

Multi-agent models and social media data: Collective dynamics and individual trajectories in linguistic populations (MACDIT)

Résumé : Le changement et la variation sont des propriétés fondamentales du langage. S'il est admis que la structure des interactions sociales influence ces propriétés, nous sommes loin de comprendre toute la complexité de ce phénomène et sa dynamique. Nous utiliserons une "double voie" pour approfondir cette compréhension : (1) nous modéliserons la structure et la dynamique des interactions en utilisant des réseaux multi-agents divers quant à leur propriétés internes et aux caractéristiques des "locuteurs" (2) nous intégrerons des "données du monde réel" en utilisant les messages twitter du corpus SoSweet et des échanges en ligne impliqués dans la construction d'articles de Wikipédia. Les données Twitter fournissent des informations sur la variation des usages linguistiques (en français en particulier) en fonction de la structure du réseau, des facteurs sociogéographiques et des domaines linguistiques, les données Wikipédia documentent les interactions des utilisateurs sur des sujets circonscrits, et l'émergence et l'évolution d'un genre de texte, l'article encyclopédique collaboratif en ligne. Une contribution majeure de ce projet est la combinaison de ces deux approches (la modélisation est contrainte par les données et informe la collecte et l'analyse des données) dans un large éventail d'expertises (sociolinguistique, dialectologie, modélisation informatique, science des données et complexité).

Mots-clés : langage, changement, variation, sociolinguistique, réseaux, données issues des médias sociaux, modélisation multi-agent, Twitter, Wikipédia.

Summary: Change and variation are fundamental properties of language. While it is widely accepted that the structure of social interactions plays a major role, we are far from understanding the full complexity of this phenomenon and the factors affecting its dynamics. We will use two approaches to advance our understanding: (1) an agent-based paradigm to modelling the structure and dynamics of the interactions as networks with different properties and types of language users, and (2) the integration of real-world data through the use of annotated twitter messages from the SoSweet corpus and online exchanges involved in the construction of Wikipedia articles. The twitter data provide information about how language usage (here, French) varies with social network structure, socio-geographical factors, and linguistic domains, while the Wikipedia data contain information about how particular users interact with respect to circumscribed topics, and about how these interactions influence the evolution of the articles. A major contribution of this project consists in the seamless combination of these two approaches (where modelling is constrained by real data and, in turns, informs the collection and analysis of such data) and in the broad range of expertise involved (sociolinguistics, dialectology, computational modelling, data science and complexity).

Keywords: Language, change, variation, sociolinguistics, networks, social media data, agent-based models, Twitter, Wikipedia.

Duration: 36 months (September 2021 - September 2024)

Involved laboratories (UMRs)

Dynamique du Langage (UMR 5596)

Interactions, Corpus, Apprentissages, Représentation (UMR 9151)

Project leaders

Marc Allasonnière-Tang (DDL), Jean-Philippe Magué (ICAR)

Consortium

The DDL (Dynamique Du Langage) laboratory explores the interface between the diversity of the world's languages and the universality of human linguistic capacity. The lab is structured around two research axes. Members of the axis DILIS (Linguistic Diversity and its Sources) conduct research in linguistic typology and the history and ecology of languages, with data collection undertaken through fieldwork in different communities around the world. The axis DENDY (Development Neurocognition Disorders) aims at investigating the development and cognitive processing of language in a lifespan perspective, in both typical and atypical populations. Members from both teams are familiar with discourse analysis, language variation, and language acquisition, which will be three of the most important theoretical aspects of the project. Furthermore, members of the DILIS team will also be able to contribute with their knowledge of quantitative computational methods, involving natural language processing and agent-based modeling.

The ICAR laboratory is a multidisciplinary laboratory which is structured around 3 axes covering a large field of research, from interactional linguistics to semiotics, as well as corpus linguistics, natural language processing and didactics. Within this laboratory, the InSitu team and more particularly the Cogcinel sub-team will bring to this project strong skills regarding the study of online interactions, and specifically a sound knowledge of the Twitter corpus, which has already been analyzed both qualitatively and quantitatively. On the other hand, the LanDES sub-team, belonging to the CEDILLES team, will bring to it his expertise in corpus linguistics and natural language processing. This project will therefore give rise to close collaboration between researchers from these two teams Cogcinel and LanDES. During this project, ICAR will provide the corpus of tweets, and will coordinate with the DDL laboratory the collection, annotation, and analysis of Wikipedia data, it will also participate in parameterization of the experiments of the agent-based models as well as in the comparative analysis of real and experimental data.

The Lidilem (Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles) laboratory is a research team of the University of Grenoble Alpes. The unity and specificity of the laboratory lie in a shared interest in the study of language in context and in situations of language transmission. The study of the interactions between language and digital science is one of the two priorities of the current strategy of Lidilem, particularly within the framework of the research action Massive data and modeling of linguistic phenomena in context and through Lidilem's participation in the Multidisciplinary Institute in Artificial Intelligence (MIAI, UGA) through the promoting of the IA & Language chair.

Five relevant publications

- Chevrot, J.-P., Drager, K., Foulkes, P. (2018). Editors' Introduction and Review: Sociolinguistic Variation and Cognitive Science. *Topics in cognitive science*, 10(4): 679-695. [\[Link\]](#)
- Ghimenton, A., Chevrot, J.-P., & Billiez, J. (2013). Language choice adjustments in child production during dyadic and multiparty interactions: a quantitative approach to multilingual interactions. *Linguistics*, 51(2): 413-438. [\[Link\]](#)
- Josserand, M., Allassonnière-Tang M., Pellegrino G., & Dediu, D. (Under review). Interindividual variation refuses to go away: A Bayesian computer model of language change in communicative networks. *Frontiers in Psychology*. [\[Link\]](#)
- 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*. Illinois, United States. [\[Link\]](#)
- Magué, J.-P., Rossi-Gensane, N., & Halté, P. (2020). De la segmentation dans les tweets : Signes de ponctuation, connecteurs, émoticônes et émojis. *Corpus*, 20. [\[Link\]](#)

People involved

DDL

Marc Allassonnière-Tang (MAT): Postdoctoral researcher in the project *Variation, change, and complexity in linguistic and health-related behaviours* (16-IDEX-0005), working on quantitative computational typology with NLP tools and agent-based models. Co-leader of the project with Jean-Philippe Magué, he will coordinate the experiments and reports with agent-based models.

Anna Ghimenton (AG, 20%+ involved): Associate Professor in Language Sciences at the Université Lumière Lyon 2 and member of the DDL research team. She is interested in investigating how language variation is perceived, categorized, acquired and used by monolinguals and bilinguals. She will be involved in the data analyses.

Dan Dediu (DD): IDEXLyon Fellow, PI of the project *Variation, change, and complexity in linguistic and health-related behaviours* (16-IDEX-0005), studies variation using mostly quantitative methods. Will have a supporting role, in the design and analysis of the agent-based models.

Florence Chenu (FC): Engineer specialized in the analysis of linguistic data. She is interested in the acquisition and development of language within written production data.

Mathilde Josserand (MJ): PhD student in Cognitive Science (NSCO doctoral school of Lyon), working on the impact of inter-individual variation on language evolution, using computational methods such as agent-based models and machine learning. She will advise and support the agent-based modelling component of the project.

ICAR

Jean-Philippe Magué (JPM): Associate professor in Language Sciences at the Ecole Normale Supérieure de Lyon. He has been Principal Investigator of the ANR funded project SoSweet, which analyzed language variation on Twitter from a computational sociolinguistics perspective. Co-leader of the project with MAT, JPM will be in charge of coordinating the analyses on the Twitter data and providing consultancy during the building and analysis of the Wikipedia data.

Louise Tarrade (LT): PhD student in Language Sciences (3LA doctoral school of Lyon), working on linguistic change and the impact of social structures on its diffusion at different scales on Twitter, using a computational approach. She will advise Wikipedia data collection and analysis.

Denis Vigier (DV): Associate professor in Language Science at the Université Lumière Lyon 2. He is PI of the Alsan funded project GEODE which studies the major changes in geographic discourses in French encyclopedias. He will be involved in the analysis of genre in Wikipedia.

Lidilem

Jean-Pierre Chevrot (JPC): Professor in Language Science, Université Grenoble Alpes, honorary senior fellow of the Institut Universitaire de France. He spent two years (2015-2017) as invited researcher at the Laboratoire de l'Informatique du Parallélisme, Dynamic Network team, INRIA & Ecole Normale Supérieure, Lyon. His interests of research concern the interfaces of sociolinguistics with cognitive science and data science.

Relevance to the call

The overarching goal of the project is to provide a better understanding of the complex dynamics at play in linguistic communities that produces linguistic variation and drives linguistic change. This goal is aligned with the societal challenge “Culture, Creativity and Inclusive Societies” identified by the European Commission in program Horizon Europe, as well as with the *Action 5 : Human being and Culture* of the French National Strategy for Research which seeks to “*apprehend human phenomena in their individual and social realities*” (Monthubert et al., 2017).

More specifically, the project will investigate how language variation reflects the negotiations in which the speakers are continuously involved, and through which they reshape the linguistic norms of their community. We will therefore consider phenomena at three distinct, but interacting levels simultaneously: *linguistic structures* (e.g. their evolution), *individuals* (e.g. bias imposed by their cognitive apparatus or their social identities) and *population* (e.g. the role of social structures). Such a multilevel approach is the essence of ASLAN which “*accounts for linguistic phenomena in all of their complexity, within an integrative, multidimensional and non-reductionist approach*”.

In order to pursue this multilevel aim, our methodological commitment is to start by building a solid computational base: we will leverage on data science and natural language processing methods applied on massive corpora of social media interactions to propose hypotheses, and on multi-agent models designed to test these hypotheses. The cross-cutting theme “Modeling & Digital Humanities”, which fosters “*data-based approaches to better understand languages and*

language as a communication system” and “*software development aimed at facilitating data understanding (exploration, visualization, analysis, prediction, etc.)*” constitutes a natural and even stimulating destination for this project.

Research statement

The question of the evolution of languages has puzzled scholars for centuries. How can it be possible both that successive generations of speakers use the same language to maintain mutual intelligibility and speak a different language to allow languages to change? This paradox goes to the core of linguistic theories, questioning the very essence of what a language is. As long as one sees a language as an idealization abstracted away from the speakers’ actual productions, such as the saussurean *langue* as opposed to *parole*, the paradox holds, because of the monolithic nature of the abstract idealization.

When, in the 1960’s, Labov shifted the focus of linguistics from an idealized system to the actual productions of speakers, he suggested a solution to this paradox. He showed that the linguistic variation among speakers is not just random noise that must be abstracted out in the course of the scientific process but, on the contrary, that this variation must be the object of scientific inquiry. Variation is the result of speakers using language to perform their identity within their community and to enact various roles depending on the social and interactional setting. This is made possible by the fact that the different variants coexisting in a population for a single linguistic function have different social meaning. When a speaker chooses one specific variant among the others available to the speech community, s/he not only uses it for its linguistic function, but transmits the associated social meaning. Hearers, or readers, then interpret this social meaning to understand the utterance and to infer the social characteristics of the speaker. In the episteme underlying this view, a language is no longer a single abstract monolithic norm, but a system of interrelated norms carrying social representations that each speaker embodies in his/her own way according to his/her socio-cognitive abilities and position in the social organization. Speakers no longer use the exact same language and the paradox of language evolution vanishes. Even better, language evolution can be more adequately addressed. If language variation is the result of speakers choosing among different possible options in competition, language evolution is the result of the dynamics of the competition. In the past half century, research on language variation and change has been fruitful. From the traditional Labovian variationist sociolinguistics to its most recent computational versions, from multi-agent models to lab experiments, it has been studied with various methods and points of view.

Yet, several fundamental questions remain open. The interplay between upward causation, i.e. the contribution of individual speakers to the population dynamic, and downward causation, the influence of the population on the speaker is not fully understood. In this project, we tackle this issue through the dynamics of linguistic innovations. From an upward perspective, we address the question of the influence of speakers in the diffusion of innovation, in particular in relation with their social status and their position in the social network that structures the speech community. From a downward perspective, we look at the dynamics of the linguistic behavior of individual speakers when acquiring, or not, linguistic innovations. To pursue these goals, we adopt a “dual route” methodology, combining real-word data with multi-agent simulations.

In terms of data and methodology, the use of agent-based models is still subject to debate in the linguistic community. On the one hand, agent-based models are now frequently applied to study language evolution. On the other hand, these models are often criticized for oversimplifying the environment of language use (more details in *1.3 Modelling language variation and change*). In an attempt to go beyond this ongoing discussion, the current project has two facets. First, we will integrate real-world data from two sources, annotated tweets from the SoSweet corpus and the history of the successive edits of Wikipedia. Second, we will model the structure and dynamics of the interactions using networks of virtual agents with different properties and types of language users. The Twitter data provides information about how the usage of various aspects of language (here, specifically French) varies with social network structure, socio-geographical factors and linguistic domains, while the Wikipedia data contains information about how particular users interact with respect to circumscribed topics, and on how these interactions lead to the emergence of a new genre, the Wikipedia article. A major contribution of this project consists in the seamless combination of these two approaches (where modelling is constrained by real data and, in turns, informs the collection and analysis of such data) and in the broad range of expertise involved (sociolinguistics, dialectology, computational modelling, data science and complexity). Based on this combination, our overarching research theme (upward and downward perspectives of language variation and change) can be translated in the following two research questions (RQ):

RQ 1: How are collective linguistic conventions constructed through interindividual interactions within social media data? For example, at what threshold of individual acceptance does an innovation generalize in the network? Does this generalization depend on the position in the network of those who have adopted an innovation?

RQ 2: How do collective linguistic conventions influence individuals? Do people get affected by the majority/established sociolinguistic conventions when they enter into a network? Or do they resist the merge? Does this top-down influence depend on the global network structure?

These two questions converge to examine how the collective and the individual interact for shaping norms, variation and change during interaction on social media. They also explore which individual (e.g. position in the network) and collective (e.g. overall shape of the network) factors influence this interaction. Answering these questions will shed light on the overall dynamics of language change, while tackling simultaneously the upward and downward perspectives.

Theoretical framework, methodology

1. Theoretical framework

1.1 Language variation and change

Sociolinguistics studies the interplay between language variation, language change, and society. For decades, sociolinguistics has investigated this interplay through the quantitative analysis of the frequency of sociolinguistic variables (e.g. realizing or not optional liaison in French, using or not the double negative in English) enabling speakers to say the same thing in different ways, with the variants being “identical in reference or truth value, but opposed in their social and/or stylistic significance” (Labov, 1972, p. 271). The study of variation is closely tied to the diachronic

evolution of language, as the alternation between variants is the starting point of change on the one hand (D'Arcy, 2013), and ongoing diachronic changes are causes of synchronic variability on the other hand. This general framework is applicable to communities where several languages are used, as both bilingual and monolingual speakers “continually engage in choices amongst alternatives which have the same referential meaning or function in specific linguistic contexts” (Poplack et al., 2012, p. 207).

Sociolinguistic studies have repeatedly shown that speakers vary in the way they talk depending on several factors. The selection of the language variants is influenced by the context of speech, such as the degree of formality of the interaction (Schilling-Estes, 2004), whereas other factors that characterize the speaker also play a role: regional background (Chambers, 2000), sex and gender (Barbu et al., 2015; Cheshire, 2004; Chevrot et al., 2011), ethnicity (Fought, 2006), socioeconomic status (Ash, 2004), and social network structure (Magué, 2007; Milroy, 1987; Nardy et al., 2014). These linguistic variants carry non-linguistic information, such as the characteristics of the speaker (social status, ethnicity) (Foulkes, 2010). As a consequence, the use of variants is a resource for expressing social identity (Lepage & Tabouret-Keller, 1985). The selection of variants is also influenced by the social context of the communication. A single language is used differently during a conversation between friends or in a philosophical essay. This kind of variation, due to the social setting in which communication takes place (the participants, their roles, their expectations, their goals...), is the subject of a long intellectual tradition that can be traced back to Aristotle's *Rhetorics* and *Poetics*. It has been addressed by different disciplines, including literary studies, and has been in the perimeter of linguistics for decades as a dimension of language variation.

Within linguistics, several branches have tackled this issue and conceptualized it with notions such as genre, style or register. The diversity of the approaches has led to a diversity of precise meaning of these words, sometimes fully synonyms, sometimes not (Biber & Conrad, 2009; Lee, 2001; Trosborg, 1997). In this project, we choose to use the term *genre* as a “socially identifiable communication activity” (Maingueneau, 2013): a speaker, or writer, producing an utterance is aware of the social context in which this production takes place. This social context imposes constraints at the level of the utterance and at the level of the organization of all the utterances produced in this communication activity, that is at the level of text or discourse. Texts or utterances of a given genre share both similar social conditions of production and linguistic features. The features range from the global structure of the text to the syntax, morphology, lexicon, phonology... of the utterances (Yates & Orlikowski, 1992). As social constructs, parts of languages, genres are not monolithic and static but vary and change (Berkenkotter & Huckin, 1995; Miller, 2016). As new communicative contexts appear, new genres can emerge.

In the decade beginning with 2010, the collaborative encyclopedia Wikipedia has been a field of choice to look at the emergence of new genre. It both has roots in the long tradition of texts and radically changes the editorial (therefore social) process of traditional encyclopedias. Emigh & Herring (2005) seems to be the first attempt to characterize the genre of Wikipedia articles. By comparing Wikipedia to another online collaborative encyclopedia, Everything2, and to a print encyclopedia, the Columbia Encyclopedia, they showed that the different technical affordances

of the online encyclopedias lead to different representations of the texts by the contributors, taking Wikipedia closer to the Columbia Encyclopedia than to Everything2. With a similar purpose, Tereszkievicz (2010) provides a quite systematic and deep analysis of genre of Wikipedia. Yet, their approach is synchronic, without paying attention to the evolution of the genre. Clark et al. (2009) adopted a diachronic point of view on the genre of Wikipedia article. Using qualitative methods on a small number of articles, they show that Wikipedia pages gain in structural complexity as they grow, without looking for change in the linguistic characteristic of the text. These studies demonstrate that data sources such as Wikipedia have a great potential for studying the evolution of genre. However, this potential has not been fully developed so far and one of the aims of the project is to fill this gap.

1.2 *The computational turn in sociolinguistics*

The emergence of social media has marked a major turning point in social science research over the last few years, notably by making available to researchers a considerable amount of data, implying a paradigm shift in most fields (Lazer et al., 2009). In fact, this influx of data has allowed the constitution of rich datasets much larger than those on which social sciences researchers used to work, especially in sociolinguistics. To leverage on such large datasets, it has therefore been necessary to move away from traditional sociolinguistics analysis methods and to seize of processing methods used in computational sciences, giving rise to the emergence of a new multidisciplinary field, computational sociolinguistics, combining methodologies of natural language processing and theoretical contributions from sociolinguistics (Nguyen et al., 2016).

The role of individual's sociodemographic variables in linguistic variation has been intensively studied. For example, Abitbol *et al.* (2018), using a corpus of French tweets, noted the correlation of linguistic variables with the income, the location, and the social network of users. In a perspective of prediction of income and age from the features of writing style, Flekova *et al.* (2016) confirm that writing style depends on these social variables. For their part, Nguyen and Eisenstein (2017) propose a method based on Hilbert-Schmidt Independence Criterion to measure the geographic dependence of linguistic variables and highlight here again the correlation between certain linguistic variants and location. More recently, Hovy *et al.* (2020) manage to capture regional variations through character-level vector representations of millions of geolocalized tweets in Europe. Johannsen *et al.* (2015) are interested in syntactic variations according to gender and age, based on a multilingual corpus of user's reviews, while Bamman *et al.* (2014) highlight the link between lexicon, network composition, and user gender on Twitter. Based on data from Facebook, Schwartz *et al.* (2013) had also shown the correlation between lexicon use and variables such as age, gender, and user personality.

As they provide rich information on the network of interactions, social media have offered the opportunity to test the validity of prior empirical trends and theoretical concepts on a large-scale. Thus, Del Tredici and Fernández (2018), from the analysis of several subreddits, have characterized the role of users during the spread of an innovation within this communities based on Granovetter's (1973) then Milroy's (1987) theories on the importance of weak ties in the diffusion of an innovation. Still based on these theories, Laitinen *et al.* (2020) note from the

observation of ego networks in a corpus of geolocated tweets that above a certain size, the fact that networks are mainly constituted of weak or strong ties doesn't seem to constitute a significant distinction in resistance to change. The importance of network size in the spread of innovations was also pointed out by the works of Lev-Ari (2018) who shows, based on data from 100 informants and computer simulations, that individuals who have smaller networks are most linguistically malleable and play a major role in the spread of innovations despite the fact that they interact with fewer people. For their part, Tamburrini *et al.* (2015) have tested the communication accommodation theory on Twitter by observing that users' linguistic behavior seems to actually change according to the community they interact with. This same theory has been validated by Danescu *et al.* (2011) and Doyle *et al.* (2016) who highlighted the fit of users' linguistic style during their interactions on Twitter. Danescu *et al.* (2013) also note from the observation of the interaction norms in two online communities that if a user's linguistic adaptation is visible on her/his arrival in a community with an initial phase of adoption of the community's norms, it is generally followed by a conservative phase during which the evolution of the norm is no longer followed.

1.3 Modelling language variation and change

Following the growth of computational sociolinguistics, multi-agent models have been used to simulate the population dynamics of interconnected speakers who either do or do not share linguistic traits subject to regulation by certain social and cognitive constraints (Hruschka *et al.*, 2009). The models explore the dynamics of language variation, simulating the effects of many different factors, whether social, spatial, linguistic or cognitive in nature. As shown in Figure 1, each agent can be represented as a speaker within a community. Agents are connected to each other (or not) through links, which represent the social connection shared between different individuals in a human population.

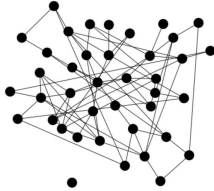
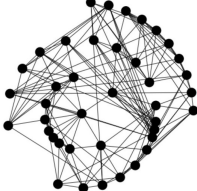
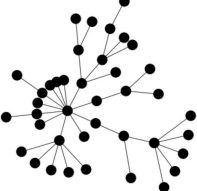
	Random	Small-world	Scale-free
<i>Visualization of the topology</i>			
<i>Degree distribution</i>	Poisson	Exponential	Power-law
<i>Average path length</i>	Short	Short	Short
<i>Clustering coefficient</i>	Small	Large	Rather small

Figure 1. A visualization of agent-based models with different network structures.

Such models can be simple or extremely complicated depending on the chosen parameters. For instance, networks can be big (e.g., more than 500 agents) or small (e.g., less than 10 agents). The connections between the agents can be dense (Figure 1, random networks) or loose (Fig 1, scale-free network), which leads to different types of networks. Variables can be assigned to agents, and the links between the agents can also have different weights. The traits can be used to represent various linguistic or social variants relevant to the study at hand. For instance, the

links between two agents speaking the same language or variety could be less costly than the link between two agents speaking different languages or varieties. Finally, each agent can have its own specific traits and interact in a way that is defined by the traits of its neighbors and interacting partners (Epstein, 2006, p. 6). The traits of the agents are not fixed, which means that the interaction with other agents can change the traits of a given agent, and eventually, the traits of all the agents in the population.

Since the 90s, there has been marked progress in the agent-based modeling of language change and diversity (Gong & Shuai, 2013). Agent-based models cannot travel back in time, but they do allow researchers to partly overcome this problem by simulating various settings relevant to language evolution and make it possible to examine hypotheses at group level, i.e. a level at which experiments are difficult to conduct. One of the most used paradigms is the *iterated learning model* (Simon Kirby et al., 2014). In its original version (Simon Kirby, 1999), agents form a transmission chain, where agents are exposed to, and learn from other agents' observed behavior. Nevertheless, several extensions exist. An agent may receive input from other agents, adapts its internal representation accordingly, and may generate an output to be used as training data for other agents. The transmission can be vertical (from one generation to the next; S. Kirby et al., 2008), horizontal (within individuals in the same society; Raviv, 2020), or a combination thereof (Dediu, 2008, 2009).¹ Thus, agent-based models have been used in the fields of linguistics and sociolinguistics to study lexical evolution (Baronchelli et al., 2006; J. Ke et al., 2002; Lupyan & Christiansen, 2002; Magué, 2005; Puglisi et al., 2008; Steels, 2005), rule-based syntactic or grammatical evolution (Gong, 2011; Gong et al., 2014; Gong & Shuai, 2013; Simon Kirby, 1999; Steels, 2005; Vogt, 2005), the patterning of linguistic diversity (Bie & de Boer, 2007; Dediu, 2008, 2009), the trajectory of language change in heterogeneously biased populations (Gong et al., 2006; J.-Y. Ke et al., 2008; Navarro et al., 2018), or the role of social structure (Gong et al., 2008; J.-Y. Ke et al., 2008; Magué, 2007).

However, such models are not without limitations. Agent-based models are often simplified, capturing only certain features of a much more complex reality (Fagiolo et al., 2007, p. 198). Even if simplifying is essential for many reasons, too much simplicity can be an issue, since some phenomena only emerge at a certain level of complexity. Also, they can be too specific: models may focus on one specific research question, and often ignore other relevant aspects of the question (Gintis, 2013; Treuil et al., 2008; Young, 2006). Some of these limitations can be addressed by creating complex social networks of agents and by manipulating many factors, but finding a good compromise between simplicity and specificity is not always easy (Gong & Shuai, 2013). This project aims at filling this gap by combining the use of agent-based models with real-world data in such a way that the models are constrained by real data, which in turns, informs the collection and the analysis of such data.

2. Materials and method

¹ An important development was the introduction of a Bayesian model of the linguistic agent (Griffiths & Kalish, 2007; S. Kirby et al., 2008), which has several advantages, including its flexibility and theoretical grounding, but which has also been criticized for its strong assumptions and unclear fit to the empirical data (Ferdinand & Zuidema, 2009).

The research will be based on two interdependent tracks. On the one hand, we will provide an in-depth analysis of the real human data (Twitter and Wikipedia) to obtain a good understanding of interaction dynamics within social media data. On the other hand, based on this knowledge, we will run agent-based iterative models to reproduce and predict the evolution of linguistic conventions during human interaction. These two tracks are expected to feedback and mutually improve each other during the process of analysis.

2.1 Materials

Twitter: We will reuse the dataset collected by ICAR during the ANR funded SoSweet project. It consists of 650 million tweets (more than 6 billion words) in French published between 2014 and 2019 by nearly 3 million users. 200 million tweets are already lemmatized and annotated with part of speech and syntactic dependencies and this work will be completed in 2021. The dataset also contains the follower/followee relations for all the users. For 110000 of them, we also have information about their socioeconomic status.

Wikipedia: Data from Wikipedia will allow us to investigate how a genre emerges from the interactions of a community of speakers (or, in that case, writers). We will focus on the French Wikipedia. Behind its interface designed for human readers, Wikipedia provides access to the full history of each of its articles in formats suitable for computational analysis. This constitutes a dataset that indicates for each modification who made it and when. As the French Wikipedia contains more than 2.2 million articles, each having on average 50 versions, the full dataset is too large (several terabytes) to be fully analyzed. We will therefore define a subset of articles based on the timeline of their edits (articles have to be created early enough in the history of Wikipedia to capture all the stages of the evolving genre) and their contributors (with a focus on high contributors, early contributors and contributors involved in the Wikipedia administration). In a similar way as the Twitter dataset, interactions between contributors such as co-appearance in discussion pages will allow us to build a network of contributors. Our goal is to obtain a dataset similar in terms of number of words and contributors to the Twitter dataset.

2.2 Method

For the analysis of real human data, we will explore the construction of linguistic conventions based on different units and domains of linguistic conventions. Two unit sizes will be considered in the project. On the one hand, we will consider phrasal units. The Twitter data is an adequate source of information for such an analysis since tweets rarely exceed the length of a phrase. On the other hand, we will consider supra-phrasal units such as paragraphs or documents. Wikipedia articles as a specific genre are an appropriate source of information for this analysis.

We will diachronically characterize the two datasets (Twitter and Wikipedia) by considering three linguistic domains, lexical, syntactic and semantic, in order to detect changes. The lexical domain relates to the study of the complexity and diversity of lexical items. For both phrasal and supra-phrasal units, we will investigate a) how people vary the diversity of lexical items during interaction. For instance, do people tend to use a similar set of words for a specific discussion topic, and do these conventions vary between contexts such as Twitter and Wikipedia b) how

people vary the complexity of lexical items during interaction. As an example, do people tend to use morphologically complex or simple lexical items for a specific discussion topic or context?

The syntactic domain refers to the variation of syntactic structure during the interaction. As an example, we will consider how the use of negation and tense vary according to different topics and contexts. For the semantic domain, we will investigate if people tend to use semantically similar words during interaction for a specific type of topics or contexts. Moreover, we will also consider the variation of sentiments within the group during the interaction. As an example, does positiveness/negativeness fluctuate more in Twitter than in Wikipedia data?

From an upward perspective, once a change of use has been identified, we will track back its inception and its diffusion in the population (Hoang & Mothe, 2018; Kawamoto, 2013; Schuster & Kolley, 2020). The goal is to understand who are the decisive users/contributors who drive that change and what are their characteristics (such as their position in the social network, their socioeconomic status or their role in Wikipedia administration). To address the downward perspective, the influence of linguistic conventions on speakers, we will pay particular attention to new users or contributors (Cazabet et al., 2013; Pierri et al., 2020). The goal is to monitor their lexical, syntactic, semantic trajectories in order to understand how they acquire, or not, the particularities of the population. Again, characteristics such as the position in the social network (and its evolution), the socioeconomic status or the involvement in Wikipedia administration (and its evolution) will act as explanatory variables. The comparison of the dynamics in the two datasets will inform us on the role of the strength of social structure, much more formal in Wikipedia than on Twitter.

The output generated from the above analysis will provide us with additional insights as to set the parameters of the agent-based models. As an example, which network size should we consider for the experiments? Small or large? How dense should be the connections between the agents to be realistic? How should we set the interaction frequency across time? among others. Based on these parameters, we will be able to simulate the interaction of agents across time and compare it with the gold data of Twitter and Wikipedia. For instance, we will be able to train the agent-based models based on the tweets of a community between year X-Y and compare the simulations of the four following years, for which we have Twitter data. This process will allow us to have a substantial improvement in comparison with conventional simulation experiments that do not have gold data to measure the level of realism of the model.

Description of work

Project Planning

The project will last 36 months (September 2021 – September 2024) and will support the recruitment of one PhD student and one postdoctoral researcher. The PhD student (PhD) will be hired during the first year and will have a contract of 36 months (2021-2024), which is expected to generate a compilation thesis (thesis by articles). The postdoctoral researcher (PdR: potentially MAT, unless he has acquired a permanent position by that time) will be hired six months after the beginning of the project and will have a contract of 24 months (2022-2024). As members of the

project already have functional knowledge of the data (JPM, LT) and agent-based models (MAT, JPM, MJ, DD), the preparation of the experiments will be relatively fast (first six months of the project) and will allow us to proceed more efficiently in gathering part of the data from which the first results may be drawn. As shown in Table 1, we are planning to conduct two main experiments across the three years of the project with each experiment focusing on one of the research questions mentioned in the *Research statement* and being applied on Twitter and Wikipedia data. Each main experiment will be subdivided into three domains of research (lexical, syntactic, semantic).

During the first six months, the PhD student will start by becoming familiar with the literature and tutorials (if needed) on agent-based models. The PhD student will also explore the structure of the Twitter data. After the first six months, the postdoctoral researcher will be hired. The PhD student will be in charge of gathering the Wikipedia data and running experiments with agent-based models. These tasks will be conducted under the day-to-day supervision and support of the postdoctoral researcher. In terms of timeline, while the PhD student is gathering the Wikipedia data, the postdoctoral researcher will use the Twitter data to construct and test the framework of agent-based models that will be used by the PhD student later on. In such a way, the preliminary testing of the postdoctoral researcher will also feedback the PhD student as to how the Wikipedia data should be shaped and constructed.

Table 1. Gantt chart of the project. The colors represent the people in charge. Yellow indicates all members, blue indicates the PhD student (PhD), red refers to the postdoctoral researcher (PdR), purple means both the PhD student and the postdoctoral researcher. The two research questions refer to the upward (RQ1) and downward (RQ2) perspectives of language change.

Action	2021		2022				2023				2024			
	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Interviews for PhD	Yellow													
Contract of PhD		Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	
Data & Literature review		Blue	Blue											
RQ1 Twitter				Purple	Purple	Purple	Purple							
RQ2 Twitter					Purple	Purple	Purple	Purple						
Publish results Twitter							Yellow	Yellow	Yellow	Yellow				
Creating Wikipedia data				Purple	Purple	Purple	Purple							
Interviews for PdR			Yellow											
Contract of PdR				Red	Red	Red	Red	Red	Red					
Mid term report							Yellow							
RQ1 Wikipedia						Purple	Purple	Purple	Purple					
RQ2 Wikipedia							Purple	Purple	Purple	Yellow				
Publish results Wikipedia									Yellow	Yellow	Yellow	Yellow		
Summary of thesis												Blue	Blue	
Final report													Yellow	

The main tasks of the postdoctoral researcher will thus be to a) supervise and support the PhD student b) conduct experiments with agent-based models based on the Twitter and Wikipedia

data c) if possible, also conduct deeper text mining analyses (for instance, with contextual dynamic word embeddings) on the Twitter and Wikipedia data.

Management (DDL, ICAR)

The project management will be done jointly between DDL and ICAR. As one of our goals is to combine two methodological approaches, we place high value on regular gathering of the whole consortium. We will leverage the Institute for Complex Systems (IXXI²) facilities to gather one day every two weeks. As an open-minded pluridisciplinary institute, the IXXI will provide a stimulating environment. These gatherings will alternate between data sessions, seminar, journal club and project management. The recruited postdoc will be in charge of the organization of the operational management. Two large meetings with all members of the project will be organized midway through the project and at the end of the project. Both meetings will serve as opportunities to open a forum for exchange and feedback and to present the results.

Analysis of real data (DDL, ICAR, Lidilem)

The Twitter and Wikipedia data sets will be analyzed in parallel. Twitter data has already been annotated (SoSweet project) facilitating the start of the first analyses. ICAR will be in charge of providing the Twitter data. The Wikipedia data are to be collected, annotated, and then analyzed. Both DDL and ICAR will participate in this process of gathering and annotating the Wikipedia data. ICAR will be in charge of coordinating the management of data collected during this process. ICAR will coordinate the quantitative and qualitative analysis of the data.

Experiments with agent-based models (DDL, ICAR, Lidilem)

The core architecture of the agent-based models has already been developed and used by the team members, facilitating the addition of new parameters and dynamics. Both ICAR and DDL will join forces in designing the parameters and dynamics of the experiments according to the existing theoretical frameworks. DDL will be in charge of conducting the agent-based experiments, of storing and of analyzing their outputs.

Valorisation (DDL, ICAR, Lidilem)

Both DDL and ICAR will participate in valorization actions for the scientific/academic and extra-academic communities. In particular, we target recruitment fairs, student fairs, partnership with engineering schools (INSA, ECL), LDigital Network and mainstream events (Nuit des chercheurs, fête de la science, etc.).

Publications program

The results will be published in national/international journals and conferences. In terms of journal, we will target journals specializing in linguistics (such as *Language Dynamics and Change*, *Langage et Société*, and *Journal of Language Evolution*), cognitive science (such as *Cognition* and *Cognitive Science*), as well as more general venues (such as *PLoS ONE*, *Science Advances*, *Scientific Reports*, and *PNAS*). In terms of conferences, we will aim at presenting our work (and receive feedback) in conferences for linguistics, cognitive science, and computer science.

² www.ixxi.fr

Examples of potential conferences are the Annual Meeting of the Societas Linguistica Europaea (SLE), the Annual Meeting of the Cognitive Society (CogSci), the Annual Meeting of the Association for Computational Linguistics (ACL) or events organized by the Wikimedia foundation. In addition to presenting the progress and results of the project at various international conferences and journals, the plan is also to organize two workshops during the mid-term and final reports of the project. The first will be proposed as a workshop at the Annual meeting of the SLE while the second will be organized in Lyon.

Results and scientific impacts

Given that the project has two relatively separate but tightly interwoven aspects, we expect that, on the one hand, each of them will have relatively specialized results, but also that there will be results that combine the two, on the other. More precisely, we project that the analysis of the real-world data will further our insights into *how* (and, much less probable but still possible, *why*) change spreads (or fails to spread) in complex networks of actual language users. Complementarily, we expect that the agent-based models, designed and constrained given existing real-world data and theories, will, in turn, not only inform the analysis of such data but suggest *causal mechanisms* that may be responsible for the observed patterns. Concretely, the project will produce a database of the analyzed real-world data and the scripts for analyzing it, as well as the code implementing the agent-based models, their results and the scripts needed for their analysis. We trust that these data, code and results will contribute to the further development of the field, by providing a solid basis for asking new and more refined questions, for performing exploratory analysis and even testing new hypotheses. The main relevance of the project will be for the language sciences in general, and sociolinguistics in particular, but we can also imagine that our methods and theories can be extended to other questions concerning the dynamics of cultural phenomena in complex networks, such as the spread of innovations or attitudes and behaviors relevant to health (e.g., in the current context, beliefs about covid-19 and the measures needed to mitigate it).

Social, economic impacts

The models elaborated throughout this project allow to better comprehend how network size and variation influence practices across genres and in new interactional computer mediated practices. In a fast-evolving society, where new and efficient ways of knowledge transmission and teaching are highly sought after, the results and the models elaborated will contribute to the educational and awareness projects involving computer mediated interactions in order to capture the most efficient ways (interactional and linguistic levels) to spread and transmit knowledge. Furthermore, this project strives for the Open Science movement making the data readily available to the scientific community. We plan to contract with the ORTOLANG³ infrastructure which offers services for corpora hosting and archiving and is one of the French nodes of CLARIN⁴, the European Research Infrastructure for Language Resources and Technology. On the one hand, it will contribute to the field of Digital Humanities and on the other hand, the structure of the project

³ <https://www.ortolang.fr/>

⁴ <https://www.clarin.eu/>

can also serve as a model for future projects as it shows how real data and agent-based models can enhance one another.

Budget

The required budget total is 229 916 euros. The details of expenses are listed in Table 2.

Table 2. An overview of the required budget. The currency is euro.

Item	Content	Cost	Total
Staff	• 1 PhD student 36 months	3078/month (including charges)	110808
	• 1 postdoc 18 months	4442/month (including charges)	106608
Travel	• 5 conference travels	1250/travel	6225
Other expenses	• 2 computers	1250/computer	2500
	• 3 open-access publications	1250/publication	3750
Total			229916

- Staff (217k€, > 80%): We will recruit one PhD student and one postdoctoral researcher. The PhD student will be in charge of gathering the Wikipedia data and running experiments with agent-based models on both Twitter and Wikipedia data. These tasks will be conducted under the day-to-day supervision and support of the postdoctoral researcher. If possible, the postdoctoral researcher will also conduct deeper text mining analyses on the Twitter and Wikipedia data.
- Travel (6k€): consortium meetings will be held in Lyon (no cost). The project will fund national and international conferences for the PhD student and the postdoctoral researcher (if travel regulations permit).
- Other expenses (6k€): computer equipment and potential open-access publication fees.

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. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 1125–1134. <https://doi.org/10.1145/3178876.3186011>
- Ash, S. (2004). Social Class. In J. Chambers, P. Trudgill, & N. Schilling-Estes, *Handbook of Language Variation and Change* (pp. 402–422). Blackwell.
- 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>
- Barbu, S., Nardy, A., Chevrot, J.-P., Guellaï, B., Glas, L., Juhel, J., & Lemasson, A. (2015). Sex differences in language across early childhood: Family socioeconomic status does not impact boys and girls equally. *Frontiers in Psychology*, 6(1874). <https://doi.org/10.3389/fpsyg.2015.01874>
- Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06), P06014–P06014. <https://doi.org/10.1088/1742-5468/2006/06/P06014>
- Berkenkotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power* (pp. xiv, 190). Lawrence Erlbaum Associates, Inc.
- Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511814358>

- Bie, P., & de Boer, B. (2007). *An agent-based model of linguistic diversity*.
- Cazabet, R., Pervin, N., Toriumi, F., & Takeda, H. (2013). Information Diffusion on Twitter: Everyone Has Its Chance, But All Chances Are Not Equal. *2013 International Conference on Signal-Image Technology & Internet-Based Systems*, 483–490. <https://doi.org/10.1109/SITIS.2013.84>
- Chambers, J. (2000). Region and language variation. *English World-Wide*, 21(2), 169–199.
- Cheshire, J. (2004). Sex and gender in variationist research. In J. Chambers, P. Trudgill, & N. Schilling-Estes, *Handbook of Language Variation and Change* (pp. 423–443). Blackwell.
- Chevrot, J.-P., Nardy, A., & Barbu, S. (2011). Developmental dynamics of SES-related differences in children's production of obligatory and variable phonological alternations. *Language Sciences*, 33(1), 180–191. <https://doi.org/10.1016/j.langsci.2010.08.007>
- Clark, M. J., Ruthven, I., & Holt, P. O. (2009). The evolution of genre in Wikipedia. *Journal for Language Technology and Computational Linguistics*, 25(1). <https://rgu-repository.worktribe.com/output/248641/the-evolution-of-genre-in-wikipedia>
- Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: Linguistic style accommodation in social media. *Proceedings of the 20th International Conference on World Wide Web - WWW '11*, 745. <https://doi.org/10.1145/1963405.1963509>
- Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 307–318. <https://doi.org/10.1145/2488388.2488416>
- D'Arcy, A. (2013). Variation and Change. In Bayley, R. Cameron, & C. Lucas, *The Oxford Handbook of Sociolinguistics* (pp. 484–502). Oxford University Press.
- Dediu, D. (2008). The role of genetic biases in shaping the correlations between languages and genes. *Journal of Theoretical Biology*, 254(2), 400–407. <https://doi.org/10.1016/j.jtbi.2008.05.028>
- Dediu, D. (2009). Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? *Journal of Theoretical Biology*, 259(3), 552–561. <https://doi.org/10.1016/j.jtbi.2009.04.004>
- Del Tredici, M., & Fernández, R. (2018). The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. *ArXiv:1806.05838 [Cs]*. <http://arxiv.org/abs/1806.05838>
- Doyle, G., Yurovsky, D., & Frank, M. C. (2016). A Robust Framework for Estimating Linguistic Alignment in Twitter Conversations. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 637–648. <https://doi.org/10.1145/2872427.2883091>
- Emigh, W., & Herring, S. C. (2005). Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 99a–99a. <https://doi.org/10.1109/HICSS.2005.149>
- Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.
- Fagiolo, G., Moneta, A., & Windrum, P. (2007). A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems. *Computational Economics*, 30(3), 195–226. <https://doi.org/10.1007/s10614-007-9104-4>
- Ferdinand, V., & Zuidema, W. (2009). Thomas' theorem meets Bayes' rule: A model of the iterated learning of language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31(31), 1069–7977.
- Flekova, L., Preoțiu-Pietro, D., & Ungar, L. (2016). Exploring stylistic variation with age and income on twitter. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 313–319.
- Fought, C. (2006). *Language and Ethnicity*. Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511791215>
- Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory*

- Phonology*, 1(1), 5–40.
- Gintis, H. (2013). Markov models of social dynamics: Theory and applications. *ACM Transactions on Intelligent Systems and Technology*, 4(3), 1–19. <https://doi.org/10.1145/2483669.2483686>
- Gong, T. (2011). Simulating the coevolution of compositionality and word order regularity. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 12(1), 63–106. <https://doi.org/10.1075/is.12.1.03gon>
- Gong, T., Minett, J., & Wang, W. (2008). Exploring social structure effect on language evolution based on a computational model. *Connection Science*, v.20, 135-153 (2008), 20. <https://doi.org/10.1080/09540090802091941>
- Gong, T., Minett, J., & Wang, W. (2006). *Language Origin and the Effects of Individuals' Popularity*. 999–1006. <https://doi.org/10.1109/CEC.2006.1688418>
- Gong, T., & Shuai, L. (2013). Computer simulation as a scientific approach in evolutionary linguistics. *Language Sciences*, 40, 12–23. <https://doi.org/10.1016/j.langsci.2013.04.002>
- Gong, T., Shuai, L., & Zhang, M. (2014). Modelling language evolution: Examples and predictions. *Physics of Life Reviews*, 11(2), 280–302. <https://doi.org/10.1016/j.plrev.2013.11.009>
- Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1360–1380. <https://doi.org/10.1086/225469>
- Griffiths, T. L., & Kalish, M. L. (2007). Language Evolution by Iterated Learning With Bayesian Agents. *Cognitive Science*, 31(3), 441–480. <https://doi.org/10.1080/15326900701326576>
- Hoang, T. B. N., & Mothe, J. (2018). Predicting information diffusion on Twitter – Analysis of predictive features. *Journal of Computational Science*, 28, 257–264. <https://doi.org/10.1016/j.jocs.2017.10.010>
- Hovy, D., Rahimi, A., Baldwin, T., & Brooke, J. (2020). Visualizing Regional Language Variation Across Europe on Twitter. In S. D. Brunn & R. Kehrein (Eds.), *Handbook of the Changing World Language Map* (pp. 3719–3742). Springer International Publishing. https://doi.org/10.1007/978-3-030-02438-3_175
- Hruschka, D. J., Christiansen, M. H., Blythe, R. A., Croft, W., Heggarty, P., Mufwene, S. S., Pierrehumbert, J. B., & Poplack, S. (2009). Building social cognitive models of language change. *Trends in Cognitive Sciences*, 13(11), 464–469. <https://doi.org/10.1016/j.tics.2009.08.008>
- Johannsen, A., Hovy, D., & Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 103–112. <https://doi.org/10.18653/v1/K15-1011>
- Kawamoto, T. (2013). A stochastic model of tweet diffusion on the Twitter network. *Physica A: Statistical Mechanics and Its Applications*, 392(16), 3470–3475. <https://doi.org/10.1016/j.physa.2013.03.048>
- Ke, J., Minett, J. W., Au, C.-P., & Wang, W. S.-Y. (2002). Self-organization and selection in the emergence of vocabulary. *Complexity*, 7(3), 41–54. <https://doi.org/10.1002/cplx.10030>
- Ke, J.-Y., Gong, T., & Wang, W. (2008). Language change in social networks. *Communications in Computational Physics (Cicp)*, v.3, 935-949 (2008), 3.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681–10686. <https://doi.org/10.1073/pnas.0707835105>
- Kirby, Simon. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford University Press.
- Kirby, Simon, Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114. <https://doi.org/10.1016/j.conb.2014.07.014>
- Labov, W. (1972). Some Principles of Linguistic Methodology. *Language in Society*, 1, 97. edsjsr.
- 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>
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N.,

- Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational Social Science. *Science*, 323(5915), 721–723. <https://doi.org/10.1126/science.1167742>
- Lee, D. Y. (2001). *Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle*. https://doi.org/10.1163/9789004334236_021
- Lepage, R. B., & Tabouret-Keller, A. (1985). *Acts of Identity: Creole-Based Approaches to Language and Ethnicity*. Cambridge University Press.
- Lev-Ari, S. (2018). Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, 176, 31–39. <https://doi.org/10.1016/j.cognition.2018.03.003>
- Lupyan, G., & Christiansen, M. H. (2002). Case, Word Order, and Language Learnability: Insights from Connectionist Modeling. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 596–601.
- Magué, J.-P. (2005). *Changements sémantiques et cognition: Différentes méthodes pour différentes échelles temporelles* [Université Lumière - Lyon II]. <http://tel.archives-ouvertes.fr/tel-00410044>
- Magué, J.-P. (2007). On the importance of population structure in computational models of language evolution. *31st Pennsylvania Linguistic Colloquium*. 31st Pennsylvania Linguistic Colloquium, Philadelphia, PA. <http://halshs.archives-ouvertes.fr/halshs-00410043>
- Maingueneau, D. (2013). Genres de discours et web: Existe-t-il des genres web ? In C. Barats (Ed.), *Manuel d'analyse du web* (Armand Colin).
- Miller, C. R. (2016). Genre Innovation: Evolution, Emergence, or Something Else? *The Journal of Media Innovations*, 3(2), 4–19. <https://doi.org/10.5617/jmi.v3i2.2432>
- Milroy, L. (1987). *Language and Social Networks*. Wiley. <http://books.google.fr/books?id=rFliAAAAMAAJ>
- Monthubert, B., Aschiéri, G., Béjean, S., Gillot, D., Guillou, M., Jégo-Laveissière, M.-N., Déaut, J.-Y., le, Masson-Delmotte, V., Pisani-Ferry, J., Plateau, B., Taddei, F., Thoury, C., & Villani, C. (2017). *Livre Blanc de l'enseignement supérieur et de la recherche*.
- Nardy, A., Chevrot, J.-P., & Barbu, S. (2014). Sociolinguistic convergence and social interactions within a group of preschoolers: A longitudinal study. *Language Variation and Change*, 26(03), 273–301. <https://doi.org/10.1017/S0954394514000131>
- Navarro, D., Perfors, A., Kary, A., Brown, S., & Donkin, C. (2018). When Extremists Win: Cultural Transmission Via Iterated Learning When Populations Are Heterogeneous. *Cognitive Science*, 42. <https://doi.org/10.1111/cogs.12667>
- Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational Sociolinguistics: A Survey. *ArXiv:1508.07544 [Cs]*. <http://arxiv.org/abs/1508.07544>
- Nguyen, D., & Eisenstein, J. (2017). A Kernel Independence Test for Geographical Language Variation. *Computational Linguistics*, 43(3), 567–592. https://doi.org/10.1162/COLI_a_00293
- Pierri, F., Piccardi, C., & Ceri, S. (2020). Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Scientific Reports*, 10(1), 1372. <https://doi.org/10.1038/s41598-020-58166-5>
- Poplack, S., Zentz, L., & Dion, N. (2012). Phrase-final prepositions in Quebec French: An empirical study of contact, code-switching and resistance to convergence. *Bilingualism: Language and Cognition*, 15(02), 203–225. <https://doi.org/10.1017/S1366728911000204>
- Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, 105(23), 7936–7940. <https://doi.org/10.1073/pnas.0802485105>
- Raviv, L. (2020). *Language and society: How social pressures shape grammatical structure* [PhD Dissertation]. Radboud Universiteit Nijmegen.
- Schilling-Estes, N. (2004). Investigating Stylistic Variation. In J. Chambers, P. Trudgill, & N. Schilling-Estes, *Handbook of Language Variation and Change* (pp. 375–401). Blackwell.
- Schuster, J., & Kolleck, N. (2020). The Global Diffusion of Social Innovations – An Analysis of Twitter

- Communication Networks Related to Inclusive Education. *Frontiers in Education*, 5. <https://doi.org/10.3389/feduc.2020.492010>
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), e73791. <https://doi.org/10.1371/journal.pone.0073791>
- Steels, L. (2005). The emergence and evolution of linguistic structure: From lexical to grammatical communication systems. *Connection Science*, 17(3–4), 213–230. <https://doi.org/10.1080/09540090500269088>
- Tamburrini, N., Cinnirella, M., Jansen, V. A. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40, 84–89. <https://doi.org/10.1016/j.socnet.2014.07.004>
- Tereszkiewicz, A. (2010). *Genre analysis of online encyclopedias. The case of Wikipedia* (1st edition). Jagiellonian University Press.
- Treuil, J.-P., Drogoul, A., & Zucker, J.-D. (2008). *Modélisation et simulation à base d'agents: Exemples commentés, outils informatiques et questions théoriques*. Dunod.
- Trosborg, A. (1997). Text typology: Register, genre and text type. In A. Trosborg (Ed.), *Text typology and translation* (pp. 3–24). John Benjamins Publishing.
- Vogt, P. (2005). On the Acquisition and Evolution of Compositional Languages: Sparse Input and the Productive Creativity of Children. *Adaptive Behavior*, 13(4), 325–346. <https://doi.org/10.1177/105971230501300403>
- Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structural approach to studying communication and media. *The Academy of Management Review*, 17(2), 299–326. <https://doi.org/10.2307/258774>
- Young, H. P. (2006). Chapter 22 Social Dynamics: Theory AND Applications. In *Handbook of Computational Economics* (Vol. 2, pp. 1081–1108). Elsevier. [https://doi.org/10.1016/S1574-0021\(05\)02022-8](https://doi.org/10.1016/S1574-0021(05)02022-8)