How do we copy and paste? The semantic drift of quotations in blogspace

SÉBASTIEN LERIQUE and CAMILLE ROTH

1. INTRODUCTION

The understanding of the mechanisms behind cultural similarity and diversity has led to a sizeable literature in the recent past, spanning over a vast area of research fields ranging from cultural anthropology to social network analysis and complex systems modelling, all diversely labelled as studies on "opinion dynamics", "cultural evolution", or "information diffusion", for the most part. Broadly, this type of research program investigates phenomena pertaining to both cognitive science and social science, and aims to understand the processing, transmission, and evolution of information both at the individual and social levels.

Several theories have been proposed and debated within these research fields, mixing social and individual cognition, notably from the side of cultural anthropology. First, the debate around the "memetic" program initiated by Dawkins [1976], for which the collection of works by Aunger [2000] provides a solid overview. A second field has focused on the development of evolutionary models of norms (see for instance Ehrlich and Levin 2005), following the seminal work of Boyd and Richerson [1985].

A third approach, "cultural epidemiology", initiated by Sperber [1996] recently attracted significant attention. Works such as Atran [2003] argue that this approach is anthropologically better suited than memetics, and the main issues in this debate are further detailed by Kuper [2000] and Bloch [2000]. The cultural epidemiology program focuses on the notion of *representation*, linking the cognitive science concept of *mental representation* to the concept of *public representation*, the latter being the former's counterpart outside the brain (i.e. in cultural artefacts: texts, utterances, etc.).[1] The key point here is that representations are not being replicated through a high-fidelity copy process, but are being interpreted and produced anew, and are thus greatly subject to change. Cultural epidemiology postulates this type of conceptual evolution and suggests that it can be appraised through the notion of "cultural attractor", seen as the attraction domain of an underlying socio-semantic dynamical system. Despite some recent modelling attempts (for instance Claidière and Sperber [2007]), the development of quantitative measurements focused on the key notion of cultural attractor has remained a relatively hard task and, to our knowledge, this hypothesis has not yet been empirically analysed.

However, the last decade has witnessed an avalanche of observable *in vivo* data in the form of online interactions. While they are not records of "physical" inter-individual interactions (in the sense of "real life" interactions), these productions and information trails still constitute a wealth of observations on the dynamics of public – albeit online – representations. Given the already significant – and rapidly growing – importance of our online interactions, these records can dramatically improve the prospects of empirical study of the individual-level processes of cultural evolution.

We aim to empirically describe the transformation of a specific type of public representation, by focusing on the possible alterations introduced by individuals when newly producing a representation. To deal with robust and simple cultural representations, we paid attention to the evolution of quotations. While these verbatim public representations should in theory not suffer any alterations as they are produced anew (as opposed to more elaborate expressions and opinions, not identified as quoted utterances), empirical observation shows that they are in fact quite often transformed. We will in particular exhibit a non-trivial process by which individual words in quotations are replaced. We will uncover some of the semantic and structural characteristics of these words and the substitutions they undergo. More generally, we contend that using this type of data is equivalent to a large-scale psycholinguistic experiment and at the same time constitutes the first step towards building empirically realistic models of cultural evolution.

The next section (Sec. 2) describes the state-of-the-art on this matter. In Sec. 3, we detail the empirical protocol and the various assumptions that were made in order to deal with the available empirical material. Sec. 6 describes the significant psycholinguistic biases observed during *in vivo* quotation reformulation as well as their epidemiological setting, followed by a discussion and general guidelines for further work in Sec. 6.

2. RELATED WORK

The relevant literature on *public representation dynamics* features two main streams. On one hand, we find studies of the macroscopic *social diffusion* of public representations, describing for instance the propagation of cultural artefacts across social networks such as blogspace [Gruhl et al. 2004], the characteristic times and diffusion cycles both within these social networks and with respect to the topical dynamics of news media [Leskovec et al. 2009a], or the reciprocal influence between the social network topology and the distribution of issues [Cointet and Roth 2009]. These studies are relatively independent from anthropology and cognition and are at the interface between data mining, complex systems and quantitative sociology (first and foremost social network analysis). Without necessarily relying on specific social science theories, this research stream is of interest for its use of large social media corpora in studying cultural dynamics.

On the other hand, the study of the *transformation* of public representations has emerged only recently. For one, models involving evolution and representations to study the notion of "cultural attractor" have appeared only a few years ago [Claidière and Sperber 2007]. Among the empirical approaches on the mutation of representations, some of the most relevant studies to date consist in a series of papers investigating *quotation* transformations in a large corpus of US blog posts, initially collected and studied by Leskovec et al. [2009a] and further analyzed by Simmons et al. [2011] and Omodei et al. [2012]. They show several types of regularities and

[1] Sperber [1996] emphasizes this distinction in his seminal work:

> A representation may exist inside its user: it is then a *mental representation*, such as a memory, a belief, or an intention. The producer and the user of a mental representation are one and the same person. A representation may also exist in the environment of its user, as is the case, for instance, of the text you are presently reading: it is then a *public representation*.

propose diffusion-transformation models of the evolution of quotations, which may nonetheless appear to be relatively simplistic from a cognitive viewpoint. One of the main conclusions of these works is that even for quotations, a type of public representation that should be among the most stable, it is still possible to observe and measure significant transformations. However, these studies address transformations by focusing on the properties of the source of the quotation (e.g. news outlet *v.* blog), or the surrounding public space (e.g. quotation frequency in the corpus), rather than the very cognitive-level features which may determine or, at least, influence these transformations.

At this level, we have to turn to the broader psycholinguistic literature which provides one of the main cognitive foundations for public representation evolution by studying the influence of word features on the ease of recall. This field is well developed and details the impact that classical psycholinguistic variables such as word frequency (see Yonelinas [2002] for a review), age-of-acquisition [Zevin and Seidenberg 2002], number of phonemes or number of syllables (see for instance Rey et al. [1998] and Nickels and Howard [2004]), have in this type of task. ***[add a ref for word frequency]***

Less classical linguistic variables, based on the study of semantic network properties, have recently appeared as an empirical investigation field, after having been heavily discussed around the notion of connectionism and its normative processual models (see for instance Collins and Loftus [1975]). Let us mention two interesting and recent studies on that matter, which demonstrate in a strictly *in vitro* framework and at the vocabulary level that word properties computed on a word network are important factors for the cognitive processes and reproduction of those words. First, Griffiths et al. [2007] analyse a task where patients are asked to name the first word which comes to their mind when they are presented with a random letter from the Latin alphabet. The authors show that there exists a link between the ease of recall of words and one of their semantic features, namely their authority position (pagerank) in a language-wide semantic network built from external word association data. A second psycholinguistic study by Chan and Vitevitch [2010] shows, in a picture-naming task, that words are produced faster when they have a higher clustering coefficient in an underlying phonological network (which, again, is defined from external phonological data).

[add *https://www.sciencedirect.com/science/article/pii/S0893608012000330, http://research.clps.brown.edu/austerweil/pdfs/papers/randomWalkNips2012.pdf, http://link.springer.com/article/10.3758/s13421-013-0312-y* ***to that picture]***

On the whole, the current psycholinguistic state-of-the-art seems to hint towards two antagonistic types of results. On one hand, part of the literature tends to show that recall is easier for the least "awkward" words; those whose age of acquisition is earlier, length is smaller, semantic network position is more central – this is particularly true in tasks where participants are asked to form spontaneous associations or utter a word in response to a given signal [Citation needed]. On the other hand, when the task consists in remembering a specific list of items, "awkward" words are actually more easily remembered, possibly as they are more informative and plausibly more discernible [Citation needed]. The jury is still out as to whether reformulation alteration, i.e. spontaneous replacement of words when asked to repeat a given utterance, is rather of the former or latter sort. Our paper additionally sheds light on this debate. ***[Show the link with epidemiology, and fitness: oddness is a kind of fitness]***

Stepping back, we observe a gap between, on one side, macro-level empirical studies of the diffusion dynamics in a social system and, on the other side, studies focused on micro-level transformations of representations — these latter studies being either strongly normative, or with results difficult to articulate with realistic cultural epidemiology models.

3. PROTOCOL

In order to start bridging this gap, we set out to *empirically* study public representation transformations at the microscopic level, aiming to stay compatible with macroscopic-level studies of these public representations. Quotations appeared to be a perfect candidate as public representations. First, they are usually cleanly delimited by quotation marks (and often with HTML markup in web pages), which greatly facilitates their detection in text corpora. Second, they stem from a unique "original" version, and could ideally be traceable back to that version. Third, and most importantly, their duplication should *a priori* be highly faithful, apart from cases of cropping: not only should transformations be of moderate magnitude, but when specific words are not perfectly duplicated, it is safe to assume that the variation is due to involuntary cognitive bias — as writers may expect any casual reader to easily verify, and thus criticize, the fidelity to the original quotation. Quotation evolution is therefore a perfect environment to measure cognition-induced transformations and relate those findings to macroscopic social dynamics.

3.1 Dataset

We used a reliable quotation dataset collected by Leskovec et al. [2009a], large enough to lend itself to statistical analysis. This dataset consists of the daily crawling of news stories and blog posts from around a million online sources, with an approximate publication rate of 900k texts per day, over a nine-month period of time (from August 2008 to April 2009) [Leskovec et al. 2009b].[2] Quotations were then automatically extracted from this corpus: each quotation is a more or less faithful excerpt of an utterance (oral or written) by the quoted person. **[C:** btw do we have an excerpt of quotations here, could we even feature a few prototypical examples?] Quotations were then gathered in a graph and connected according to their similarity: either because they differ by very few words (in that case, no more than one word) or because they share a certain sequence of words (in that case, at least ten consecutive words). A community detection algorithm was applied to that quotation graph to detect aggregates of tightly connected, i.e. sufficiently similar, groups of quotations (see Leskovec et al. [2009a] for more detail). This analysis yielded the final data we had access to, with a total of about 70,000 sets of quotations; each of these sets allegedly contains all variations of a same parent utterance, along with their respective publication URLs and timestamps.

3.2 Word-level measures

To keep the analysis palatable, we restricted the analysis to quotation transformations which consisted in the *substitution* of a word by another word (and only those cases) in order to unambiguously discuss single word replacements. To quantify those substitutions, we decided to associate a number of features to each word, the variation of which we can statistically study. The following sections detail the features we used.

[2] Unfortunately, the original article [Leskovec et al. 2009a] does not provide additional details on the source selection methodology.

3.2.1 *Standard psycholinguistic indices.* We first introduce some of the most classical psycholinguistic measures on words:
[Add some bibliography about those features' known effects]

—**Word frequency**: the frequency at which words appear in our dataset, ***[add ref on why it's important]***

—**Age of Acquisition**: the average age at which words are learned, obtained from Kuperman et al. [2012],

—The average **Number of Phonemes** for all pronunciations of a word, obtained from the Carnegie Mellon University Pronouncing Dictionary [Weide 1998],[3]

—The average **Number of Syllables** for all pronunciations of a word, also obtained from the CMU Pronouncing Dictionary,

—The average **Number of Synonyms** for all meanings of a word, obtained from WordNet [WordNet 2010]. ***[add ref on why it's important]***

We also considered grammatical types within quotations by detection of *Part-of-Speech* (POS) categories, using the Penn TreeBank Project typology [Santorini 1990] and thereby distinguishing verbs, nouns, adjectives and adverbs. The results were however extremely similar across the various categories, exhibiting no specific effect of words belonging to different POS categories. ***[See #8 for this fact-check]***

3.2.2 *Network-based measures.* Aside from classical psycholinguistic measures, we also considered more recently studied variables based on semantic network properties. We relied on the "free association" norms collected by Nelson et al. [2004] which naturally embed information on the idea association process underlying transformation of quotations.

Free association (FA) norms record the words that come to mind when someone is presented with a given cue (that is the "free association" task). As Nelson et al. explain,

> free association response probabilities index the likelihood that one word can cue another word to come to mind with minimal contextual constraints in effect. [Nelson et al. 2004]

Following Griffiths et al. [2007], we first build a directed unweighted network based on association norms, where words are nodes and edges are directed from cue to target word whenever a target word is being produced when this particular cue word was presented. This network is of particular interest since it measures the *in-vitro forced-choice* version of a substitution whereas the data we analyse is the *in-vivo spontaneous* version of what we otherwise hypothesize to be the same process.

We introduce three standard network-based measures to be used on the FA network:

—**Centrality** k, initially measured by the number of incoming edges to a given node, i.e. the number of cues for which a given word is triggered as an association, which strongly relates to word polysemy. However in the present case there is a quasi-perfect correlation between node incoming degree and node *pagerank* [Page et al. 1999], which will lead us to favour the latter later on. Word pagerank on the FA network had already been used by Griffiths et al. [2007]; it may be interpreted as a generalized and recursive measure of word polysemy: central nodes in the pagerank sense are words often selected as targets when presented with cues themselves often selected as targets, and so on recursively.

—**Clustering coefficient** c, which measures the extent to which a node belongs to a local aggregate of tightly connected nodes, and defined as the ratio between the number of actual *v.* possible edges between a node's neighbours [Watts and Strogatz 1998]. We compute the clustering coefficient on the undirected version of the FA network; we thus measure if a word belongs more or less to a local aggregate of equivalent words (from a "free association" point of view).

—**Betweenness coefficient** b, another measure of node centrality describing the extent to which a node tends to connect otherwise remote areas of the network [Freeman 1977]. More technically, it corresponds to the normalized number of shortest paths connecting dyads which pass through that node; the higher the coefficient, the more important that node is in ensuring the connectedness of the rest of the network. This quantity tells us if some words behave like unavoidable waypoints on the path associating one word to another.

3.2.3 *Variable correlations.* An important question arises concerning the possible correlations between all the variables we use.

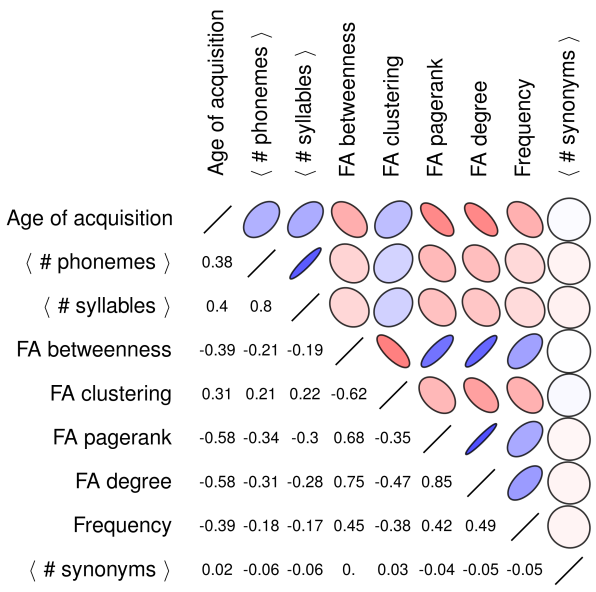


Fig. 1: Spearman correlations in the initial set of features

[Correlation talk needs to be redone for new set of features]

Age of acquisition is a key variable which appears as a usual suspect in psycholinguistic studies and is also usually correlated to many of the other variables. This relates to an ongoing debate suggesting that age of acquisition encodes a variety of phenomena, difficult to disentangle from more specific phenomena which could be captured by more independent variables [Citation needed]. Here however, as can be seen in Figure 1, age of acquisition has a relatively low correlation to the other variables (absolute value not above 0.42 if we exclude the centrality measures), leading us to keep the variable in the rest of the analysis.

[3]The CMU Pronouncing Dictionary is included in the NTLK package [Bird et al. 2009], the natural language processing toolkit we used for the analysis.

Number of phonemes and number of syllables naturally exhibit a strong linear correlation (0.83). The analysis showed a better prediction effect of number of phonemes over number of syllables, which is consistent with Nickels and Howard [2004], and we therefore chose to focus the presented results on the former only.

Frequency and number of meanings both have relatively low levels of correlation to the other variables; we therefore also keep them in the rest of the analysis.

Network properties, on the other hand, are strongly dependent on one another. As mentioned earlier, word degree and word pagerank have a very strong correlation (0.89) and, degree being generally more correlated to other variables, we chose to remove this variable from the results presented. Finally betweenness centrality also exhibits strong correlation levels to the other network properties (0.62, 0.64, and 0.72 in absolute value), leading us to drop this final feature due to its redundancy.

The final set of variables we consider, as well as their cross-correlations, can be seen in Figure 2.

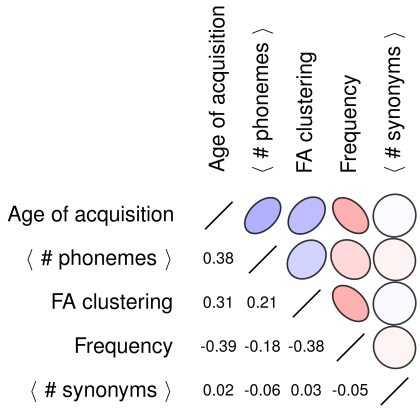


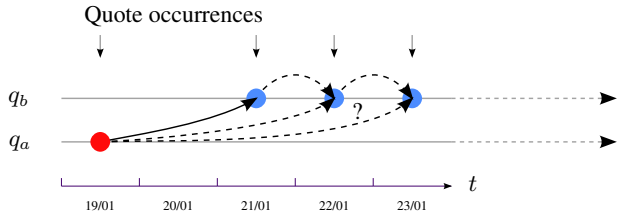
Fig. 2: Spearman correlations in the filtered set of features

3.3 Temporal binning

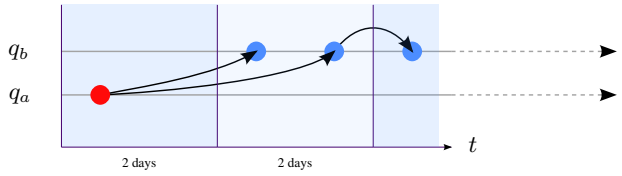
The data we use presents an additional challenge: each set of quotations bears no explicit information either about the authoritative original quotation, or about the source quotation(s) each author inspired himself from when creating a new post and reproducing (possibly altering) those sources. Quote-to-quote transformations, and much less substitutions, are therefore not explicitly encoded in the dataset.

We face an inference problem where, given all quotations and their occurrence timestamps, we should estimate which was the originating quotation for each instance of each quotation. We therefore model the underlying quotation selection process by making a few additional assumptions which let us define quote-to-quote substitutions from the available data. The main issue at hand is deciding whether a later occurrence is a strict copy of an earlier occurrence, or a substitution of an even earlier occurrence, or perhaps even a substitution or copy from quotes appearing outside the dataset, i.e. from a source external to the data collection perimeter.

Let us give an example: say the quotation "These accusations are false and **absurd**" (q_a) appears in a blog on January 19, and the slightly different quotation "These accusations are false and **incoherent**" (q_b) appears in other blogs on the 21st, 22nd and 23rd of January. If q_a was sufficiently prominent when q_b first appeared, we can safely assume that the first author of q_b on the 21st based himself on q_a as is shown in Figure 3a. But what about the second and third occurrences of q_b, on the 22nd and 23rd? Should we consider them to be substitutions based on q_a or accurate reproductions of the previous occurrences of q_b? (Options shown in Figure 3a.)



(a) Possible paths from occurrence to occurrence



(b) Binned quotation family

Fig. 3: Temporal binning of quotation families

To settle this question we bin the quote occurrences into fixed *time bags* spanning Δt days (2 days in the implementation), each one representing a unit of time evolution. Then when a quotation q appears in time bag n, it is counted as a substitution from each quote q^* in the preceding time bag $(n-1)$ from which it differs by only one word. If no quote in the preceding time bag can qualify as a source in a substitution (i.e. q differs from all the quotes in the preceding time bag by more than one word), the occurrence of q is not considered to be an instance of substitution. Such a model defines how many times quote occurrences can be counted as substitutions: in Figure 3b, occurrences of q_b on the 21st and 22nd are counted as substitutions, whereas the occurrence on the 23rd is not.

The assumptions embedded in this model are only a subset of a wider set of possibilities, each leading to alternative substitution inferences.[4] These various flavours of an ideal substitution detection model essentially change whether occurrences are considered as substitutions from another quote, repetitions of the original quote, or introduction of information external to the dataset. We identified and implemented eleven other such models, and they all yielded essentially the same results.

4. MACROSCOPIC EVOLUTION OF QUOTATION FAMILIES

We first examine the evolution of quotation families under the repeated action of substitutions. Our goal in this step is to identify

[4] In particular, time can be sliced into bins to build fixed time bags as is done here, or kept fine-grained by using sliding time bags.

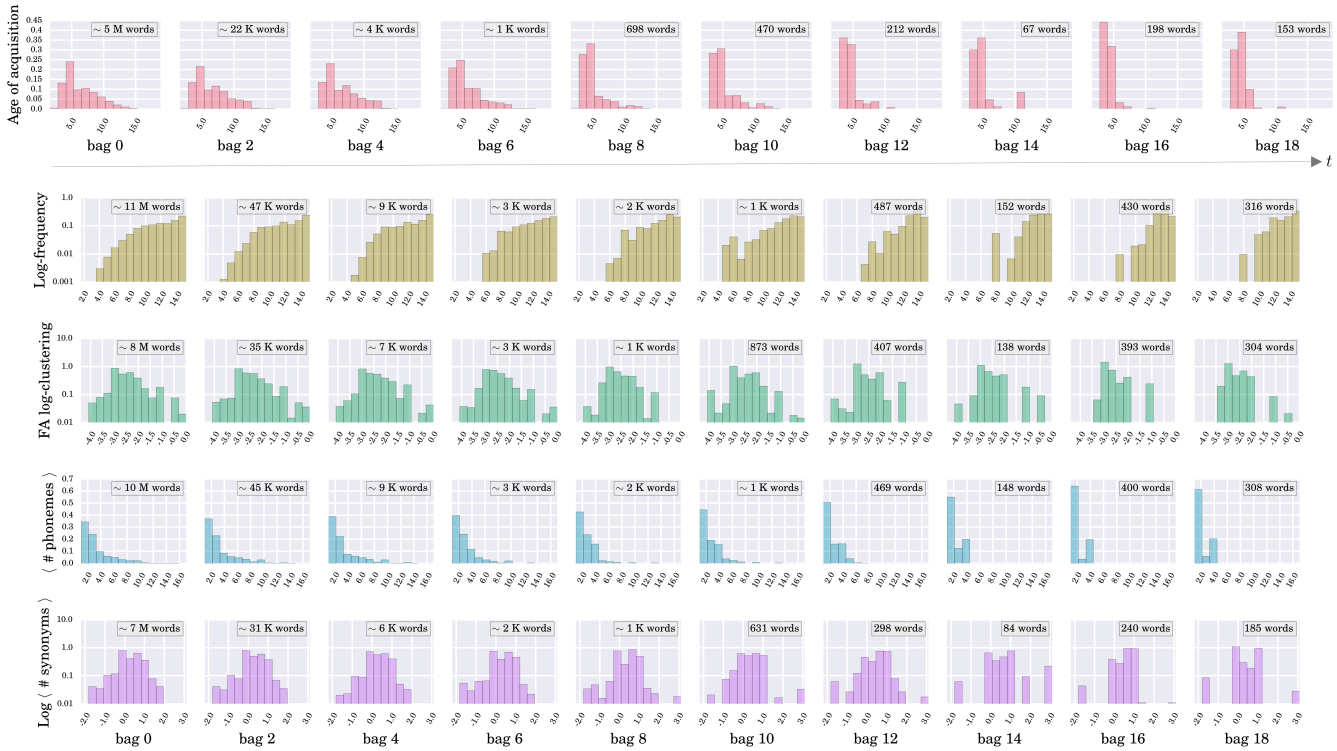


Fig. 4: **Feature distribution evolution:** evolution of the distribution of feature values in substitution chains over successive 2-day time bags (bags 0 to 18, i.e. days 0 to 37). The legends indicate the number of words left in each time bag; these decrease exponentially since only a fraction of the quotes undergo substitution at each step. After a period of time, each feature becomes concentrated in a specific range of its own.

the long-term effect of cognitive bias over the lifetimes of quotation families and thus over the framing of public information.

To do so we compute the distribution of word features for each time bag $\mathcal{B}_n$ of each quotation family, and sum those distributions over all quotation families. This yields a distribution of feature values for each n, which is the simplest possible view of the state of an average quotation family in its n-th time bag, i.e. after $n \times \Delta t$ days.[5] Such a computation, based solely on the binning of quotation families, makes no assumptions on the way quotations undergo substitutions over time.

The raw distributions built with this computation are stationary. That is, the substitutions on quotations over time have no global effect on quotation families, either because external quotations are continuously fed into the family and compensate for the effect of substitutions, or because substitutions operate on a marginal portion of the quotation families, or both.

To narrow this view to the specific effects of cognitive bias we consider substitution *chains*: using the substitution detection model described in the previous section, we filter time bags to include only new quotes produced by substitutions themselves based on quotes produced by substitutions, and so on recursively. The first time bag is therefore untouched, the second contains the quotations produced by substitutions from the first, the third by substitutions from the second, and so on (see Figure 5 for an illustration of this process). As a result, the number of observed words drastically decreases across time, yet it unambiguously focuses on successive mutations.

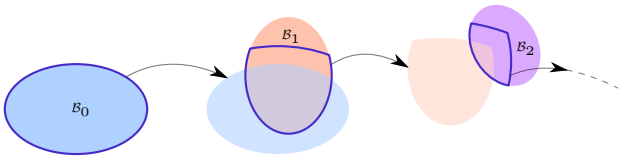


Fig. 5: XXXXX

We thus observe only the repeatedly-affected portion of the quotation families, and obtain a simple view of the longitudinal effect of cognitive bias, i.e. how quote features are evolving in the long term and, perhaps, converging towards specific attractors.

[Comment results and transition to micro process]

5. MICROSCOPIC EVOLUTION: THE SUBSTITUTION PROCESS

We then focus on the individual substitution process at work when authors transform quotations, by examining the features of the substituted and substituting words in each substitution.

To do so we first extend the substitution detection model used in the previous section to reduce possible false positives. Indeed, when in the previous section we were forced to detect substitutions with a very permissive model to make sure chains of substitutions

[5] For consistency, if we do this for n going up to N, we only include quotation families that span long enough to have at least N time bags.

could be extracted (at the expense of possible false positives, the only effect of this being additional noise in the results, making our conclusions more conservative), here we can afford to detect substitutions more precisely. We thus add the constraint that substitutions can only stem from the most frequent quote in the preceding time bag. Since it is fairly certain that authors will have come upon the most frequent quote in any given time bag (*v.* all the other quotes in that time bag), restricting substitutions to the most frequent quote makes sure we only detect substitutions that really occurred, greatly reducing the number of false positives.

Let us illustrate this "majority" rule by going back to the example described in Section 3.3 and extending it, in Figure 6. In time bag 3, q_b now holds the majority and is considered the unique basis for the last occurrence of q_a (in time bag 4). This is despite the fact that q_a also appears in time bag 3 alongside q_b, and despite it having appeared earlier at the very beginning of the quotation family (indeed in the situation shown in Figure 6, this seems to be the most likely scenario). Conversely, if q_a had been the most frequent quote in time bag 3, the last occurrence of q_a in time bag 4 would have been considered a faithful copy of the occurrence in time bag 3.

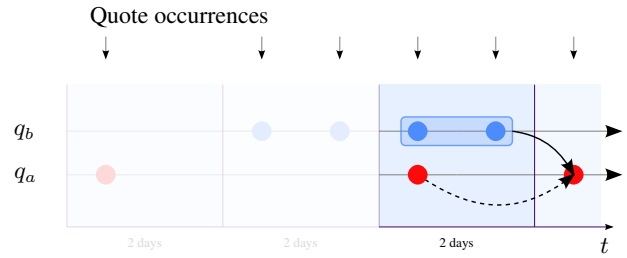


Fig. 6: XXXXX

With the substitution detection model thus refined, we build two main observables for each word feature. First, we measure the susceptibility for words to be the source of a substitution, knowing that there has been a variation, in order to show which semantic features are the most likely to attract a substitution under this condition. Second, we measure the variation of word feature over a substitution, looking at the variation of a given feature between start and arrival words.

Note that since we only consider substitutions and not faithful copies, we measure the features of an alteration *knowing that there has been an alteration*, and we do not take invariant quotations into account. Indeed, in the first case we know there has been a human reformulation, whereas in the second case it is impossible to know whether there has been perfect human reformulation or simply digital copy-pasting of a source ("CTRL-C/CTRL-V").

5.1 Susceptibility

For a given feature ϕ, the protocol lets us compute substitution *susceptibilities* for each feature value f. We say that a word is *substitutable* if it appears in a quote which undergoes a substitution, whether that substitution operates on the considered word or on another. Word substitution susceptibility is computed as the ratio of the number s_w of times a word is substituted to the number p_w of times that word appears in a substitutable position, i.e. s_w/p_w.

Now averaging over all words such that $\phi(w) = f$ (only taking into account words that are substituted at least once), we obtain the mean susceptibility for the feature value f: [6]

$$\sigma_\phi(f) = \left\langle \frac{s_w}{p_w} \right\rangle_{\{w|\phi(w)=f\}}$$

This measure focuses on the selection of start words involved in substitutions, measuring the effect of features at the moment preceding the substitution when it is not yet known which word in the quotation – if any – will be substituted.

5.2 Alteration

Next, we measure how a word w's feature varies as w is substituted by w', i.e. $\phi(w') - \phi(w)$. Averaging this value over all start words such that $\phi(w) = f$ yields the mean variation for that feature value f:

$$\Delta_\phi(f) = \langle \phi(w') - \phi(w) \rangle_{\{(w,w')|\phi(w)=f\}}$$

We introduce a null hypothesis $\mathcal{H}_0$ to compare the actual variation of a word's feature to its expected variation, assuming the arrival word w'_0 had been chosen randomly from the pool of free association words. The new quantity under $\mathcal{H}_0$ is: [7]

$$\Delta^0_\phi(f) = \langle \phi(w'_0) - \phi(w) \rangle_{\{(w,w'_0)|\phi(w)=f\}}$$

We also considered an alternative null hypothesis, denoted $\mathcal{H}_{00}$, where the arrival word is chosen randomly *among immediate synonyms of the start word*, i.e. an arrival word chosen among semantically plausible though still random words.[8] The results were not qualitatively changed with this second null hypothesis, so for the sake of clarity we chose to present the results using the simpler $\mathcal{H}_0$.

Using this method we obtain the mean variation of feature for each start feature value, and can compare the variations to a situation where arrival words are chosen randomly. This gives us a fine-grained view of how word features evolve upon substitution.

5.3 Results

[Comment results]

6. CONCLUSION

Main contributions:

—large *in vivo* psycholinguistics experiment, emphasizing the importance of the semantic network structure (Wordnet: Pagerank, degree, clustering[, distance?]), finely describing the impact of classically-influent psycholinguistic variables (*aoa*, number of phonemes, etc.);

—new in the sense that it does not focus on ease of recall but rather bias of substitution, and that in this respect it not only provides a finer description of the bias but also corresponds to an "input-output" reformulation couple describing the joint properties of (substituted→substituting) terms.

[6] To avoid any auto-correlation effect due to the number of substitutions in a cluster (possibly leading to an overly optimistic estimation of confidence intervals), we first average substitutions over each cluster, by considering the average of arrival word features for a given start word. Indeed, substitutions occurring in the same cluster are likely not statistically independent.

[7] Note that $\phi(w'_0)$ is in fact a constant in this averaging, since by definition w'_0 does not depend on w.

[8] In this case w'_{00} does depend on w.

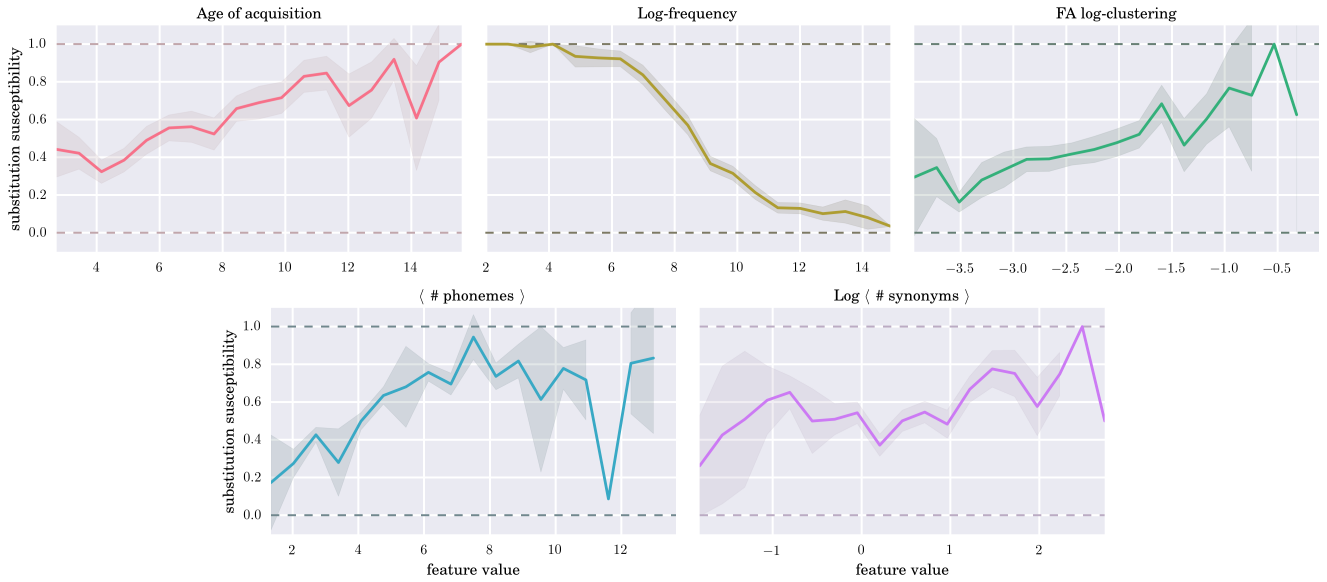


Fig. 7: **Substitution susceptibility:** average susceptibility to substitution *v.* average feature value of a candidate word for substitution, with 95% asymptotic confidence intervals. Each feature exhibits a specific and significant pattern favouring either high- or low-valued words for substitution.

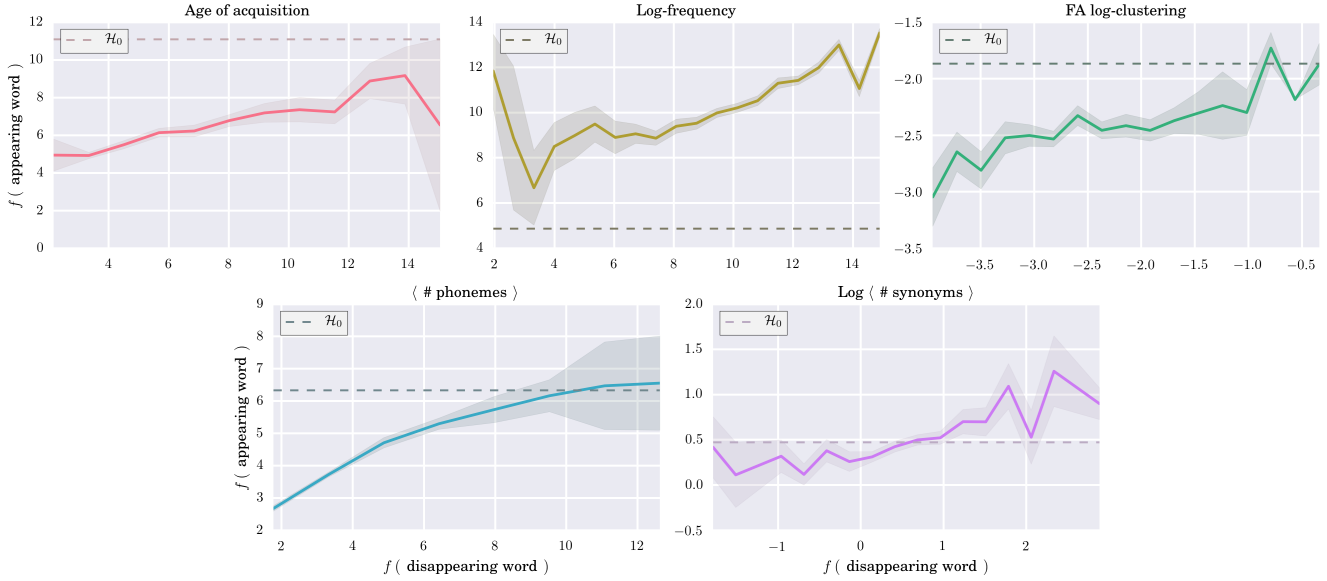


Fig. 8: **Feature variation upon substitution:** average feature of the appearing word minus $\mathcal{H}_0$ *v.* average feature of the disappearing word in a substitution, with 95% asymptotic confidence intervals. The overall position of the curve with respect to $y = 0$ indicates the direction of the cognitive bias. The fact that all the curves have slopes smaller than 1 means that the substitution operation is contractile on average: each feature will converge towards its own specific asymptotic range, which is consistent with the evolution observed in Figure 4.

—beyond that, provide the first bricks of an empirical *fitness landscape* for the epidemiology of representations

Acknowledgements

We are warmly grateful to Ana Sofia Morais for her precious feedback and advice on this research.

REFERENCES

Scott Atran. Théorie cognitive de la culture. *L'Homme*, 166(2): 107–143, 2003.

Robert Aunger, editor. *Darwinizing Culture: The Status of Memetics as a Science*. Oxford University Press, Oxford, 2000.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Incorporated, 2009.

Maurice Bloch. A well-disposed social anthropologist's problems with memes. In Robert Aunger, editor, *Darwinizing Culture: The Status of Memetics as a Science*, chapter 10, pages 189–203. Oxford University Press, 2000.

Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, 1985.

Kit Ying Chan and Michael S Vitevitch. Network structure influences speech production. *Cogn Sci*, 34(4):685–97, 2010. doi: {10.1111/j.1551-6709.2010.01100.x}.

Nicolas Claidière and Dan Sperber. The role of cultural attraction in cultural evolution. *Journal of Cognition and Culture*, 7(1): 89–111, 2007. doi: {10.1163/156853707X171829}.

Jean-Philippe Cointet and Camille Roth. Socio-semantic Dynamics in a Blog Network. In *2009 International Conference on Computational Science and Engineering*, pages 114–121, 2009. doi: {10.1109/CSE.2009.105}.

Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

Richard Dawkins. *The Selfish Gene*, chapter 11, pages 189–201. Oxford University Press, 1976. "Memes: The New Replicator".

Paul R Ehrlich and Simon A Levin. The evolution of norms. *PLoS Biol.*, 3(6):e194, 2005. doi: {10.1371/journal.pbio.0030194}.

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

Thomas L Griffiths, Mark Steyvers, and Alana Firl. Google and the mind: predicting fluency with PageRank. *Psychol Sci*, 18 (12):1069–76, 2007. doi: {10.1111/j.1467-9280.2007.02027.x}.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information Diffusion Through Blogspace. In *Proceedings of the 13th International World Wide Web Conference (WWW'04)*, pages 491–501, 2004.

Adam Kuper. If memes are the answer, what is the question? In Robert Aunger, editor, *Darwinizing Culture: The Status of Memetics as a Science*, pages 180–193. Oxford University Press, 2000.

V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert. Age-of-acquisition ratings for 30 thousand english words. *Behavior Research Methods*, 2012.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the Dynamics of the News Cycle. *KKD'09*, (June 28-July 1):497–505, 2009a.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. MemeTracker: tracking news phrase over the web. `http://memetracker.org/`, 2009b. Retrieved on August 19, 2012.

D.L. Nelson, C.L. McEvoy, and T.A. Schreiber. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407, 2004.

Lyndsey Nickels and David Howard. Dissociating effects of number of phonemes, number of syllables, and syllabic complexity on word production in aphasia: It's the number of phonemes that counts. *Cognitive Neuropsychology*, 21(1):57–78, 2004. doi: 10.1080/02643290342000122. URL `http://www.tandfonline.com/doi/abs/10.1080/02643290342000122`. PMID: 21038191.

Elisa Omodei, Thierry Poibeau, and Jean-Philippe Cointet. Multi-level modeling of quotation families morphogenesis. In *Proc. ASE/IEEE 4th Intl. Conf. on Social Computing "SocialCom 2012"*, 2012.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, November 1999. URL {`http://ilpubs.stanford.edu:8090/422/`}. Previous number = SIDL-WP-1999-0120.

A. Rey, A.M. Jacobs, F. Schmidt-Weigand, and J.C. Ziegler. A phoneme effect in visual word recognition. *Cognition*, 68(3): B71–B80, 1998.

Beatrice Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*, 3rd revision, 2nd printing edition, 1990.

Matthew Simmons, Lada Adamic, and Eytan Adar. Memes Online: Extracted, Subtracted, Injected, and Recollected. In Nicolas Nicolov and James G. Shanahan, editors, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2011.

Dan Sperber. *Explaining Culture: A Naturalistic Approach*. Oxford: Blackwell Publishers, 1996.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

RL Weide. The cmu pronunciation dictionary, release 0.6, 1998.

WordNet. Princeton University "About WordNet.". `http://wordnet.princeton.edu`, 2010. Retrieved on August 19, 2012.

A.P. Yonelinas. The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, 46(3): 441–517, 2002.

J.D. Zevin and M.S. Seidenberg. Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, 47 (1):1–29, 2002.

Garbage / Results

We may see that $\mathcal{H}_0$ and $\mathcal{H}_{00}$ are slightly translated yet not qualitatively different.

Results for global evolution of features:

1/ no effect on global distribution => either continuous reinjection with base distribution, or substitutions are marginal (we can't separate those two with the information we have), or they are on marginal quotations (but no, because in that case we would see a difference when dropping quote frequency, which we don't). So the global distribution is stationary.

2/ we do see an effect when looking at chains, i.e. restricting the source quotes to be the product from previous step.

3/ also no effect when allowing more complex transformations than 1-word substitutions => they are either reinjections, or the process is not as constrained as 1-word substitutions.

6.1 Location of arrival words

Where in the network are arrival words, are they related to the origin word, when relying on a given semantic network? (In other words, would a semantic word be a good predictor of the arrival of wrds?). **[*tout un blabla sur les distances dans les substitutions.*]**

[*is this really necessary?*]

6.2 Known psycholinguistic effects

We first looked at well-known (and well-studied) psycholinguistic features, and doing so we were able to determine which of the two alternative hypotheses presented in the introduction is valid; the features we examined were age of acquisition norms[9] and number of phonemes[10]. **[*repetition*]** First, by looking at substitution susceptibilities for those features, we can see that words with lower age of acquisition (figure 0**??**) and words with lower number of phonemes (figure 0**??**) are more likely to be substituted than words with higher values for either of those features. This shows us that, although words learned later in development (as well as words with larger number of phonemes) may be cognitively harder to recall, they seem to be substituted according to hypothesis [X]: their relevance and specificness makes them less susceptible to change, whereas words learned earlier or with smaller number of phonemes are more easily substituted.

Now looking at what type of arrival words are selected upon substitution, we can see that both words learned earlier in development (figure 0**??**) as well as words with lower number of phonemes (figure 0**??**) are substituted for words with roughly the same feature values (they tend to augment a little, but stay way below the values expected under $\mathcal{H}_0$ or $\mathcal{H}_{00}$). Conversely, words learned later in development and words with higher number of phonemes, when substituted, go towards words with lower feature values, but no lower than what the null hypotheses predict. This shows again that words learned early on are easily interchangeable for other words learned at about the same age. On the other hand, it seems the cognitive system puts little constraint on the arrival word when substituting a word learned later in development (that is, those words are rarely substituted, but when a substitution does occur the arrival word does not show constraints related to the features we examined). We also tested these results for dependence on the grammatical category of words, i.e. POS tags, and found no effect.

Though not originating from the better-known image-naming or controlled lexical-retrieving tasks, these results shed a new light on the strength of the effect of these features depending on their exact value, and on how these features behave in the context of *substitution* and not only free recall.

6.3 Epidemiological setting

Secondly, and coming back to our primary goal of providing a first empirical testing means for epidemiological models of culture, we added to these two features a number of more abstract properties computed from the Free Association norms used by Griffiths et al. [2007]: namely, the PageRank, betweenness coefficient and clustering coefficient of the words. **[*repetition*]**

These new features are classical measures for network-related matters, but – to our knowledge – have seldom been applied to word characterization. They are only a few among many other features that can be used to characterize words based on the network they form (be it the Free Association norms network, another semantic network, or even a phonological network). We consider the complete set of these features (which includes the two mentioned in the previous section, the new network-based ones, as well as any other network-based feature one can compute) as traits characterizing words in an empirical epidemiological evolution of quotes. The benefit we get from measuring the evolution of such features is in how we can use this information to falsify epidemiological culture models. Indeed, in such a setting the information on how a feature is modified upon substitution (detailed to the individual feature value) is in fact a *fitness landscape* for that particular feature. This fitness landscape can in turn be re-injected into epidemiological models to see if those models account for the empirical distribution of quotes, their life-span and relative success. The data for PageRank, betweenness coefficient and clustering coefficient are shown in figures 0**??**, 0**??** and 0**??**.

[*some talk about what this confirms in Griffith's work*]

Moreover, the substitution susceptibility corresponds to a *mutation probability* in this setting (shown in figures 0**??** for PageRank, 0**??** for betweennes coefficients, and 0**??** for clustering coefficient). To sum up, the data obtained here are empirical measures of (some of) the parameters of epidemiological models of cultural evolution, i.e. essential information to allow empirical testing of these models. If quotes, as a first example of cultural representation, do indeed follow epidemiological rules in their evolution, we may be able to empirically prove the validity – or falsify – the existing models of cultural evolution.

[*this is very rough and brute-like. It needs to be better presented and further developed.*]

[*A big question to be addressed is the "why": what underlying aspect of words with these features make them words to be substituted or not, and why does the brain do that? How was that selected for in evolution?*]

[9] The Age-of-Acquisition data is obtained from **?**.

[10] The number of phonemes are obtained from the CMU Pronouncing Dictionary for U.S. English [Weide 1998], included in the NLTK package [Bird et al. 2009]. We note that these two measures are likely to be linked, since words with larger numbers of phonemes are likely to be learned later in development.