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

Sébastien Lerique, Camille Roth

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

Sébastien Lerique
Centre d'Analyse et de Mathématique Sociales, UMR 8557 CNRS/EHESS
190 av. de France, F-75013 Paris
and Centre Marc Bloch Berlin, UMIFRE 14 CNRS/MAEE/HU
Friedrichstr. 191, D-10117 Berlin

Camille Roth
CNRS
Centre Marc Bloch Berlin, UMIFRE 14 CNRS/MAEE/HU
Friedrichstr. 191, D-10117 Berlin

Author Note

Correspondence should be directed to lerique@cmb.hu-berlin.de and roth@cmb.hu-berlin.de

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.

Keywords: word production; recollection bias; semantic network; cultural evolution; cultural attraction; data mining; big data; *in vivo* psycholinguistics

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

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 & 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 kinds 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 & 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, Guha, Liben-Nowell, & Tomkins, 2004) or the email network (Liben-Nowell & Kleinberg, 2008), the characteristic times and diffusion cycles both within these social networks and with respect to the topical dynamics of news media (Leskovec, Backstrom, & Kleinberg, 2009b), or the reciprocal influence between the social network topology and the distribution of issues (Cointet & 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

[1] Works 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).

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 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 describes our hypotheses along with the relevant state-of-the-art on this psycholinguistic matter. Then, we detail the empirical protocol and the various assumptions that were made in order to deal with the available empirical material. We further describe 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 the final section.

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 & Sperber, 2007, and Claidière, Scott-Phillips, & Sperber, 2014, as well as a hybrid empirical-theoretical protocol in MacCallum, Mauch, Burt, & Leroi, 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. (2009b) and further analyzed by Simmons, Adamic, and Adar (2011) and Omodei, Poibeau, and Cointet (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 & Seidenberg, 2002), number of phonemes or number of syllables (see for instance Nickels & Howard, 2004; Rey, Jacobs, Schmidt-Weigand, & Ziegler, 1998), have on 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 & 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, Steyvers, and Firl (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 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. Austerweil, Abbott, and Griffiths (2012) further develop this idea by showing that random walk on such a semantic network, that is 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). D. L. Nelson, Kitto, Galea, McEvoy, and Bruza (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 assocation).

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, that is 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.

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 study the individual transformation process at work when authors alter quotations, by examining the modified words in each transformation. 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.

In vivo utterances

We used a quotation dataset collected by Leskovec et al. (2009b), 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, Backstrom, & Kleinberg, 2009a).[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, that is sufficiently similar, groups of quotations (see Leskovec et al., 2009b, for more details). 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 casual expressions, while we are interested in transformations of news-related quotes causally linked to an original, identifiable utterance. To filter them out, we exclude all quotes having less than 5 words long or lasting more than 80 days (as well as quotes not written in English). If an entire cluster still lasts more than 80 days

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

after this screening (because of short-lived but unrelated quotes far apart in time), we also exclude it. We eventually keep 45,749 clusters (out of 71,568; i.e. 63.9%), containing a total of 127,778 unique quotes (out of 310,457; i.e. 41.2%) making up about 2.43m occurrences (out of 8.16m, i.e. 29.8%).[3] Even if we lose some real event-related utterances which are present in clusters lasting more than 80 days (such as "the city is tired of me and the organization and I have run our course together"), we check that our approach essentially fulfills its goals by manually coding a random subsample of 100 excluded clusters: a solid 71% appear to be entirely irrelevant to our analysis (everyday language rather than quotations), and all but one of the remaining clusters were of relevance to the protocol set out below.

Word-level measures

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, Stadthagen-Gonzalez, & Brysbaert, 2012), known to have different effects than word frequency (Dewhurst, Hitch, & Barry, 1998; Morrison & Ellis, 1995),
- 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.

[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.

[4] The CMU Pronouncing Dictionary is included in the NTLK package (Bird, Klein, & Loper, 2009), the natural language processing toolkit we used for the analysis.

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 D. Nelson, McEvoy, and Schreiber (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 D. 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, Brin, Motwani, & Winograd, 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 & 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).

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 (.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.

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 Fig. 1: age of acquisition has a limited correlation to the other variables (absolute value not above .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 (.85), and are also redundant with betweenness centrality (with correlation levels at .75 and .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,

[5]The Clustering coefficient is formally defined as the ratio between the number of actual versus possible edges between a node's neighbors.

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 .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 Fig. 2.[6]

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, that is 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

[6]Note 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.

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 Fig. 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 Fig. 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. Note that these assumptions are admittedly a subset of a much wider set of possibilities, each leading to alternative substitution inferences.[7] It is however not feasible to try them all and, for the sake of simplicity, we decided to go with a sensible set of assumptions, and stick to them without trying alternative options.

Put shortly, such a model defines how many times quote occurrences can be counted as substitutions: in Fig. 3b, occurrences of q' on the 20th are counted as substitutions, whereas the occurrences on the 21st are not. In practice, from the 2.43m initial occurrences spread into 45,749 classes of quotes, with significant redundancy (many quotes are indeed simple duplicates), we manage to mine 6,172 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 1,051 valid substitutions.

[7] In 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.

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 — for instance 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.

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, that is 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 .33 for low-frequency words down to nearly 0 for very high-frequency words. To make things clear, this value of .33 means that low-frequency words, when present in a quote undergoing a substitution, are the ones being substituted 33% 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. If we set aside the values for low Number of phonemes, 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 Age of acquisition, where susceptibility values remain within quite small intervals (respectively $[.13 - .18]$ and $[.16 - .20]$). We may notice a slight effect for the lowest values of Number of synonyms and Number of phonemes, where the mean susceptibility is almost half as high as the average of the other values (respectively .09 vs. .16, and .11 vs. .17). Keeping in mind the poor statistical significance of this effect, we could still wonder if the shortest words and words with fewest synonyms are significantly less 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). Even if there are

a few outlier cells, values tend to navigate around the mean value (.16) with little obvious regularity (except for a low number of synonyms, consistent with the unidimensional graph). On the whole, this makes it relatively hard to draw any conclusion as regards the direction of an effect, except for the least populated value ranges (which as a result are also less significant).

All in all, apart from Word frequency and despite some local tendencies, in general these results do not allow us to conclude 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.

Variation

We can thus show how words are modified once we know they are substituted, that is how their features are modified by said substitution. Considering a word w substituted for w', we measure how the feature of w varies when it is replaced with w', that is we look at $\phi(w')$ 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, that is:[8]

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

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

[8] To avoid possible autocorrelation effects due to substitutions belonging to the same cluster (which are likely not statistically independent and may lead to overly optimistic confidence intervals), we first average substitutions over each cluster, by considering the average of arrival word features for a given start word.

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]

$$\nu^0_\phi(f) = \langle \phi \rangle$$

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

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^0_ϕ shows that there are two classes of attractors, depending on whether:

1. there is a triple intersection (of $y = x$, ν^0_ϕ and ν_ϕ);
2. or ν_ϕ always remains above or below ν^0_ϕ.

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

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 ($\mathcal{H}_0$).

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

[10] We additionally considered an alternative null hypothesis, denoted $\mathcal{H}_{00}$, where the arrival word is randomly chosen *among immediate synonyms of the start word*, that is 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 $\mathcal{H}_0$.

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 (D. L. Nelson et al., 2013; Zevin & Seidenberg, 2002). Clustering coefficient has the additional particularity that, on average, the destination word does not depend on the start word; that is on average, substitutions will always produce words with a clustering coefficient around $\exp(-2.4) \simeq .1$.

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 .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 .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^0_ϕ), 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 slightly below 8, and a Clustering coefficient of .1).

Concluding remarks

We aimed to contribute to the empirical understanding of representation 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 susceptible 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 Telmo Menezes, Jean-Philippe Cointet, Jean-Pierre Nadal, Sharon Peperkamp, and Nicolas Baumard for useful suggestions and comments.

References

Atran, S. (2003). Théorie cognitive de la culture. *L'Homme*, *166*(2), 107-143.

Aunger, R. (Ed.). (2000). *Darwinizing Culture: The Status of Memetics as a Science*. Oxford: Oxford University Press.

Austerweil, J. L., Abbott, J. T., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems* (pp. 3041–3049).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly Media, Incorporated.

Bloch, M. (2000). A well-disposed social anthropologist's problems with memes. In R. Aunger (Ed.), *Darwinizing Culture: The Status of Memetics as a Science* (pp. 189–203). Oxford University Press.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. University of Chicago Press.

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cogn Sci*, *34*(4), 685-97.

Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1642), 20130368.

Claidière, N., & Sperber, D. (2007). The role of cultural attraction in cultural evolution. *Journal of Cognition and Culture*, *7*(1), 89–111.

Cointet, J.-P., & Roth, C. (2009). Socio-semantic dynamics in a blog network. In *Proc. IEEE 4th Intl. Conf. Social Computing* (pp. 114–121).

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Dawkins, R. (1976). Memes: The New Replicator. In *The Selfish Gene* (pp. 189–201). Oxford University Press.

Dewhurst, S. A., Hitch, G. J., & Barry, C. (1998). Separate effects of word frequency and age of acquisition in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 284.

Ehrlich, P. R., & Levin, S. A. (2005). The evolution of norms. *PLoS Biol.*, *3*(6), e194.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*, 35–41.

Gregg, V. (1976). Word frequency, recognition and recall.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: predicting fluency with PageRank. *Psychol. Sci.*, *18*(12), 1069-76.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information Diffusion Through Blogspace. In *Proceedings of the 13th International World Wide Web Conference (WWW'04)* (pp. 491–501).

Kuper, A. (2000). If memes are the answer, what is the question? In R. Aunger (Ed.), *Darwinizing Culture: The Status of Memetics as a Science* (pp. 180–193). Oxford University Press.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009a). *MemeTracker: tracking news phrase over the web.* `http://memetracker.org/`. (Retrieved on August 19, 2012)

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009b). Meme-tracking and the dynamics of the news cycle. In *Proc. ACM SIGKDD'09 15th Intl. Conf. on Knowledge Discovery and Data Mining* (pp. 497–506).

Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, *105*(12), 4633-4638.

MacCallum, R. M., Mauch, M., Burt, A., & Leroi, A. M. (2012). Evolution of music by public choice. *Proceedings of the National Academy of Sciences*, *109*(30),

12081–12086.

Morrison, C. M., & Ellis, A. W. (1995). 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.

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

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & cognition*, *41*(6), 797–819.

Nickels, L., & Howard, D. (2004). 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.

Omodei, E., Poibeau, T., & Cointet, J.-P. (2012). Multi-level modeling of quotation families morphogenesis. In *Proc. ASE/IEEE 4th Intl. Conf. on Social Computing "SocialCom 2012".*

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November). *The PageRank Citation Ranking: Bringing Order to the Web.* (Tech. Rep. No. 1999-66).

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

Simmons, M. P., Adamic, L. A., & Adar, E. (2011). Memes online: Extracted, substracted, injected and recollected. In *Proc. 5th ICWSM - AAAI Intl Conf Weblogs & Social Media* (pp. 353–360).

Sperber, D. (1996). *Explaining culture: A naturalistic approach.* Oxford: Blackwell Publishers.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.

Weide, R. (1998). *The CMU Pronouncing Dictionary, release 0.6.* Carnegie Mellon University.

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

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

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

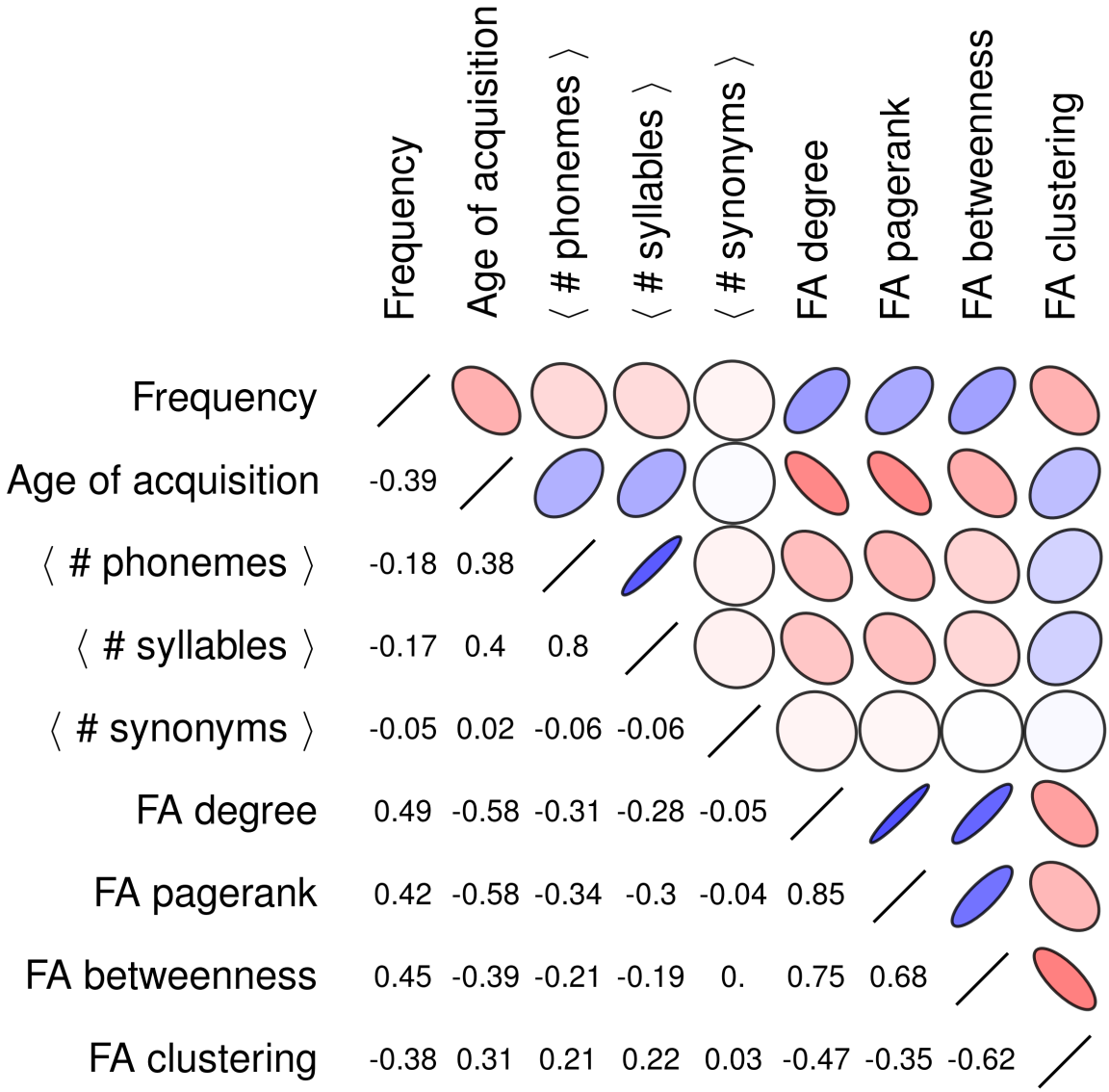


Figure 1. Spearman correlations in the initial set of features

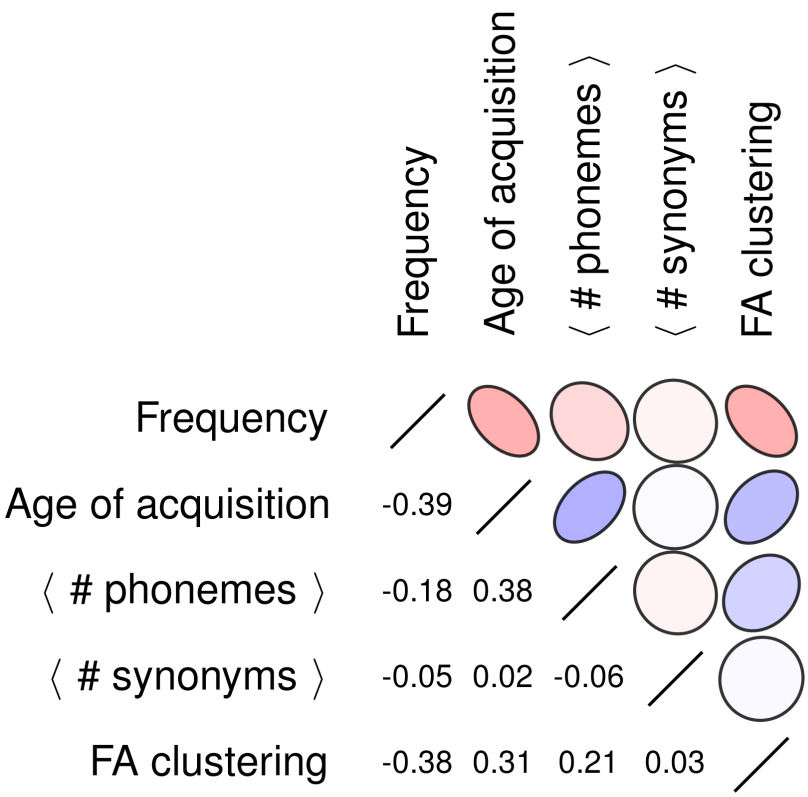
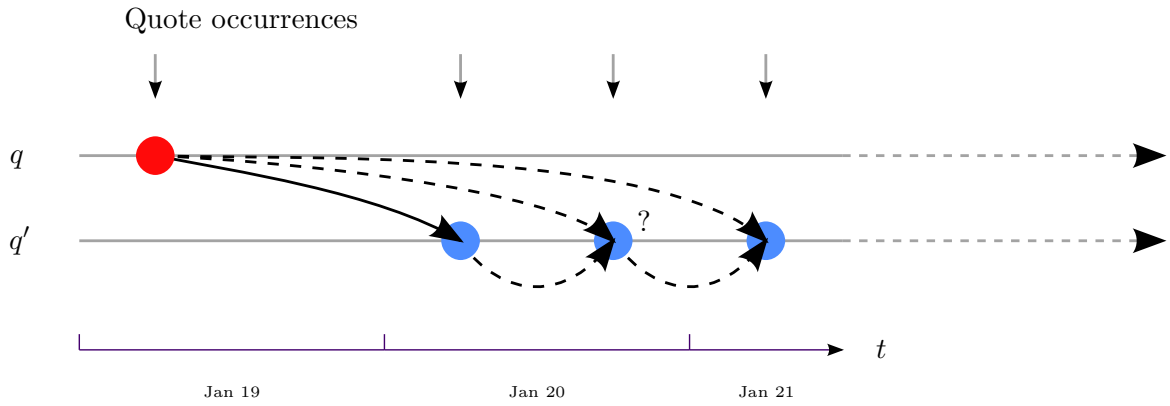
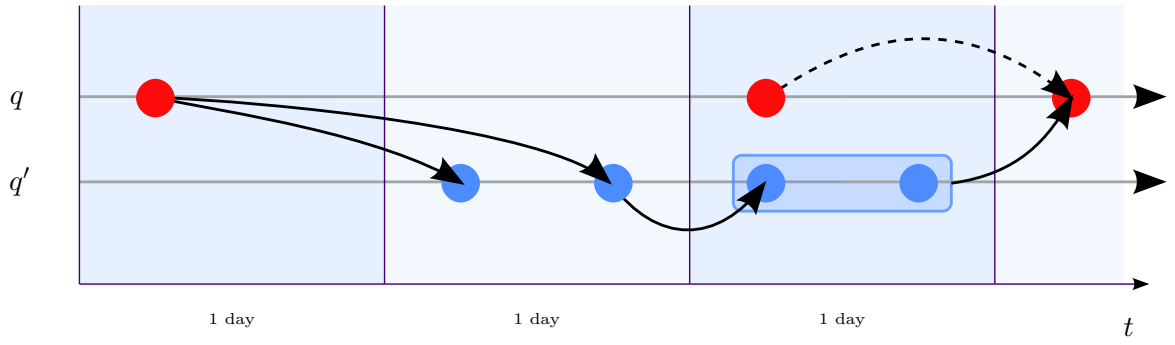


Figure 2. Spearman correlations in the filtered set of features



(a) Possible paths from occurrence to occurrence



(b) Binned quotation family with majority rule

Figure 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 Fig. 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.

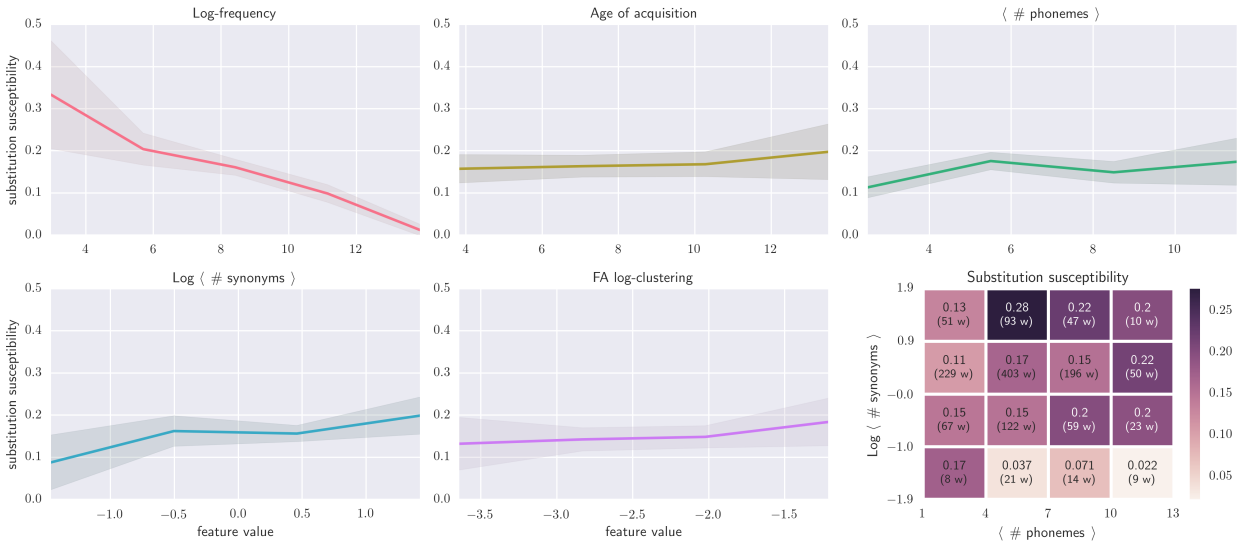


Figure 4. **Substitution susceptibility:** average susceptibility to substitution versus average feature value of a candidate word for substitution, with 95% asymptotic confidence intervals. The heatmap on the lower-right shows the joint effect of Number of synonyms and Number of phonemes on susceptibility, averaged over the respective single-variable ranges, with sample size (word numbers) in parentheses.

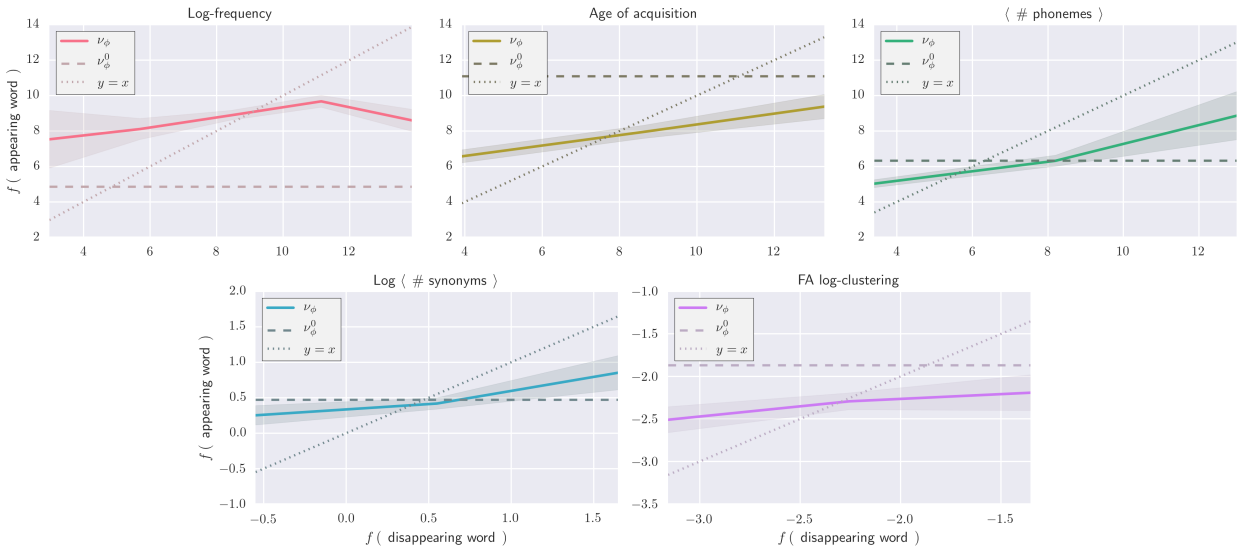


Figure 5. **Feature variation upon substitution:** ν_ϕ, average feature value of the appearing word as a function of the feature value of the disappearing word in a substitution, with 95% asymptotic confidence intervals. The overall position of the curve with respect to the dashed line representing $\mathcal{H}_0$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