française*. Paris : Gallimard, coll. TEL.

Dürrenmatt J. (2015). *La ponctuation en français*. Paris : Ophrys.

Feuillard C. (1989). *La syntaxe fonctionnelle dans le cadre des théories linguistiques contemporaines*, Thèse d'État, Université Paris V.

Gadet F. (2007). *La variation sociale en français*. Paris : Ophrys.

Gadet F. (2007-2008). « L'oral et l'écrit dans les changements technologiques et idéologiques », *in* E. Galazzi & C. Molinari (éds) *Les français en émergence*. Berne : Peter Lang, 131-142.

Gosselin L. (2010). *Les modalités en français. La validation des représentations*. Amsterdam-New York : Éditions Rodopi B. V.

Halté P. (2017). « Positionnement syntaxique des interjections et des émoticônes : modalisation, portée, visée », *Cahiers de praxématique* 69. Consulté à l'adresse https://journals.openedition.org/praxematique/4680

Halté P. (2018). *Les émoticônes et les interjections dans le tchat.* Limoges : Lambert Lucas.

Halté P. (2019). « Iconicité et signification modale : l'émoticône, de l'icône du corps au geste énonciatif », *MEI* 47. Consulté à l'adresse https://www.mei-info.com/en/revue/47/159/

Klinkenberg J.-M. (2009). « La relation texte-image. Essai de grammaire générale ». Consulté à l'adresse http://gemca.fltr.ucl.ac.be/docs/cahiers/20090128_Klinkenberg.pdf

Koch P. & Oesterreicher W. (2001). « Langage parlé et langage écrit », *in* G. Holtus, M. Metzeltin & C. Schmitt (éds) *Lexikon der Romanistischen Linguistik*, vol I,2. Tübingen : Max Niemeyer Verlag, 584-627.

Lodge A. (2011). « La question de la "langue commune" en français. Normes "sociales" *vs* normes "communautaires" », *in* S. Branca-Rosoff, J.-M. Fournier, Y. Grinshpun & A. Régent-Susini (éds) *Langue commune et changements de normes*, *Actes* du Colloque International *Langue commune et changements de normes*, Paris, février 2009. Paris : Champion, 77-92.

Morel M.-A. & Danon-Boileau L. (1998). *Grammaire de l'intonation. L'exemple du français oral*. Paris : Ophrys.

Paolacci V. & Rossi-Gensane N. (2012). « Quelles images de la phrase dans les écrits d'élèves de fin d'école primaire française ? Description linguistique et réponses didactiques aux difficultés des élèves », *in* F. Neveu, V. Muni Toke, P. Blumenthal, T. Klingler, P. Ligas, S. Prévost & S. Teston-Bonnard (éds) *Actes* du 3ème Congrès Mondial de Linguistique Française (CMLF), Lyon, juillet 2012, 341-359.

Popin J. (1998). *La ponctuation*. Paris : Nathan.

Roukhomovsky B. (2001). *Lire les formes brèves*. Paris : Nathan.

Sébillet T. (1548). *Art poétique françoys*. Consulté à l'adresse https://gallica.bnf.fr/ark:/12148/bpt6k50945f

Söll L. (1974). *Gesprochenes und geschriebenes Französisch*. Berlin : Schmidt.

Trésor de la Langue Française informatisé (TLFi), Centre National de Ressources Textuelles et Lexicales (CNRTL). Consulté à l'adresse https://www.cnrtl.fr/definition/

Vaillant P. (2013). « Sémiologie des pictogrammes », *texto !* 18(4). Consulté à l'adresse http://www.revue-texto.net/index.php?id=3336

Annexe

Liste des émoticônes (298) :

!!^^!!	!:(	!:-)	!:@	!=D	!>:(	#_#	(*_*)	(-_-)	(...!...)
(0_o)	(:	(:-(	(:>	(=^,..,^=)	(?_?)	(^-^)	(^^)	(^_^)	):
-	*.*	*0*	*3*	*^*	*^-^*	*^_^*	*^o^*	*_*	*_^
o	*p*	*~*	*•*	*○*	--'	-.-	-.-)	-_-	-_-'
/:O	0-o	0.o	0.○	0^0	0_0	0_O	0_o	3:)	8=)
:$	:'$	:'(	:')	:'-(	:'-)	:'-D	:'D	:'O	:']
:'o	:'p	:(	:)	:*	:*(	:*-(	:-$	:-(	:-)
:-/	:-@	:-D	:-O	:-P	:-S	:-[	:-]	:-o	:-p
:-s	:-{	:-|	:-}	:/	:0(	:0)	:3	:=(	:=)
:>>	:@	:D	:O	:O»	:O(	:O)	:P	:S	:[
:]	:^(	:^)	:d	:o	:o(	:o)	:o/	:oD	:op
:p	:poop:	:s	:|	:~)	;$	;'(	;')	;'D	;(
;()	;)	;-(	;-)	;-0)	;-D	;-]	;-o)	;-p	;-}
;0)	;3)	;=)	;@	;D	;O	;O)	;P	;S	;]
;d	;o	;o(	;o)	;p	;s	;~)	<'3	<3	<3'
<3*	<:-(	=$	='(	=')	='D	=(	=)	=-(	=-)
=-)	=-=	=.=	=/	=0	=3	=3	=3=	=D	=D
=O	=O)	=P	=[	=]	=^.^=	=^=	=^_^=	=_=	=o
=p	=~=	>.<	>.<'	>.>	>:(	>:-(	><	>=(	>?
>_<	>_<'	?(	?)	?:	?:-)	?_?	@-@	@.@	@^@
@_@	D:	D:<	D=	O.o	O_0	O_O	O_o	Oo	O‿O
T-T	T.T	T^T	T_T	ToT	T~T	XD	Xd	^*^	^*^
^-^	^.^	^0^	^3^	^=^	^O^	^^	^^»	^^'	^^'«
^^*	^^`	^^o	^_*	^_^	^_^»	^_^'	^_^)	^_^.	^o)
^o^	^p^	^~^	^•^	n_n	o.0	o.O	oO	o^o	o_0
o_O	o_o	u_u	xD	xd	|:(	~.~	~_~	~o~	¬.¬
¬_¬	¬‿¬	·-·	×3	è-é	è.é	è^é	è_è	è_é	é-è
é-é	é.è	é.é	é^è	é_è	é_é	•-•	•.•	∩.∩	‹3
•3•	•^•	•_•	◉_◉	◔_◔	◕_◕	⊙_⊙	＞_＜	・_・`	•~•
◡̈	•̀_•́	≧∇≦	◕‿◕	╤︿╤	◟˟◞˟	(・∀・)	｡•́︿•̀｡		

Liste des émojis (90) :

😍	😭	👌	😘	😏	👍	💕	👋	😄	😔
😑	✋	😊	😱	🙌	😬	🙏	😩	💪	💘
😉	😒	😢	💖	😡	😄	👊	😴	😞	😪
😌	😹	😜	💜	💚	💞	✨	😚	💙	💋
😤	😎	😋	😅	😕	😻	💗	😫	😀	😃
😐	😨	💓	😰	😥	😝	💛	😳	😓	👅
😆	😶	😈	😇	😛	😂	❤️	🙂	🙏	🎉
🙏	🌸	✅	👑	🤔	💔	👌🏻	🙈	🌹	❤️
☺️	😁	👍🏻	💪🏻	🔥	🤗	🙃	⭐	👌🏻	🙄

Detecting and categorising lexical innovations in a corpus of tweets

Psychology of Language and Communication 2022, Vol. 26, No. 1

sciendo

DOI: 10.2478/plc-2022-15

Louise Tarrade[1], Jean-Philippe Magué[1], Jean-Pierre Chevrot[2]
[1] ICAR laboratory (UMR 5191), École Normale Supérieure de Lyon, France
[2] LIDILEM laboratory (EA 609), Université Grenoble Alpes, France

Detecting and categorising lexical innovations in a corpus of tweets

In this paper, we present the methodology we have developed for the detection of lexical innovations, implemented here on a corpus of 650 million of French tweets covering a period from 2012 to 2019. Once detected, innovations are categorized as change or buzz according to whether their use has stabilized or dropped over time, and three phases of their dynamics are automatically identified. In order to validate our approach, we further analyse these dynamics by modelling the user network and characterising the speakers using these innovations via network variables. This allows us to propose preliminary observations on the role of individuals in the diffusion process of linguistic innovations which are in line with Milroy & Milroy's (1997) theories and encourage further investigations.

Key words: computational sociolinguistics, linguistic innovations, S-curve, Twitter, big data, network, diffusion of innovations

Address for correspondence: Louise Tarrade
ICAR laboratory (UMR 5191), École Normale Supérieure de Lyon, 15 parvis René Descartes, 69007, Lyon, France.
E-mail: louise.tarrade@ens-lyon.fr

The diffusion process of linguistic innovations has long been a topic of interest in sociolinguistics (Weinreich et al., 1968). Many studies have highlighted the influence of social structures on this process (Labov, 2010; Milroy & Milroy, 1997). The recent access to massive social network data and the advent of computational sociolinguistics (Nguyen et al., 2016) allow an approach to this phenomenon that combines a large amount of data and a fine-grained temporality. It is from this perspective that we present an approach for detecting and categorising linguistic innovations, more specifically lexical innovations. We relied on the idealized S-shaped trajectory of successful innovations (Blythe & Croft, 2012; Feltgen et al., 2017; Rogers, 2003) to identify lexical innovations in a corpus of French tweets and categorize them according to whether their use has stabilised (change) or not (buzz) over time. We then automatically detected the three phases of diffusion (Chambers, 2013; Fagyal et al., 2010) of these new forms: innovation, propagation, fixation for a change, or decline for a buzz.

In order to validate our approach for detecting lexical innovations, categorizing according to their fate, and identifying the successive phases of their dynamics, we tested the hypothesis that changes and buzzes spread differently across the network of users.

The first section presents relevant previous works which will allowed us to formulate our hypothesis more precisely. The second section presents in detail the data, and describes the method we propose to detect, categorizes, and delimit the diffusion phases of lexical innovations as well as our validation method. The third section presents the results obtained, which are discussed in the last section. All the codes used and the results obtained are available on our GitHub repository.[1] However, the corpus of tweets cannot be made available in order to respect the privacy of users.

Previous Work

A topic of interest in sociolinguistics for many years, linguistic change corresponds to the outcome of a process in several steps described by Weinreich et al. (1968). First, a speaker introduces a new form in their use of language, then this form is taken up and used by other speakers, and finally, the use of this form stabilizes in a community. We can consider this to be a change. These phases in the establishment of a linguistic innovation as a change are variously named *innovation*, *propagation*, and *fixation* by Fagyal et al. (2010) or initial stasis, rapid rise, and tailing off by Chambers (2013). The idealised S-shaped trajectory observed by Rogers (2003), confirmed at the linguistic level by Blythe and Croft (2012), and later validated on a large scale by Feltgen et al. (2017), accounts for these phases in the case of successful innovations. The role of social structures and individuals in this process of diffusion has also been discussed, and

1 https://github.com/LTarrade/lexical_innovation_detection

sociolinguistics has tried to identify the innovators in the process of diffusion of a linguistic innovation, as well as how their position in the network influences this process. Labov (2010) describes leaders of linguistic change as generally female, middle-class individuals who are both very central to their local community and have a large number of connections outside of it. Milroy and Milroy (1997) observed phonological variation in three different neighbourhoods of the city of Belfast and noted that people who were very central to their neighbourhoods were also very conservative with respect to vernacular norms, but also, that innovations were mainly introduced by young women who worked in shops where the different communities met, and who were therefore in regular contact with them, but had no strong ties to them. For Milroy and Milroy (1997), the innovators are therefore people with weak ties, on the periphery of the communities. They are the ones who bring the different variants into the community, but for that variant to be adopted by a community, it is necessary that it is first adopted by people with strong ties and who are very central to the community. Earlier, Granovetter (1973) had also highlighted the importance of weak ties in the transmission of innovations. However, these studies were most often conducted on phonetic changes, on populations of hundreds of individuals at most.

Over the past decade, access to massive digital data has allowed the emergence of computational sociolinguistics, which approaches the issues of sociolinguistics by combining the methodologies of natural language processing and data science (Nguyen et al., 2016). In addition to the methodological renewal it has brought about, the interest of computational sociolinguistics also lies in the nature of the data it uses. Often derived from social media, these data document language varieties that are not very standardized, showing a high variability and a high rate of innovation. Computational sociolinguistics has thus been able to address the issue of language change and some of its theories have been partly transposed into social media studies. Thus, Del Tredici and Fernández (2018) highlighted in the observation of the diffusion of linguistic innovations within Reddit communities that the innovators seem to correspond to the hubs of the community, that is, individuals with many but weak connections. They also demonstrated that the adoption of an innovation by the community seems to be conditioned by the fact that it is adopted by members with strong ties. Moreover, Laitinen et al. (2020), using a corpus of tweets, underlined the importance of the size of the network and showed that, above a certain size, the fact that networks are mostly composed of weak (more propitious to innovation) or strong ties (more conservative and resistant to change) no longer seems to constitute a significant distinction in the resistance to change. Fagyal et al. (2010) used multi-agent simulations to study the role of individuals in the diffusion of a linguistic innovation and its adoption as a norm. Using a number of tests carried out by varying the parameters of their network, they highlighted that, in general, leaders (hubs) push forward the change in progress and are indispensable for establishing it as a norm, and loners are the repositories of old or new variants

and their absence from the network results in a lack of innovation. One of the most promising avenues of research in the field of linguistic variation and change on social media is the consideration of psychological and social factors through the relationship between personality and social network. The structure and size of the communities of contacts established by Facebook users depend on their degree of extraversion (Friggeri et al., 2012), a trait that is also associated with a greater general tendency to innovate (Ali, 2019), which could also manifest itself in the field of language. While these studies partly addressed the way in which social structure influences the process of linguistic diffusion, they often remain confined to one aspect and do not provide a comprehensive view of this diffusion process in action at different population levels.

However, before being able to study this mechanism, one must first identify the linguistic innovations to be studied, and the very noisy nature of social media corpora does not facilitate this task. The methods used for the detection of linguistic innovations often focus on the detection of semantic changes, in particular because of the possibilities offered by the first word embedding techniques allowing to represent words in vector spaces with, for example, the word2vec algorithm proposed by Mikolov et al. (2013) and in particular since the appearance of language models based on deep learning such as BERT (Devlin et al., 2019) and its successors. For our part, we were more interested in lexical innovations, that is, the appearance of new words. The majority of studies on lexical changes in computer mediated corpora rely on frequency analysis to detect them. Among them, Eisenstein et al. (2014), in their work on linguistic diffusion networks on Twitter between different metropolitan areas in the USA, selected the words used in their analysis based on the 100,000 most frequent terms. After imposing a minimum frequency of use, they calculated the variance of the algorithmic probability of each of them to select only the words with a variance above a certain threshold. From the 5,000 words obtained, they manually eliminated both named entities and non-English words, and determined for each remaining word whether its usage is similar to that of an English word based on examples of word usage in context. Meanwhile, Costin-Gabriel and Rebedea (2014) used word evolution images provided by the Google Books N-gram Viewer and principal component analysis techniques to find the general patterns of three types of words: common words, neologisms, and archaisms, which they listed manually. They calculated the proximity of the evolution of the word to be classified to each of the three trends to determine which type of word it was. Tjong Kim Sang (2016) tested two methods of detecting neologisms and archaisms in a corpus of magazine texts as well as a corpus of tweets, one calculating a score from the ratio of the starting frequency to the ending frequency, and another, less effective on tweets, whose score depended on the correlation coefficients between word frequencies and time. At the same time, Kershaw et al. (2016) relied on two methods originally intended for lexicographers to measure the acceptance of linguistic innovations. They implemented three statistical

tests based on the variations of the frequency, meaning, and morphology of the forms, which they applied to a corpus of tweets and a corpus of Reddit posts. To measure the importance of tie strength in the adoption of an innovation within communities, Del Tredici and Fernández (2018) used an already existing online lexicon of slang terms from the internet to identify these innovations. They then categorized innovations as either successful or unsuccessful based on the slope of diffusion of each term. Stewart and Eisenstein (2018) highlighted the importance of linguistic rather than social diffusion in the maintenance or decline of nonstandard words, and showed that one of the factors determining their stabilisation is their membership to a wider variety of lexical contexts. To do so using a Reddit corpus, they used the frequency of words over time to identify the words with increasing frequency using the Spearman coefficient and the words with decreasing frequency by fitting the frequency series of words with a two-phase piecewise linear regression and with a logistic distribution for the more discrete growth and decay trajectories. On the other hand, Kerremans and Prokić (2018), chose a semi-automatic detection of neologisms on the web using correspondence dictionaries.

The approaches for detecting lexical innovations mentioned above almost systematically using English corpora, only partially respond to our needs. Often, they either require manual steps that are quite time-consuming, they require the use of dictionaries, or they seem to focus more on innovations in the growth phase rather than on innovations that have stabilised, and do not seem to seek to identify the three phases of diffusion of innovations. Furthermore, as with the detection of semantic changes (Schlechtweg et al., 2019), it is difficult to assess the performance of these methods given the diversity of the datasets and the lack of evaluation of this task.

Methodology

Data Presentation

Our data was collected in two steps. The first corpus of tweets was first collected as described in (Abitbol et al., 2018), spanning from June 2014 to March 2018. Afterwards, the second collection of tweets was carried out, which consisted in updating the first corpus by taking each of its users and retrieving their last tweets using the Twitter API. The tweets thus obtained were filtered according to the language detected by Twitter (French) and the client used (in order to filter out robots). In the end, the cleaned corpus data included about 650 million tweets in French from just over 2.5 million users, covering a period from 2007 to February 2019, with 98% of the corpus concentrated between 2012 and 2019. For each tweet, we gathered a set of metadata such as the creation date or the identifier as well as information about the user who produced the tweet such as the identifier, the number

of followers or the number of followees,[2] that is, the other accounts they follow.

However, for about 20,000 accounts, this information could not be retrieved for various reasons, as in the case when the Twitter account has been deleted in the meantime. This collection allowed us to reconstruct the network of the corpus users by modelling it as a directed graph composed of nodes (users) that can include incoming ties (followers) and outgoing ones (followees). This network is thus a directed, static, and closed graph.

Detection and Categorization of Lexical Innovations

From the tokenized tweets, we retrieved all of the new tokens in the corpus as follows. Since we were interested in the dynamics of innovation after their appearance, we selected the tokens that appeared between March 2012 and February 2014 (which were totally absent from the corpus for the whole of the previous year) in order to have at least a five-year period to observe the evolution of their use. We filtered these forms by keeping only those that have been used by at least 200 different users, and using regular expressions, we excluded hashtags, emojis, or punctuation marks in order to focus the rest of our analysis exclusively on words, which were then defined as any sequence of alphanumeric characters that may contain an apostrophe or a dash.

For each of the new forms thus recovered, we retrieved their usage rate each month over a period of five years. The usage rate of a form in a given month corresponds to the ratio between the number of people who used this form that month and the number of people who tweeted in the month. Therefore, for each emerging form, and for each linguistic innovation, we obtained the trajectory of its use among the users of our corpus during its first five years of use.

In order to categorise each innovation as either a change (an innovation whose usage rate stabilised after experiencing exponential growth) or a buzz (an innovation that also experienced a phase of exponential growth, but whose usage eventually declined to a very low rate), we used a curve fitting method. To reduce the influence of accidental peaks in the trajectory of the usage rate of each observed form, we considered the rolling average of this rate with a three-month window. Then, using the LMFIT[3] library for Python, we fit the usage trajectories of each form to two reference functions:

- The logistic function (the S-shaped curve followed by the changes), defined as $f(x, A, \mu, \sigma) = A[1 - \frac{1}{1+e^{\alpha}}]$, where $\alpha = (x - \mu)/\sigma$ and where A is the amplitude, μ is the center, and σ is the sigma parameter (which influences the steepness of the inclination of the curve slope).
- The lognormal function (a skewed bell-shaped curve followed by the buzzes), which is defined by $f(x; A, \mu, \sigma) = \frac{A}{\sigma\sqrt{2\pi}} \frac{e^{-(\ln(x)-\mu)^2/2\sigma^2}}{x}$, where A is the amplitude, μ is the center, and σ is the sigma, that is, the characteristic width of the peak.

2 https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet
3 https://lmfit.github.io/lmfit-py/index.html

Forms for which the reduced chi-square (a measure of quality of fit) was above a certain threshold (0.00005) for both functions were ruled out. Otherwise, the best of the two fits determined whether the form was a change or a buzz. In order to observe the trajectory of innovations in their entirety, we also sorted out the values of some output parameters of the fit.[4]

As the latter influenced each other, the choice of the limit attributed to their value was decided following observations of the impact of each of these parameters on the others. Following this selection, we ended up with just over 500 words whose use seems to follow a logistic or lognormal trajectory very closely and whose trajectory is almost entirely observable over the five-year period.

As mentioned in the Introduction section, a successful language innovation follows three phases of diffusion. During the first phase, the innovation phase, the form spreads only very slowly through the population. In the second phase, propagation, the form spreads among users exponentially until it reaches a threshold at which the use of the form stops growing but remains high, which is the fixation phase. When its use stabilizes, a change can be considered to have occurred. If an innovation fails to stabilise and instead experiences a third phase of decline, we consider it to be a buzz.

To delimit the three phases of diffusion of innovations, we searched for the maximums of the third derivative, that is, the moments when the acceleration in the spread of the form varied the most. The first maximum marks the boundary between the innovation and the propagation phases, the second marks the boundary between the propagation and the fixation or decline phases.

Validation of the Method for Detecting and Classifying Linguistic Innovations

In order to validate our approach of detecting lexical innovations and our classification as buzzes or changes, we hypothesized that they are distinguished from each other by the position in the network of the users who adopted them at different phases of diffusion. Based on what is reported in the literature, in particular the theory on the process of diffusion of an innovation within a language community proposed by Milroy and Milroy (1997), lexical innovations should be introduced by people who are rather on the periphery and in contact with several different social groups. Similarly, they should only start to stabilize after being adopted by people who are very central to the community, who, transposed on our corpus of tweets, could be identified as people with many incoming ties and, therefore, with a certain prestige. To confirm this, we characterized each user of our corpus by network variables and looked at the distribution of these variables for each of the phases defined above.

4 Sorting on the output parameters of the logistic curve fit: (((center>=16) & (center<=31) & (sigma<=8)) | ((center>31) & (center<=46) & (sigma<=7))) & (redchi<0.00005) & (amplitude>0.02) & (center_err<5); Sorting on the output parameters of the lognormal curve fit: (fwhm>=4) & (fwhm<=40) & (redchi<=0.00005) & (amplitude<=1.1) & (maxPoint>=21) & (maxPoint<=46) & (((center<=3.6) & (sigma<=0.65)) | ((center>3.6) & (center<=3.8) & (sigma<=0.35)) | ((center>3.8) & (sigma<=0.15)))

To extract network variables, we relied on the user network, modelled as described in the Data Presentation section. For each user, we calculated a PageRank score (Brin & Page, 1998) using the SNAP network analysis library (Leskovec & Sosič, 2016). The PageRank score corresponds to a measure of user prestige. More concretely, it is calculated from the frequency of visits of each node (user) by a random walk. Therefore, the score will be influenced both by the number of incoming ties (followers) of each node, but also by the respective prestige of the incoming nodes. Thus, the higher the score, the more prestige a node has. In the same way, we characterized each node by its local clustering coefficient. The clustering coefficient of a node is calculated by considering the graph as an undirected graph (all links between users are considered, regardless of whether they are incoming or outgoing) and by calculating the number of effective triads out of the number of possible triads for each node, thus looking at the proportion in which the neighbors of a node are connected to each other. Therefore, this measure is an indicator of the openness of each user network. The higher the clustering coefficient of a user, the more closed their own network is, so the more their friends are also friends with each other.

Our goal was to understand how and when these variables affect the process of acceptance or nonacceptance of innovations. For each form and for each user using that form, we recorded the phase when they used the form for the first time. Thus, for each type of innovation (change or buzz), we recovered all the users who adopted a form of this type at each phase (innovation, propagation, and fixation/decline). We then compared the distribution of the variables across the type of innovation and the phases. These six distributions were compared to the distribution in the whole population as well. A user may appear twice since they may have used one form in one phase and another form during another phase. This nonindependence constrained the statistical tests used to compare the distributions. However, only 18% of the users have used at least one buzz and one change. The great majority of users used either only buzzes or only changes, and on average, they used two different forms, the median being one form.

Since the distributions of the variables were not normal, we based our comparisons on the median and the first and third quartiles rather than on the means. We also used the nonparametric Kruskal-Wallis and Wilcoxon-Mann-Whitney statistical tests to ensure the significance of our observations. Finally, as the number of individuals in the observed samples varied considerably, we also ensured that this parameter did not influence our results by using a bootstrapping technique. More precisely, for each observed sample (e.g., the distribution of the number of incoming ties of users who used a buzz for the first time during the period of innovation of the form), we carried out 1,000 random samples without replacement, of the same number of individuals as in the observed sample, and then ensured that the median of this sample did not lie within the 95% confidence interval of the medians observed on the 1,000 random samples.

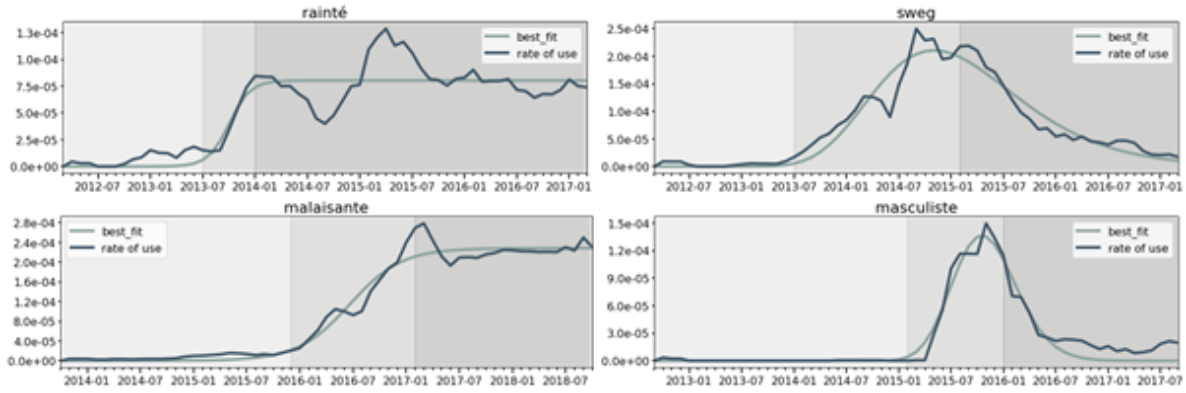
Results

Identifying Changes and Buzzes

The method of detection and categorization according to the trajectory of use of new linguistic forms over the months described in the previous section allowed us to identify a set of lexical innovations, and categorize them as changes (C) or buzzes (B). Figure 1 shows examples of linguistic forms identified as changes (left) or as buzzes (right) and for which the monthly rate of use over five years follows either a logistic function or a lognormal function, respectively. However, note that while the method correctly recognized the different types of curves we were looking to identify, the usage trajectory of the words was never ideally fitted with the reference curve due to the noisy nature of the data.

Focusing on the changes and buzzes identified by our method, we can observe a large number of neologisms, most of them linked to new realities (fullstack (C), émoji (C), mutuals (C), snapé (C), streameur (C)) or practices (twerké (C), dabé (C), binge-watching (C)), as well as to social phenomena (islamogauchiste (C), féminazis (C), masculiste (B), agenre (C)). We can also find archaic reactualized linguistic forms (malaisante (C)), new linguistic forms (enlové (B), sweg (B), baé (C), ggwp (C), oklm (C)) with a large number of morphological derivations (cuissance (B), mignonance (B), coulance (C), génance (C)), but also borrowings (mskn (C), mutuals (C), kehba (B), sadlife (B)) or slang words (bresom (C), rainté (C), kecho (B), peufra (B)). Many variations of new linguistic forms were also present, especially in the buzzes (miskinou (B), oklmus (B), tchuips (B)), but also phonological variations (chumor (C), aoé (C), aeq (B), caley (B)), agglutinations (heinquoi (C), balecouilles (B)), abbreviations (batrd (B), qtv (C)), or simple spelling variations (parasyte (C), embiancer (B)). The buzzes also included a very large number of lengthenings (oklmm (B), mddddddrr (B),

Figure 1. *The Usage Rate per Month of Two Changes (Left) and Two Buzzes (Right) Represented by a Rolling Average with a Three-Month Window (Blue), as well as the Result of the Curve Fitting (Green). The Three Diffusion Phases are Represented by the Grey Shading in the Background.*



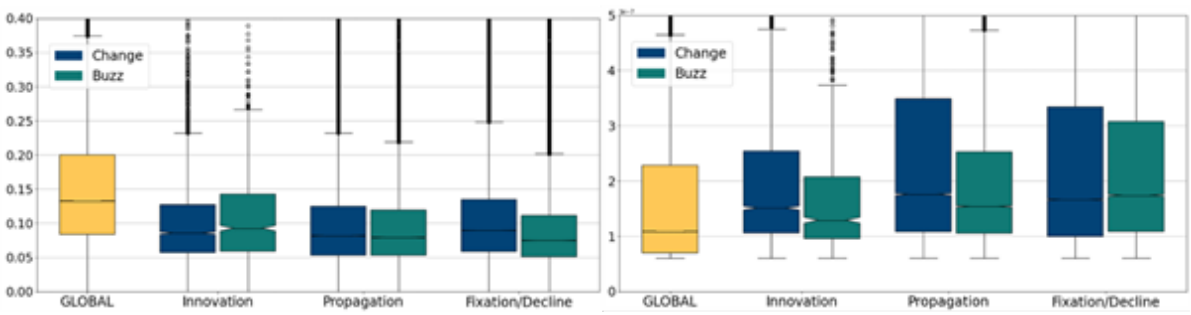
flmmm (C)). Finally, despite an initial automatic filtering of proper nouns, there remained a large number of named entities (just over 200 in the two categories of innovations combined), which we set aside after manual filtering. To conclude, we have identified 141 changes and 251 buzzes.

Analysis of Network Variables

A first comparison of the distribution of the two network variables of the buzz and change users at the different phases with the distribution in the global population allowed us to perceive some interesting phenomena. First, the users who used buzzes or changes were both clearly different from the overall population, even if not to the same degree and not always following the same dynamics. Second, the comparison of the network variables was in line with our hypothesis, especially for the changes.

The left panel of Figure 2 shows the distributions of the clustering coefficient of the overall population and of the users of the buzzes and changes during the different diffusion phases. As a reminder, the clustering coefficient of a user is an indicator between 0 and 1 of how much their friends are also friends with each other. The higher this coefficient, the more the user is part of a closed network, the less new information has the possibility to reach this subnetwork. The clustering coefficient distributions of the users of changes and buzzes in the innovation phase had much lower values than those of the overall population (Mann-Whitney = 3.9e+10; $p < .01$ (C); Mann-Whitney = 1e+10, $p < .01$ (B)), which would suggest that the innovations appear in both cases in more open networks than normal, in which new information can more easily enter. In the propagation phase, if buzzes and changes follow the same dynamics with a clustering coefficient that continues to decrease, buzzes experience a much stronger decrease (median at -1.4e-02 compared to the innovation phase; Mann-Whitney = 1.6e+08, $p < .01$) than changes (median at -3.9e-03 compared to the innovation phase; Mann-Whitney = 1e+09, $p < .01$). Finally, users who adopted changes during their last diffusion

Figure 2. *Distributions of the Clustering Coefficients (Left) and PageRank Scores (Right) for the Whole Corpus in Yellow and for the Users of the Changes (Blue) and Buzzes (Green) at the Three Different Phases of Diffusion (from Left to Right: Innovation, Propagation, Fixation/Decline).*



phase (fixation) had a clustering coefficient that rose and tended to get closer to that of the global population, with a median at 8.9e-02 and, above all, a greater dispersion of the data towards higher clustering coefficient values (thus, a more closed network), contrary to those who adopted buzzes during the corresponding phase, with a median at 7.5e-02 (Mann-Whitney = 4.6e+09, $p < .01$).

The right panel of Figure 2 shows the distributions of the PageRank score, an indicator of user prestige, influenced both by the number of incoming degrees of a user and by the prestige of these incoming degrees. The higher the PageRank score, the more prestige the user has. To illustrate this, the Twitter account in our corpus with the highest PageRank score (6.55e-04) was the official Twitter account of the French newspaper Le Monde. Here too, the dynamics between the different phases of diffusion of changes and buzzes differed. While users of the two categories of innovations (changes and buzzes) who adopted the innovation during its first phase of diffusion had a much higher distribution of their PageRank score than that of the overall population (Mann-Whitney = 4.6e+09, $p < .01$ (C); Mann-Whitney = 1.2e+09, $p < .01$ (B)), that of users of changes was even higher (with a median of 1.5e-07, i.e., 1.17 times the median of buzzes, Mann-Whitney = 2.4e+06, $p < .01$, in the same phase and 1.39 times that of the overall population). In the propagation phase, the distribution of PageRank scores for changes continued to expand strongly towards higher values (median 1.62 times that of the overall population and third quartile at 3.48e-07, i.e., 1.53 times higher than that of the overall population; Mann-Whitney = 4.2e+10, $p < .01$) and that of buzzes followed the same dynamic, but with less expansion (median 1.42 times that of the overall population but with a third quartile at 2.5e-07, i.e., only 1.11 times higher than that of the overall population; Mann-Whitney = 3.1e+10, $p < .01$). Users who used a change for the first time during the fixation phase had lower median PageRank scores than during the previous phases (Mann-Whitney = 4.1e+09, $p < .01$), which would suggest the beginning of a decline in the distribution, while those who used a buzz for the first time during its decline period had higher median PageRank scores (Mann-Whitney = 9e+08, $p < .01$), almost reaching the values of that of the change propagation phase (1.73e-07). These observations suggest that if innovations are initially employed by more prestigious users than the average in both cases, those who adopt the changes in the propagation phase are even more prestigious, and one might suppose that they thus favor their diffusion. Also interesting was the decrease in the PageRank score during the last phase of diffusion of changes, with the opposite phenomenon in the phase of decline of buzzes. This could be interpreted as an indicator of the introduction of the innovation to a wider part of the population, with users with a lower prestige value who would in turn adopt the innovation, unlike the buzz.

The overall observation of the distributions of the two observed network variables, that is, the clustering coefficient, a measure of network openness, and the PageRank score, a measure of prestige, is in line with previous sociolinguistics studies. Indeed, the decrease of the clustering coefficient during the first two

phases of diffusion and its slight increase during the last phase suggests that the individuals described by Milroy and Milroy (1997) and Granovetter (1973) with weaker ties who are at the intersection of several communities (and therefore with less closed networks) are also at work here, before the innovation is appropriated by people with tighter networks. Similarly, the inverse dynamic observed during the diffusion of change phases with the PageRank score is in line with Milroy and Milroy (1997), for whom an innovation must first be adopted by the central (i.e., more prestigious) members of their community before being passed on to the rest of the community.

Discussion and Future Perspectives

We have presented a method for automatically detecting lexical innovations based on the trajectory of their rate of use in a corpus of Tweets over a period of five years, categorized them as changes or buzzes according to this same criterion, and then determined for each of them their three phases of diffusion: innovation, propagation, and diffusion for changes or decline for buzzes. This method is semiautomatic, in the sense that it required two manual interventions: one to restrict the values of the parameters of the curves fitted during the automatic categorization of the innovations, the other for a final filtering of the words obtained in order to remove the named entities. Although these two steps need to be improved to avoid any manual recourse afterwards, it should be noted that they are not extremely costly in terms of time.

Using two network variables reflecting, respectively, the degree of network openness of each user and their level of prestige, we validated the relevance of this detection method by showing that the distributions of these variables for the users of the two classes of lexical innovations both differed clearly from that of the overall population of the corpus, and that the evolution of the distribution of these variables, in particular for the changes, was in line with what had been observed in earlier sociolinguistic work. Nevertheless, from a more distant perspective, the rather similar dynamics observed in particular at the level of the first two phases of buzz and change diffusion suggests that what determines the adoption or not of a lexical innovation could also be located at other levels, which should be explored in parallel, in particular at the level of user communities. We can legitimately question the possibility whether the diffusion of innovations (whether buzzes or changes) that we observed is not really the diffusion of innovations in the overall population of our corpus of tweets, but rather the diffusion of these innovations within communities. In order to know whether these innovations are accepted within the communities in which they were created or whether have spread outside of them, we believe it is essential to model the different communities in our user network. Therefore, the next immediate step of our work will be to detect these communities in two ways: on the one hand, with community detection algorithms such as the Louvain algorithm (Blondel et al., 2008), and on the other, by finding communities of interest in a more traditional way, for example, by exploiting hashtags.

In order to continue to explore the importance of social structure in the diffusion and acceptance of innovations, it would be interesting to broaden the field of variables to be observed in order to characterize users more completely. Thus, it would be interesting to complete the network variables characterizing each user, in particular by calculating centrality scores such as the betweenness centrality, which gives an indication of the extent to which the observed node is a passage point for the other nodes of the graph, the centrality of proximity, or the Katz centrality, which would allow us to obtain an exposure index for each user (the higher the index, the more the user is exposed to the information). Similarly, we can consider including linguistic variables such as the number of tweets, measures of lexical diversity, and so forth. Another short-term objective would be to complete this characterisation with social variables, such as age or gender, which could be inferred using machine learning algorithms as has already been done, for example, by Wang et al. (2019) for gender, age, and organization status, Bamman et al. (2014) for gender, or Flekova et al. (2016) for age and income. Thus, we can also question the strength of the influence of other variables in the acceptance of a linguistic innovation. For example, does it depend more strongly on the variety of linguistic contexts in which it is used, as suggested by Stewart and Eisenstein (2018), or is the duration of the innovation or propagation phases decisive in the acceptance process?

Finally, we also plan in the near future to complete our study by including the detection and classification of semantic innovations, but also by addressing the diffusion of lexical change through a more qualitative approach, extracting the different contexts of use of a number of innovations at different periods of their diffusion to analyze the evolution of their usage and lexical contexts.

Conclusion

From a corpus of a hundred million tweets in French, produced by more than 2.5 million users and spanning several years, we automatically identified the lexical innovations present in the corpus and categorized them in the same way as changes (innovations whose use stabilised in the corpus over time) or buzzes (innovations whose use, after a period of growth, declined). We used the speed of diffusion of the rate of use of each form to determine the three characteristic phases of the diffusion of these linguistic innovations. We also modeled the network of users in the corpus and characterized each of these speakers with network variables indicating the level of openness of their own network and their level of prestige. The first observations of the distribution of these user variables at the different phases of buzz and change diffusion validated the efficiency of our method by bringing out, in particular for the changes, a diffusion dynamic described in the literature (Milroy and Milroy, 1997). These results encourage us to continue in this direction and to explore new directions to study the influence of social structure on the process of diffusion of linguistic innovation.

Conflict of Interest Disclosure

The authors do not have any conflicts of interest to report.

Funding

Funding source: LabEx ASLAN (ANR-10-LABX-0081) of the Université de Lyon.

Research Ethics Statement

The whole project including design, data collection and processing, data handling, storing and sharing, privacy protection were screened and approved by the ethics committee of INRIA (National Institute for Research in Digital Science and Technology) (favorable opinion, reference 2017-005, IRB00013144).

Authorship Details

Louise Tarrade: research concept and design, collection and/or assembly of data, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article.

Jean-Philippe Magué: research concept and design, collection and/or assembly of data, data analysis and interpretation, critical revision of the article, final approval of the article.

Jean-Pierre Chevrot: research concept and design, collection and/or assembly of data, data analysis and interpretation, critical revision of the article, final approval of the article.

References

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2018). Socioeconomic dependencies of linguistic patterns in Twitter: A multivariate analysis. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 1125–1134). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186011

Ali, I. (2019). Personality traits, individual innovativeness and satisfaction with life. *Journal of Innovation & Knowledge, 4*(1), 38–46. https://doi.org/10.1016/j.jik.2017.11.002

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160. https://doi.org/10.1111/josl.12080

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language, 88*(2), 269–304.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Chambers, J. K. (2013). Patterns of variation including change. In J. K. Chambers & N. Schilling (Eds.), *The handbook of language variation and change* (pp. 129–297). Wiley Blackwell.

Costin-Gabriel, C., & Rebedea, T. E. (2014). Archaisms and neologisms identification in texts. In: *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference* (pp. 1–6). IEEE. https://doi.org/10.1109/RoEduNet-RENAM.2014.6955312

Del Tredici, M., & Fernández, R. (2018). The road to success: Assessing the fate of linguistic innovations in online communities. *ArXiv:1806.05838* [cs.CL]. https://doi.org/10.48550/arXiv.1806.05838

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv:1810.04805* [cs.CL]. https://doi.org/10.48550/arXiv.1810.04805

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS One, 9*(11), e113114. https://doi.org/10.1371/journal.pone.0113114

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua, 120*(8), 2061–2079. https://doi.org/10.1016/j.lingua.2010.02.001

Feltgen, Q., Fagard, B., & Nadal, J.-P. (2017). Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language

change. *Royal Society Open Science, 4*(11), 170830. https://doi.org/10.1098/rsos.170830

Flekova, L., Preoţiuc-Pietro, D., & Ungar, L. (2016). Exploring stylistic variation with age and income on Twitter. In: K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 313–319). Association for Computational Linguistics

Friggeri, A., Lambiotte, R., Kosinski, M., & Fleury, E. (2012). Psychological aspects of social communities. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 195–202). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.104

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360–1380. https://doi.org/10.1086/225469

Kerremans, D., & Prokić, J. (2018). Mining the web for new words: Semi-automatic neologism identification with the NeoCrawler. *Anglia, 136*(2), 239–268. https://doi.org/10.1515/ang-2018-0032

Kershaw, D., Rowe, M., & Stacey, P. (2016). Towards modelling language innovation acceptance in online social networks. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 553–562). Association for Computing Machinery.

Labov, W. (2010). *Principles of linguistic change. 2: Social factors*. Wiley-Blackwell.

Laitinen, M., Fatemi, M., & Lundberg, J. (2020). Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence, 3*, 46. https://doi.org/10.3389/frai.2020.00046

Leskovec, J., & Sosič, R. (2016). SNAP: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology, 8*(1), 1–20. https://doi.org/10.1145/2898361

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Curran Associates, Inc.

Milroy, J., & Milroy, L. (1997). Network structure and linguistic change. In: N. Coupland & A. Jaworski (Eds.), *Sociolinguistics* (pp. 199–211). Springer.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics, 42*(3), 537-593. https://doi.org/10.1162/COLI_a_00258

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed). Free Press.

Schlechtweg, D., Hätty, A., Del Tredici, M., & im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. ArXiv, 1906.02979 [cs.CL]. https://doi.org/10.48550/arXiv.1906.02979

Stewart, I., & Eisenstein, J. (2018). Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. *ArXiv:1709.00345* [cs.CL]. https://doi.org/10.48550/arXiv.1709.00345

Tjong Kim Sang, E. (2016). Finding rising and falling words. In: E. Hinrichs, M. Hinrichs, & T. Trippel (Eds.), *Proceedings of the workshop on language technology resources and tools for digital humanities* (LT4DH) (pp. 2–9). The COLING 2016 Organizing Committee

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. In: L. Loiu & R. White (Eds.), *WWW' 19: The World Wide Web Conference* (pp. 2056–2067). https://doi.org/10.1145/3308558.3313684

Weinreich, U., Labov, W., & Herzog, M. (1968). *Empirical foundations for a theory of language change (Vol. 58)*. University of Texas Press Austin.