

How do we copy and paste?

The semantic drift of quotations in blogspace

SÉBASTIEN LERIQUE and CAMILLE ROTH

ABSTRACT

We describe reformulation processes within a large distributed system such as blogspace; showing how some specific features of public representations may be altered by bloggers when they freely reproduce them. To deal with robust and simple cultural representations, we focus on the evolution of quotations. In particular, we uncover some of the semantic and structural characteristics of individual words and the substitutions they undergo. Our work amounts to a large *in vivo* experiment where we appraise the impact of classically-influent psycholinguistic variables in the accuracy of the reproduction. We show that all variables remarkably exhibit a single attractor and are generally contractile. Even though the observed convergence patterns only partially explain quotation evolution, we shed light on a class of phenomena which are prone to constitute a key element of a broader empirically-grounded, attractor-based theory of cultural evolution.

1. INTRODUCTION

The understanding of knowledge transmission mechanisms has led to a sizable literature in the recent past, spanning over numerous research fields ranging from cultural anthropology to social network analysis and complex systems modeling; from social cognition to data mining. These works are diversely labeled as studies on “opinion dynamics”, “cultural evolution”, or “information diffusion” and, for the most part, investigate phenomena pertaining to both cognitive science and social science, both at the individual and social levels.

Broadly speaking, we may distinguish two main research streams, depending on whether the focus lies on cognitive processes or on social dynamics. A first stream is largely structured around cultural anthropology and essentially addresses cultural similarity, diversity and its evolution. It features several theories mixing social and individual cognition including, to cite a few, the debated “memetic” program initiated by Dawkins [1976] (for which the collection of works by Aunger [2000] provides a solid overview), the development of evolutionary models of norms (see for instance Ehrlich and Levin [2005]) following the seminal work of Boyd and Richerson [1985]); or the “cultural epidemiology” program proposed by Sperber [1996], which links the concept of mental representation to the concept of public representation (the latter being the counterpart of the former outside of the brain, i.e. in all sorts of cultural artifacts: texts, utterances, etc.).

One of the core claims of this literature consists in emphasizing that not all knowledge is equally fit for being reproduced, although the various approaches have a different take on how exactly this notion of fitness should be operationalized. Sperber’s cultural epidemiology classically opposes Dawkins’ memetics by insisting 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 that this conceptual evolution can be appraised through the notion of “cultural attractor”, seen as the attraction domain of an underlying

socio-semantic dynamical system.¹ Despite some recent modeling attempts (e.g. Claidière and Sperber [2007]), the development of quantitative measurements relying on the concept of cultural attractors has remained a relatively hard task and, to our knowledge, this hypothesis has not yet been empirically analyzed in an extensive manner.

Another research stream deals with rather macroscopic studies of knowledge diffusion. Here, one of the focal points is that not all knowledge gets propagated identically along the same routes, within the same communities, at the same pace. The various approaches usually feature a minimalistic description of cognitive processes, strongly reminiscent of biological epidemiology (a single, atomic piece of information may or may not be adopted by each individual). This research program nonetheless exhibits a particularly interesting empirical track record — largely owing to a recent avalanche of observable *in vivo* data which, for a good decade now, have mainly come from online interaction contexts. While these information trails are not records of “physical” inter-individual interactions (in the sense of “real life” interactions), they still constitute a wealth of observations on the dynamics of public — albeit online — representations. Some authors could describe for instance the propagation of cultural artifacts 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 latter studies are at the interface between data mining, complex systems and quantitative sociology (first and foremost social network analysis) and are relatively remote from cognitive science; for a significant part, they rely rather marginally on specific social science theories. They nonetheless show us the added value of using these rapidly growing records towards radically improving the empirical understanding of (individual-level) cultural evolution processes.

Stepping back, we thus observe a gap between, on one side, empirical studies of diffusion dynamics in social systems and, on the other side, more theoretical works focused on knowledge transformation processes. Our research lies at the intersection of these two programs, aiming to shed light on micro-level information transformation by leveraging the empirical wealth of (*in vivo*) social diffusion phenomena. More precisely, we hope to describe reformulation processes within a large distributed system such as blogspace; showing how some specific types and features of public representations may be altered by bloggers when they freely reproduce them.

We focus on simple linguistic modifications, thereby connecting our research to the broader psycholinguistic literature. To deal with robust and simple cultural representations, we paid attention to the evolution of quotations. While these verbatim public rep-

¹Works such as Atran [2003] argue that this approach is anthropologically better suited than memetics, and some of the main issues in this debate are further detailed by Kuper [2000] and Bloch [2000].

representations should in theory not suffer any alterations when they are produced anew (as opposed to more elaborate expressions and opinions, not identified as quoted utterances), empirical observation shows that they are occasionally 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. In a way using this type of data is equivalent to a large-scale psycholinguistic experiment and at the same time constitutes a first step towards building empirically realistic models of cultural evolution.

The next section (Sec. 2) describes our hypotheses along with the relevant state-of-the-art on this psycholinguistic 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. 4 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. 5.

2. RELATED WORK

The practical 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 (see Claidière and Sperber [2007]; Claidière et al. [2014] as well as a hybrid empirical-theoretical protocol in MacCallum et al. [2012]). Among the empirical approaches, 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]. One of the main observations in these works is that even for quotations, a type of public representation that should be among the most stable, it is still possible to witness significant transformations. They essentially examine the effect of some properties of the quotation source (e.g. news outlet vs. blog) or of the surrounding public space (e.g. quotation frequency in the corpus). Some diffusion-transformation models have been proposed, yet the very cognitive features which may determine or, at least, influence these transformations, are overlooked; which may appear to be relatively unsatisfying from a cognitive viewpoint.

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.

Less classical linguistic variables, based on the study of semantic network properties, have recently started to be used, in the context of connectionism and its normative processual models (see for instance Collins and Loftus [1975]). Let us mention four interesting studies on that matter, which demonstrate in a strictly *in vitro* framework and at the vocabulary level that properties computed on a word network are important factors for the cognitive processes and reproduction of those words. First, Griffiths et al. [2007] analyze a task where subjects are asked to name the first word which comes to their mind when they are presented with a random letter from the alphabet. The authors show that there ex-

ists 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. Austerweil et al. [2012] further develop this idea by showing that random walk on such a semantic network, i.e. the exact process measured by the pagerank index, gives a parsimonious account of some semantic retrieval effects (namely, related items being retrieved together). A third psycholinguistic study by Chan and Vitevitch [2010] shows, in a picture-naming task, that words are produced faster and with fewer mistakes when they have a lower clustering coefficient in an underlying phonological network (which, again, is defined from external phonological data). Nelson et al. [2013], finally, show the importance of clustering coefficient in a semantic network by studying the role it plays in a variety of recall and recognition tasks (extralist and intralist cuing, single item recognition, and primed free association).

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. On the other hand, when the task consists in recognizing a specific item in a list, “awkward” words are actually more easily remembered, possibly as they are more informative and plausibly more discernible (see again Yonelinas [2002] for a review). 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. We also aim here at shedding some light on this debate, considering oddness as a dimension of the purported fitness of utterances.

3. METHODS

Quotations appeared to be a perfect candidate to propose a first *in vivo* measure of low-level cognitive bias in a reformulation task. First, they are usually cleanly delimited by quotation marks 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.

We could therefore examine the individual substitution process at work when authors transform quotations, by examining the features of the substituted and substituting words in each substitution. To keep the analysis palatable, we focused on quotation transformations consisting 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 next subsections describe the dataset and measures we used to assess this cognitive bias.

3.1 In vivo utterances

We used a quotation dataset collected by Leskovec et al. [2009a], large enough to lend itself to statistical analysis. This dataset con-

sists 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. For instance,

The Bank of England said, “these operations are designed to address funding pressures over quarter-end.”

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). We find for example the following variation of the above quote:

“these operations are **intended** to address funding pressures over quarter-end.”

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.

Manual inspection of this dataset revealed that it contains a significant number of everyday language quotations (such as “it was much better than I expected”, “did that just happen”, as well as many simple expletive-based sentences). Their presence is largely due to random variations around a casual expression, while we are interested in transformations of news-related quotes causally linked to an original, identifiable utterance. We thus filter out all clusters lasting more than 80 days, and all quotes less than 5 words long (as well as quotes not written in English). We eventually keep 19,621 clusters (out of 71,568; i.e. 27.4%), containing a total of 60,246 unique quotes (out of 310,457; i.e. 19.4%) making up ~838k occurrences (out of 8.12m, i.e. 10.3%). The significantly larger loss in occurrences indicates that, on average, the clusters we lose contain more occurrences than those we keep, which is expected for everyday language utterances.³

3.2 Word-level measures

3.2.1 Psycholinguistic indices. We first introduce some of the most classical psycholinguistic measures on words.

- Word frequency:** the frequency at which words appear in our dataset, known to be relevant for both recognition and recall [Gregg 1976],
- Age of Acquisition:** the average age at which words are learned (obtained from Kuperman et al. [2012]), known to have different effects than word frequency [Morrison and Ellis 1995; Dewhurst et al. 1998],

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

³While we lose real event-related utterances such as “the terrorists have used automatic weapons and in some places grenades have been lobbed the encounters are still going on and we are trying to overpower them”, we contend that the use of the unfiltered dataset would strongly jeopardize results related to linguistic evolution or information diffusion. We manually checked on a subsample of clusters that a large proportion of remaining quotes seem to be relevant, while filtered quotes are essentially irrelevant.

- The average **Number of Phonemes** and **Number of Syllables** for all pronunciations of a word (obtained from the Carnegie Mellon University Pronouncing Dictionary [Weide 1998])⁴ as a proxy to word production cost,

- The average **Number of Synonyms** for all meanings of a word (obtained from WordNet [2010]) as an *a priori* indicator of how easy it would be to replace a word.

The number of synonyms is related to a notion of the word connectivity in a semantic network. To go a bit further in this direction, we appraise the possible role of network-based variables which have received special attention in the recent related literature, following the blooming interest in networks from many disciplines over the last decade.

We relied on the “free association” (FA) norms collected by Nelson et al. [2004] which naturally embed information on the idea association process underlying transformation of quotations. FA norms record the words that come to mind when someone is presented with a given cue. As Nelson et al. [2004] 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.” Following Griffiths et al. [2007], we first build a directed unweighted network based on association norms, where nodes are words and edges are directed from cue to target word whenever the considered target word was produced in response to the considered cue word. This network is of particular interest since it measures the *in-vitro forced-choice* version of a substitution whereas the data we analyze is the *in-vivo spontaneous* version of what we otherwise hypothesize to be the same process.

Three standard network-based measures are to be used on the FA network:

- Degree centrality**, measured by the number of cues for which a given word is triggered as a target, and a corresponding generalized measure, node *pagerank* [Page et al. 1999], which has already been used on the FA network by Griffiths et al. [2007]. In the present case these two polysemy-related measures are quasi-perfectly correlated.
- Betweenness centrality**, another measure of node centrality describing the extent to which a node connects otherwise remote areas of the network [Freeman 1977]. This quantity tells us if some words behave like unavoidable waypoints on association chains connecting one word to another.
- Clustering coefficient**, which measures the extent to which a node belongs to a local aggregate of tightly connected nodes [Watts and Strogatz 1998], computed on the undirected version of the FA network.⁵ This tells us if a word belongs more or less to a local aggregate of equivalent words (from a “free association” point of view).

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

The number of phonemes and the number of syllables naturally exhibit a strong linear correlation (0.8). Our analysis showed clearer results with number of phonemes over number of syllables, which is consistent with Nickels and Howard [2004], and we therefore chose to only present results for the former.

⁴The CMU Pronouncing Dictionary is included in the NTLK package [Bird et al. 2009], the natural language processing toolkit we used for the analysis.

⁵The Clustering coefficient is formally defined as the ratio between the number of actual vs. possible edges between a node’s neighbors.

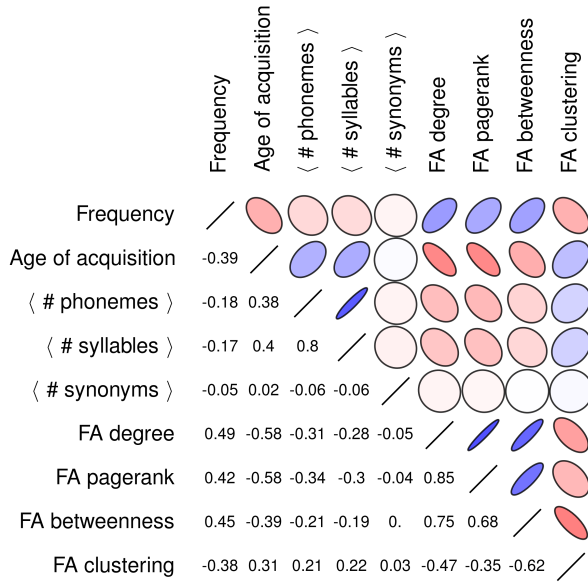


Fig. 1: Spearman correlations in the initial set of features

Age of acquisition is a key variable which appears as a usual suspect in psycholinguistic studies. Despite it being usually difficult to disentangle from many of the other variables, it is known to have independent effects, which is consistent with what we see on Figure 1: age of acquisition has a limited correlation to the other variables (absolute value not above 0.39 if we exclude the number of syllables and the network properties), leading us to keep the variable in the rest of the analysis.

Frequency and number of synonyms both have relatively low levels of correlation to the other variables (excluding again the network properties); we therefore also keep them in the rest of the analysis.

Network centrality properties, on the other hand, are strongly dependent on one another. As mentioned earlier, degree centrality and pagerank have a very strong correlation (0.85), and are also redundant with betweenness centrality (with correlation levels at 0.75 and 0.68 respectively). Furthermore, the three variables are also strongly related to age of acquisition, which leads us to keep the latter as the sole indicator for centrality. This may trigger a chicken-and-egg issue where a strong centrality may be due, or be the result, of an early age of acquisition; in any case, the age of acquisition seems to partially capture centrality-based network properties.

Conversely, clustering coefficient exhibits low correlation levels with all the variables we kept (maximum absolute value 0.38), leading us to include it in the rest of the analysis.

The final set of variables we consider, as well as their cross-correlations, can be seen in Figure 2.⁶

⁶Note that feature values stem from different datasets which do not always encode the same words. Indeed, we have data on frequency for about 22.6k words, on age of acquisition for 30.1k words, on number of phonemes for 123.4k words, number of synonyms 111.2k, and clustering coefficient 5.7k words. Quite often then, not all features are available for all words in our dataset; however this is not problematic since the analysis is done on a per-feature basis, and not all words need be encoded in all features.

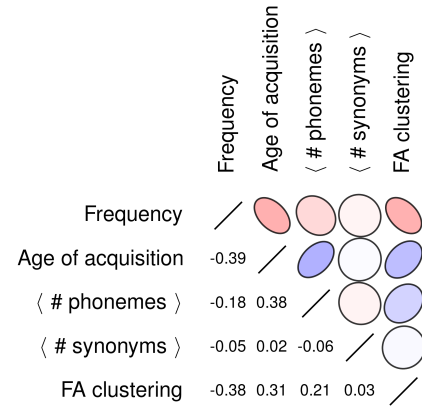


Fig. 2: Spearman correlations in the filtered set of features

3.3 Substitution model

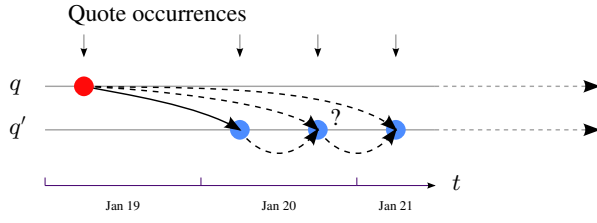
We finally need a substitution detection model, for the utterance data we use presents a challenge: quote-to-quote transformations, and much less substitutions, are not explicitly encoded in the dataset. More precisely, each set of quotations bears no explicit information about either the authoritative original quotation, or the source quotation(s) each author relied on when creating a new post and reproducing (and possibly altering) that source. We thus 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. The main issue 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.

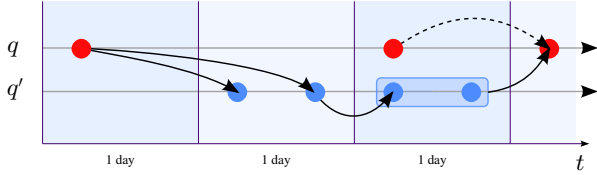
Let us give an example: say the quotation “These accusations are false and **absurd**” (q) appears in a blog on January 19, and the slightly different quotation “These accusations are false and **incoherent**” (q') appears in other blogs twice on the 20th and once on the 21st of January. If q was sufficiently prominent when q' first appeared, we can safely assume that the first author of q' on the 20th based himself on q as is shown in Figure 3a. But what about the second and third occurrences of q' , on the 20th and 21st? Should we consider them to be substitutions based on q or accurate reproductions of the previous occurrences of q' ? (Options shown in Figure 3a.)

To settle this question we group quote occurrences into fixed bins spanning Δt days (1 day in the implementation), each one representing a unit of time evolution. When a quotation q' appears in bin $t + 1$, it is counted as a substitution if it differs from the most frequent quote q of the preceding bin t (or a substring thereof) by only one word. If not, q' is not considered to be an instance of substitution. These assumptions are only a subset of a wider set of possibilities, each leading to alternative substitution inferences.⁷ These various flavors of an ideal substitution detection model essentially

⁷In particular, the criterion of the most frequent quote in the preceding bin may be replaced with the most frequent quote overall, or the oldest quote; time can be sliced into fixed bins as is done here, or kept fine-grained by using sliding bins.



(a) Possible paths from occurrence to occurrence



(b) Binned quotation family with majority rule

Fig. 3: **Temporal binning of quotation families.** q and q' are two versions of a quotation belonging to the same cluster. In the bottom panel (b), q' holds the majority in the 3rd bin and is considered the unique basis for the last occurrence of q (in the 4th bin). This is despite the fact that q also appears in bin 3 alongside q' , and despite it having appeared earlier at the very beginning of the quotation family (indeed in the situation shown in Figure 3b, this seems to be the most likely scenario). Conversely, if q had been the most frequent quote in bin 3, the last occurrence of q in bin 4 would have been considered a faithful copy of the occurrence of q in bin 3.

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.

Put shortly, such a model defines how many times quote occurrences can be counted as substitutions: in Figure 3b, occurrences of q' on the 20th are counted as substitutions, whereas the occurrences on the 21st are not. In practice, from the 838k initial occurrences spread into 19,621 classes of quotes, with significant redundancy (many quotes are indeed simple duplicates), we manage to mine 2,873 real substitutions obeying to this model. From these substitutions we remove those featuring stop words, minor spelling changes (e.g. center/centre, November/Nov, Senator/Sen), abbreviations, spelled out numbers; this eventually yields 476 valid substitutions.

4. RESULTS

We may now use this substitution model to formulate a family of psycholinguistic hypotheses describing the role of each feature in the accuracy of the reformulation. To this end, we build two main observables for each word feature. First, we measure the susceptibility for words to be the target of a substitution in a quote, 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 change in word feature upon 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 former case we know there has been a human reformulation, whereas in the latter 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”). Furthermore, perfect human reformulation possibly involves different practices than those involved in alteration — e.g. drafting before publishing, double-checking sources, proof-reading — and may not be representative of the cognitive processes at work during alteration. The two situations are different enough to be studied separately, and we focus here on the latter.

4.1 Susceptibility

We say that a word is *substitutable* if it appears in a quote which undergoes a substitution, whether that substitution operates on that word or on another one. Word substitution susceptibility is computed as the ratio of the number of times s_w a word is substituted to the number of times p_w that word appears in a substitutable position, i.e. s_w/p_w . In other words, it measures how often a word w actually gets substituted, compared to how often it could have been substituted (because it appears in quotes undergoing substitution).

Now, for a given feature ϕ , we obtain the mean susceptibility $\sigma_\phi(f)$ for the feature value f by averaging this ratio over all words such that $\phi(w) = f$, that is:

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

Put shortly, susceptibility 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 will be substituted.

Results for this measure are gathered in Fig. 4. They first show an obvious strong effect of Word frequency: the more frequent a word, the less likely it is to attract substitutions. Indeed, susceptibility goes from nearly 0.4 for low-frequency words down to 0 for very high-frequency words. To make things clear, this value of 0.4 means that low-frequency words, when present in a quote undergoing a substitution, are the ones being substituted 40% of the time on average.

The other features — Age of acquisition, Number of phonemes, Clustering coefficient and Number of synonyms — do not seem to exhibit any particularly significant effect on susceptibility. For each of these features, it is indeed possible to draw a constant line which always remains within the respective confidence intervals. If these variables have an effect, it is by no means as strong as it is for Word frequency. This is remarkably clear for Clustering coefficient and Number of synonyms, where susceptibility values remain within quite small intervals (respectively [0.09-0.13] and [0.11-0.15]). We may notice a somewhat marked effect for the highest values of Age of acquisition and Number of phonemes, where the mean susceptibility is almost twice as high as in the rest (respectively 0.19 vs 0.13, and 0.21 vs 0.12). Keeping in mind the poor statistical significance of this effect, we could still suggest that the longest words and words learned late are more susceptible to substitution. To further examine this phenomenon, we plotted the two-dimensional map of susceptibility values for these two features (see heatmap at the bottom right of Fig. 4). We see that outlier values (the two darkest cells) are both due to cells where the number of items is only slightly above the significance threshold (11 items, while we decided to cut at 10). If we focus on cells based on at least several dozens of words, values range from 0.11 to 0.22 with little obvious regularity, making it hard to draw any conclusion as regards the

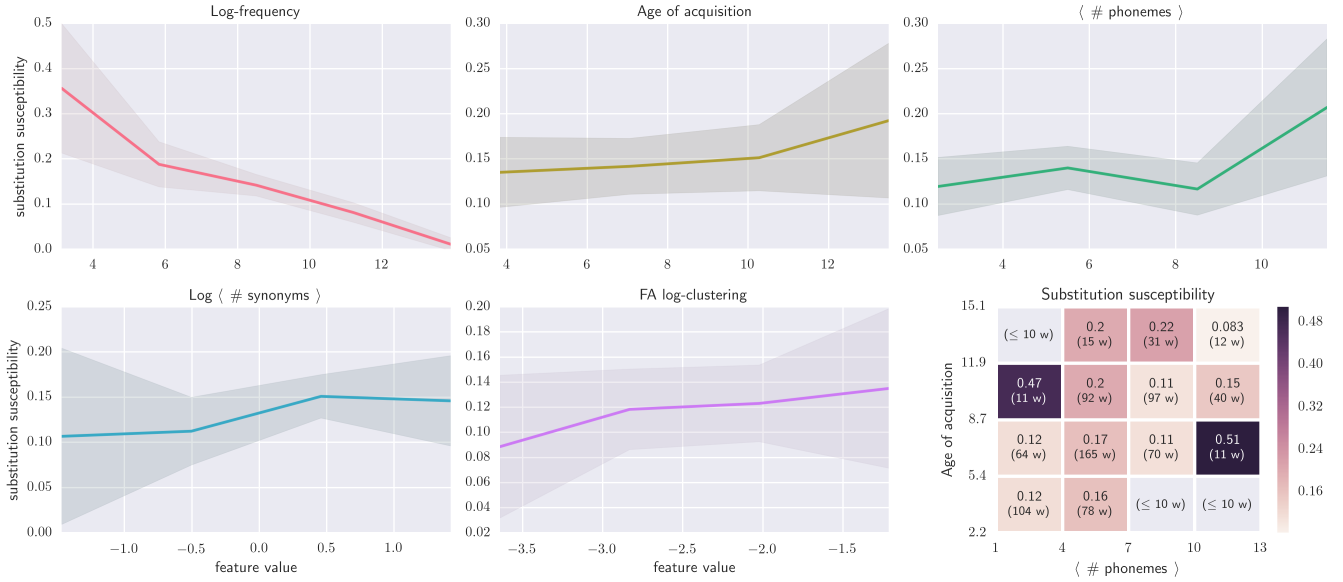


Fig. 4: **Substitution susceptibility:** average susceptibility to substitution vs. average feature value of a candidate word for substitution, with 95% asymptotic confidence intervals. The heatmap on the lower-right shows the joint effect of Age of acquisition and Number of phonemes on susceptibility, averaged over the respective single-variable ranges, with sample size (word numbers) in parentheses. Without the rather non-significant cells (11 words or less), the heatmap shows no specific trend in one direction or the other.

direction of an effect. This confirms the fact that the slight late increases observed for Age of acquisition and Number of phonemes are rather non-significant.

All in all, apart from Word frequency and despite some tendencies these results do not allow us to conclude either to a marked effect of the selected psycholinguistic features on substitution susceptibility. We may therefore globally assume that substitution targets are chosen in a more or less uniform way with respect to these features.

4.2 Variation

We can thus show how words are modified once we know they are substituted, i.e. how their features are modified by said substitution. We therefore measure how a **word w 's feature** varies as w is substituted for w' , i.e. $\phi(w')$ as a function of $\phi(w)$. Averaging this value over all start words such that $\phi(w) = f$ yields the mean variation for that feature value f , i.e.:⁸

$$\nu_\phi(f) = \langle \phi(w') \rangle_{\{w \rightarrow w' | \phi(w) = f\}}$$

Of prime interest is the comparison of the value of $\nu_\phi(f)$ with respect to f , as it shows whether there is an attraction (or a repulsion) effect towards (respectively from) some values of each feature. In other words, plotting the $y = x$ line, we can see if substitutions tend to converge towards some typical value of a word feature or not — as is classically done in the study of dynamical systems.

We also 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' was randomly chosen from the whole pool of words available in the dataset for that feature.⁹ In this case, since $\phi(w')$ becomes a constant value in the above averaging (by definition w' does not depend on w anymore), the baseline variation under \mathcal{H}_0 may be rewritten as:¹⁰

$$\nu_\phi^0(f) = \langle \phi \rangle$$

This approach yields a fine-grained view of how word features evolve upon substitution, on average, with respect to (i) the original feature (vs. $y = x$) and (ii) a random arrival (vs. ν_ϕ^0).

Results are gathered in Fig. 5. We can do a first striking observation: all graphs show the existence of a unique intersection of ν_ϕ with $y = x$, while the slope of ν_ϕ is smaller than 1, independently of the feature considered. In other words, beyond individual variation patterns, the substitution process is contractile for all the features, and each of them therefore exhibits a unique attractor. Second, the comparison with ν_ϕ^0 shows that there are two classes of attractors, depending on whether:

- (i) there is a triple intersection (of $y = x$, ν_ϕ^0 and ν_ϕ);
- (ii) or ν_ϕ always remains above or below ν_ϕ^0 .

The first class (Number of phonemes and Number of synonyms) are features for which the substitution process only brings words slightly closer to ν_ϕ^0 , and no uniform bias can be observed.

⁹For instance, when considering the feature “Clustering coefficient”, the arrival word is randomly chosen among words present in the dataset of FA norms.

¹⁰We additionally considered an alternative null hypothesis, denoted \mathcal{H}_{00} , where the arrival word is randomly chosen *among immediate synonyms of the start word*, i.e. an arrival word chosen among semantically plausible though still random words. In this case w'_{00} does depend on w . Our conclusions hold under this second null hypothesis, so for the sake of clarity we chose to keep the simpler \mathcal{H}_0 .

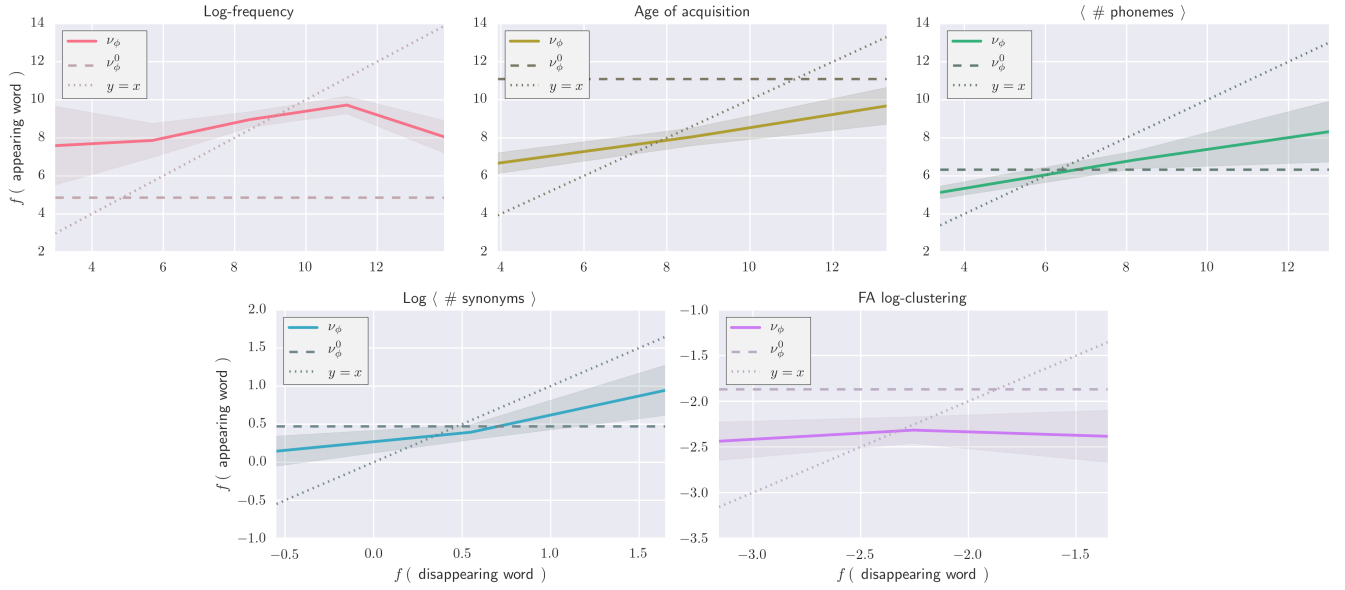


Fig. 5: **Feature variation upon substitution:** ν_ϕ , average feature value of the appearing word as a function of the feature value of the disappearing word in a substitution, with 95% asymptotic confidence intervals. The overall position of the curve with respect to the dashed line representing \mathcal{H}_0 (constant ν_ϕ^0) indicates the direction of the cognitive bias. The intersection with $y = x$ marks the attractor value.

On the other hand, the second class (comprising Word frequency, Age of acquisition, and Clustering coefficient) are features for which the substitution process has a clear bias, positive or negative, with respect to the purely random situation (\mathcal{H}_0).

Word frequency, with ν_ϕ always significantly above ν_ϕ^0 , exhibits a strong bias towards more frequent words. This, in turn, is consistent with the hypothesis that substitution is a recall process, since common words are favored over awkward ones, while it goes against the idea that it could be a familiarity process, where awkward terms would be favored.

Age of acquisition and Clustering coefficient, on the other hand, exhibit a clear negative bias for the substitution process. Both curves are significantly below their respective ν_ϕ^0 values, which is consistent with the literature on recall: words learned earlier and words with lower clustering coefficient are easier to produce than average [Nelson et al. 2013; Zevin and Seidenberg 2002]. Clustering coefficient has the additional particularity that, on average, the destination word does not depend on the start word; i.e. on average, substitutions will always produce words with a clustering coefficient around $\exp(-2.4) \simeq 0.1$.

To make things concrete, here is an example substitution taking place in the dataset. At the end of January 2009, many media websites reported the following quote,

“The massive economic upheaval being experienced across the globe is sparing no one in the consumer electronics world.”

and a smaller number of media websites, and blogs, reported the following,

“The massive economic upheaval being experienced across the **world** is sparing no one in the consumer electronics world.”

The word *globe* is acquired at an average of 6.5 years old, appears about 3.5k times in the dataset, and has a Clustering coefficient of

0.24. The word it was replaced with, *world*, is acquired on average at 5.3 years old, appears about 146k times in the dataset, and has a Clustering coefficient of 0.05. (Both words have four phonemes.) Such a change, though minor in appearance, is a typical example of alteration along the lines shown by our results.

We thus observe a clear convergence pattern for each feature, with two different classes corresponding to the psychological relevance of each feature for the substitution process. Taken as a dynamical system where substitutions are repeatedly applied, Number of phonemes and Number of synonyms will simply converge towards their average value in the FA corpus (i.e. ν_ϕ^0), while Word frequency, Age of acquisition and Clustering coefficient, consistent with the literature, will converge towards significantly biased values indicated by the intersection with $y = x$ (respectively, a frequency of $\exp(9.1) \simeq 9000$, an acquisition age of 8, and a Clustering coefficient of 0.1).

5. CONCLUDING REMARKS

We aimed to contribute to the empirical understanding of **knowledge transformation processes** by studying a simple task where individuals are *implicitly* trying to reproduce textual content. To some extent, our work amounts to a large *in vivo* experiment where we appraise the impact of classically-influent psycholinguistic variables in the accuracy of the reproduction. In more detail, we describe the joint properties of the substituted and substituting terms in the reformulation by individuals of a specific type of utterances (quotations).

For each of the selected psycholinguistic variables, we demonstrate the existence of attractor values in the underlying variable spaces. More precisely, beyond the interpretation of our results for each variable, we notice that all variables remarkably exhibit a single attractor and are generally contractile — as such, even though the observed convergence patterns only partially explain quotation evolution, we shed light on a class of phenomena which are prone

to constitute a key element of a broader empirically-grounded, attractor-based theory of cultural evolution.

Acknowledgements

We are warmly grateful to Ana Sofia Morais for her precious feedback and advice on this research, and to Jean-Pierre Nadal, Sharon Peperkamp, and Nicolas Baumard for useful suggestions and comments.

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.
- Joseph L Austerweil, Joshua T Abbott, and Thomas L Griffiths. Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2012.
- 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 Viteitch. 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}.
- Nicolas Claidière, Thomas C Scott-Phillips, and Dan Sperber. How darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1642):20130368, 2014.
- 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".
- Stephen A Dewhurst, Graham J Hitch, and Christopher Barry. Separate effects of word frequency and age of acquisition in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(2):284, 1998.
- 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.
- Vernon Gregg. Word frequency, recognition and recall. 1976.
- 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. *KDD'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.
- Robert M MacCallum, Matthias Mauch, Austin Burt, and Armand M Leroi. Evolution of music by public choice. *Proceedings of the National Academy of Sciences*, 109(30):12081–12086, 2012.
- Catriona M Morrison and Andrew W Ellis. Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):116, 1995.
- 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.
- Douglas L Nelson, Kirsty Kitto, David Galea, Cathy L McEvoy, and Peter D Bruza. How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & cognition*, 41(6):797–819, 2013.
- 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.
- 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.