Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis

Jacobo Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, Eric Fleury

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).

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.

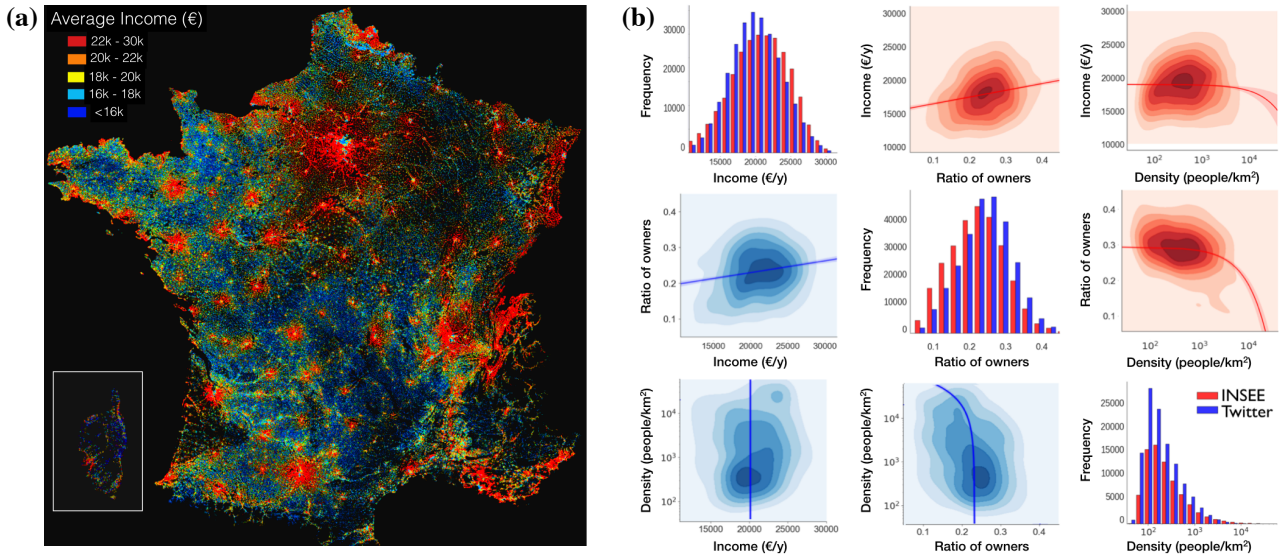

Figure 1: Distributions and correlations of socioeconomic indicators. (a) Spatial distribution of average income in France with $200m \times 200m$ resolution. (b) Distribution of socioeconomic indicators (in the diag.) and their pairwise correlations measured in the INSEE (upper diag. panels) and Twitter 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

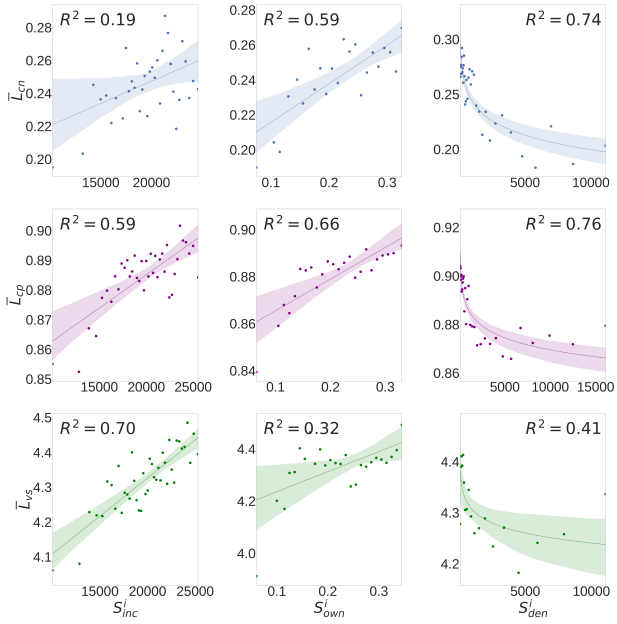

Figure 2: Pairwise correlations between three SES indicators and three linguistic markers. Columns correspond to SES indicators (resp. S^i_{inc}, S^i_{own}, S^i_{den}), while rows correspond to linguistic variables (resp. $\overline{L}_{\text{cn}}$, $\overline{L}_{\text{cp}}$ and $\overline{L}_{\text{vs}}$). 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).

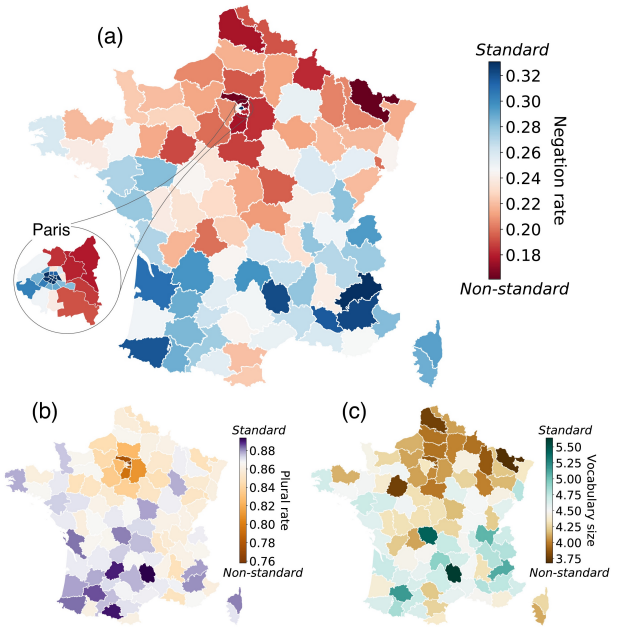

Figure 3: Geographical variability of linguistic markers in France. (a) Variability of the rate of correct negation. Inset focuses on larger Paris. (b) Variability of the rate of correct plural terms. (c) Variability of the average vocabulary size set. Each plot depicts variability on the department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

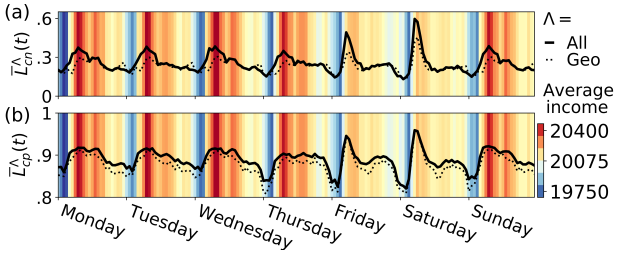

Figure 4: Temporal variability of (a) $\overline{L}^{\Lambda}_{\text{cn}}(t)$ (resp. (b) $\overline{L}^{\Lambda}_{\text{cp}}(t)$) average rate of correct negation (resp. plural terms) over a week with one hour resolution. Rates were computed for $\Lambda = all$ (solid line) and $\Lambda = geo$located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

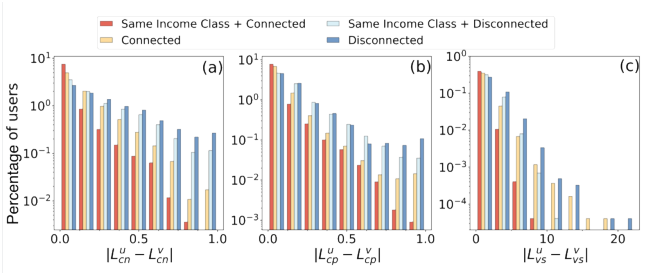

Figure 5: Distribution of the $|L^u_* - L^v_*|$ absolute difference of linguistic variables $* \in \{\text{cn}, \text{cp}, vs\}$ (resp. panels (a), (b), and (c)) of user pairs who were connected and from the same socioeconomic group (red), connected (yellow), disconnected and from the same socioeconomic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

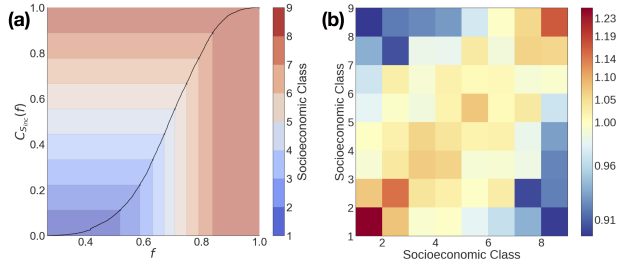

Figure 6: (a) Definition of socioeconomic classes by partitioning users into nine groups with the same cumulative annual income. (b) Structural correlations between SES groups depicted as matrix of the ratio $|E(s_i, s_j)|/|E_{rand}(s_i, s_j)|$ between the original and the average randomized mention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