The semantic drift of quotations in blogspace: a case study in short-term cultural evolution

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

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. Update after introduction rework.

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

Introduction

#20: check text/flow/definitions for clarity against Gureckis' edited pdf

Since the very beginnings of both social science and psychology, scientists have tried to capture the way cognition and culture influence each other. While it has been the subject of intense debate in the social sciences in the 20th century (starting with Durkheim's initial works, 1912, later tackled in earnest by e.g. Mauss' *Techniques of the Body*, 1936, Giddens' *Structuration Theory*, 1984, and Bourdieu's *Sens Pratique*, 1980), today's discussion is mostly structured by proponents from cognitive science.

These construe culture as an evolutionary process analogous and parallel to biological evolution (and especially the modern synthesis' account of it). That analogy can be traced a long way back in the 20th century and earlier, with milestones such as Kroeber's works (1952), Dawkins' *Memetics* (2006), and later the development of *Dual Inheritance Theory* by Boyd and Richerson (1985) and Cavalli-Sforza and Feldman (1981) among others. More recently, Dan Sperber has drawn on this principle to explicitly connect anthropology and cognitive science through the theory of *Epidemiology of Representations* (Sperber, 1996), and the study of cultural evolution has been growing steadily since.

The collection of works by Aunger (2000) (in particular Bloch, 2000, and Kuper, 2000) has shown how the theory of memetics cannot account for the levels of transformation culture undergoes as it is transmitted. Mesoudi and Whiten (2008) have discussed the uses of transmission chain experiments to test what dual inheritance theory can explain about cultural evolution. Morin (2013) and Miton, Claidière, and Mercier (2015), by carefully compiling a series of anthropological works, show how cognitive biases have influenced the evolution of cultural artifacts over several centuries. Kirby, Cornish, and Smith (2013; 2008) have shown how evolutionary pressures lead to the emergence of structured and expressive artificial languages in simulations and laboratory experiments. Such transmission chain experiments have also been explored in non-human primates by Claidière, Smith, Kirby, and Fagot (2014).

The theory of epidemiology of representations proposes a unifying framework for all these works by recasting them as questions of spread and transformation of representations: these are alternatively located in the mind ("mental representations" in Sperber's terminology), or in the outer world

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

("public representations") as expressions of mental representations in diverse cultural artifacts (pieces of text, utterances, pictures, building techniques, etc.). A human society is then modeled as a large dynamical system of people constantly interpreting public representations into mental representations, and producing new public representations based on what they have previously interpreted. Two key points are that (a) transmission is not reliable (representations change significantly each time they are interpreted and produced anew, as opposed to e.g. memetics), and (b) the reciprocal influences of cognition and culture can be captured by studying the evolution of public representations themselves, which is what the studies cited above are doing.

The theory makes an additional strong hypothesis, which this paper focuses on: as transformations accumulate, some representations evolve to be very stable and spread throughout an entire society without changing any more (they are called "cultural representations", because they characterize a given culture). This process should manifest itself as attractors (called "cultural attractors") in the dynamical system that models cultural evolution, that is: there should be areas of the representation space where any cognitive effect in transformation brings representations closer to a given stable asymptotic point.[1]

This hypothesis, a cornerstone of the theory because of the intelligibility it gives to cultural evolution, has been hard to test in concrete situations: quantitative data on out-of-laboratory cultural artifacts is not easy to collect. One approach, as mentioned above, has been the meta-analysis of large bodies of anthropological studies (see Miton et al., 2015, for instance). This paper exemplifies a second approach, taking advantage of the ever-increasing avalanche of available digital footprints since the 2000's. Indeed, tools and computing power to analyze such data are now widespread, and the body of research aimed at describing online communities and content is growing accordingly. For instance, the propagation of cultural artifacts across social networks has been studied in blogspace (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004) and in the email network (Liben-Nowell & Kleinberg, 2008); Cointet and Roth (2009) have described the reciprocal influence between the social network topology and the distribution of issues; Leskovec, Backstrom, and Kleinberg (2009) detailed the characteristic times and diffusion cycles both within these social networks and with respect to the topical dynamics of news media, and Danescu-Niculescu-Mizil, Cheng, Kleinberg, and Lee (2012) have studied the characteristics of particularly memorable quotes that circulate in those networks. We believe those works can connect the field of cultural evolution with psycholinguistics to advance the testing of cultural attractors.

To show this we analyze the way quotes in blogs and media outlets are modified when they are copied from website to website. These public representations should normally not change as they spread on the Web (as opposed to more elaborate expressions or opinions, not identified as quoted utterances), but empirical observation shows that they are in fact occasionally transformed (Simmons, Adamic, & Adar, 2011): authors spontaneously transform quotes, not only cropping them but also replacing words, when in fact they are implicitly required to copy them exactly. We can therefore assume that most transformations, especially the simple ones, are the result of automatic (i.e. hard to control) low-level cognitive biases of the authors.

Our question is as follows: given such representations that seem to evolve precisely because of the kind of automatic cognitive biases referred to in the theory of epidemiology of representations, do cultural attractors appear and how do cognitive biases participate in them? We chose to restrict our analysis to substitutions (one word being replaced by another), both to keep the analysis tractable and because of missing information in our data set.[2] While this restricts the scope of our observations for the particular data set we use, the methodological point we also make is left intact. By characterizing words with 6 well-known features, we identify what makes a substitution more likely, and how a word changes when it is substituted. This exploratory approach uncovers a number of transmission biases consistent with known effects in linguistics. While the transformations we describe are not the only ones at work in this data set, our analysis also indicates that feature-specific attractors could exist because of the substitution process. This study can be viewed as analyzing part of the transmission step operating in transmission chains of artificial languages like those studied by Kirby et al. (2008), but with natural language out of the laboratory.

The next section describes our hypotheses along with a review of the psycholinguistics literature. Then, we describe the data set and detail the various assumptions that were made in order to analyze it. Next, we describe the measures built to observe the cognitive biases operating in quote transmission. Finally, we discuss the relevance of these results for the study of cultural evolution, followed with general guidelines for further work.

Related work

#15: add (1) sentence recall (Potter & Lombardi), (2) false memories (Deese), (3) subjective organization (Tulv-

[1] Attractors need not be points in fact, they can also be sub-areas; in that case any transformation brings representations in the area closer to (or maintained inside) the target sub-area.

[2] As explained further down, source-destination links between quotes must be inferred from the data set, an operation which is much more reliable if we restrict our analysis to substitutions. This also impedes us from observing the effect of accumulated transformations in the long term, limiting our results to a view of the individual evolutionary step.

ing, Zaromb), (4) working memory and attention (Jefferies), (5) iterated learning (Kirby)

#20: check text/flow/definitions for clarity against Gureckis' edited pdf

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 **?**, and **?**, as well as a hybrid empirical-theoretical protocol in **?**). 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 **?** and further analyzed by **?** and **?**. 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 **?**, for a review), age-of-acquisition (**?**), number of phonemes or number of syllables (see for instance **??**), 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 **?**). 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, **?** 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. **?** 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 **?** 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). **?**, 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 **?**, 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

#20: check text/flow/definitions for clarity against Gureckis' edited pdf

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.

#19: better explain choice of single-substitution case, referring to further down in (1) substitution detection and (2) susceptibility measure

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

#17: go into more detail on how quotes are selected by Leskovec et al. (or explain there's no info if there's none)

We used a quotation dataset collected by **?**, 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 — **?**).[3] 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."

#20: make clear we didn't do this, it's Leskovec et al.

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 **?**, 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.

#17: precision/recall analysis on anti-spam and language

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 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%).[4] 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. #16: (1) update the set of features, (2) better explain feature selection: no PCA or all-feature regression (a predictive model doesn't differentiate correlated features), it's robust feature selection with domain knowledge (3) note clustering values have changed because of weights

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 (**?**),
- **Age of Acquisition**: the average age at which words are learned (obtained from **?**), known to have different effects than word frequency (**??**),
- The average **Number of Phonemes** and **Number of Syllables** for all pronunciations of a word (obtained from the Carnegie Mellon University Pronouncing Dictionary, **?**)[5] as a proxy to word production cost,
- The average **Number of Synonyms** for all meanings of a word (obtained from **?**) 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.

#20: FA don't encode reformulation explicitly, they did norms. We hypothesize it's what's at work.

We relied on the "free association" (FA) norms collected by **?** which naturally embed information on the idea association process underlying transformation of quotations. FA

[3] Unfortunately, the original article (**?**) does not provide additional details on the source selection methodology.

[4] 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.

[5] The CMU Pronouncing Dictionary is included in the NTLK package (**?**), the natural language processing toolkit we used for the analysis.

norms record the words that come to mind when someone is presented with a given cue. As **?** 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 **?**, 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* (**?**), which has already been used on the FA network by **?**. 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 (**?**). 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 (**?**), computed on the undirected version of the FA network.[6] 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 **?**, 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. **??**: 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, 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. **??**.[7]

Substitution model

#19: show more clearly why this is more robust with single-substitution, justifying our choice of this case

#17: explain all models and the work involved, with graphs and explanatory figures for each model in annex

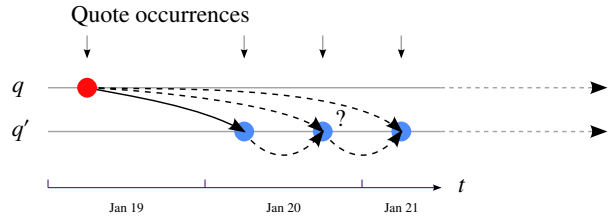
We finally need a substitution detection model, for the utterance data we use presents a challenge: quote-to-quote transformations, and much less substitutions, are not explicitly encoded in the dataset. More precisely, each set of quotations bears no explicit information about either the authoritative 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.

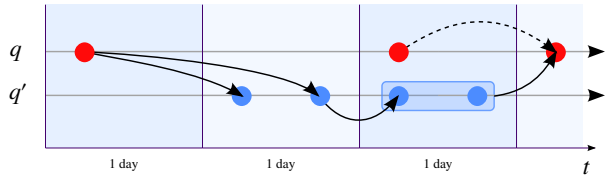
[6] The Clustering coefficient is formally defined as the ratio between the number of actual versus possible edges between a node's neighbors.

[7] 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.

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 Fig. 1a. 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. 1a.)



(a) Possible paths from occurrence to occurrence



(b) Binned quotation family with majority rule

Figure 1. **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. 1b, 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.

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.[8] 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.

#12: fix substitution counts

Put shortly, such a model defines how many times quote occurrences can be counted as substitutions: in Fig. 1b, 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.

Results

#16: add a quick description of what types of substitutions we see, then moving on to the observables since it's not what principally interests us.

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

#19: explain if we're susceptible to bias from single-substitution

[8] 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.

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.

#16: (1) show POS don't differ, (2) note that stopwords are excluded, which excludes POS open/closed discussion

#16: explain we do this on quartiles, rework results discussion

#16: explicit all significance claims, referring to the CIs

#17: show other feature variations and other models in annex

#15: relate to missed literature

Results for this measure are gathered in Fig. **??**. 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. **??**). 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:[9]

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

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 the whole pool of words available in the dataset for that feature.[10] 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

[9] 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.

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

as:[11]

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

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).

#16: (1) add H00 lines to account for semantic similarity (nothing more fancy since it involves a threshold), (2) rework results discussion

#17: show other feature variations and other models in annex

#17: explain binning

#18: tone down claims about contractile: it's a possible hypothesis if this were the only process, but not observed with the mix of all other processes

#15: relate to missed literature

Results are gathered in Fig. **??**. 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 ($\mathcal{H}_0$).

Word frequency, with ν_ϕ always significantly above ν^0_ϕ, 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^0_ϕ values, which is consistent with the literature on recall: words learned earlier and words with lower clustering coefficient are easier to produce than average (**??**). 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.

#16: add the taking-context-into-account parts, potentially moving the non-relative results discussions to here: (1) explain and add sentence-relative susceptibility by quartiles, which has a perfectly clear interpretation + results discussion, (2) explain and add sentence-relative variations + results discussion (3) link those results to non sentence-relative values

#17: show other feature variations and other models in annex

#15: relate to missed literature

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).

[11] 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 $\mathcal{H}_0$.

Discussion

Discuss related to introduction. Attractors, lineage with specification, what we couldn't observe, how it fits into Kirby.

#15: relate to missed literature

ADDTHIS: We also chose exploratory vs. predictive to give a detailed view of what happens and because there's too many possible things to predict.

ADDTHIS: By characterizing substitutions with 6 features on the disappearing and appearing words, we identify what makes a substitution more likely, and how a word changes when it is substituted. Consistent with known effects in linguistics, we observe that low-frequency words and words learned later in development are more susceptible to substitution than other words. Looking at the context those words appear in, we observe a marked effect for substitution of extreme words in a sentence (either very high-valued or very low-valued features compared to sentence average, except for word frequency). Focusing on how words are transformed, we see that the appearing words have significantly higher frequency and lower age-of-acquisition than synonyms of the disappearing word. Finally, the patterns we observe are also consistent with an attraction of each of the features towards a (feature-specific) asymptotic value.

ADDTHIS: It is possible however, that these attractors appear due to an interaction between biases and sentence context, making it a contingency rather than a rule. This is not really dealt with (context, aside from relevance) by Sperber.

Concluding remarks

#14: link to introduction discussion: (1) this can be a model system, (2) convergence can be looked for in any dimension, but that doesn't make a theory, so Epidemiology of Representations is all nice, but: (3a) taken simplistically it predicts obviously simplistic stuff (all quotes CV. to a single quote), (3b) taken more realistically (many dimensions in life) it gives some ideas, but it's not clear it's a core principle (3c) we need more controlled investigation to fuel the discussion and see how relevant it is.

ADDTHIS: On the other side an enactive proposition which anthropologists like Ingold, in line with Mauss' works, are calling for [Citation needed], is being developed by Froese, Di Paolo, and De Jaegher among others [Multiple citations needed].

The question is also gaining relevance in other fields, as work in evo-devo and non-genetic inheritance is accumulating evidence not accounted for by the modern synthesis [Citation needed]; these discoveries are creating demand for new or extended approaches to life evolution that unify its different levels, as well as creative empirical methods to test the predictions these approaches make [Citation needed].

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).

#18: tone down claims about contractile: it's a possible hypothesis if this were the only process, but not observed with the mix of all other processes

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, Nicolas Claidière and Nicolas Baumard for useful suggestions and comments.

This work has also been partially supported by the French National Agency of Research (ANR) through the grant Algopol (ANR-12-CORD-0018).

References

Aunger, R. (2000). *Darwinizing culture: the status of memetics as a science*. Oxford; New York: Oxford University Press. (OCLC: 44518383)

Bloch, M. (2000). A well-disposed social anthropologist's problems with memes. In *Darwinizing culture: the status of memetics as a science* (pp. 189–204).

Bourdieu, P. (1980). *Le sens pratique*. Paris: Editions de Minuit. (OCLC: 299354015)

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press. (OCLC: 11496588)

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: a quantitative approach*. Princeton, N.J.: Princeton University Press. (OCLC: 6863128)

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014, December). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society of London B: Biological Sciences*,

281(1797), 20141541. Retrieved 2015-03-31, from http://rspb.royalsocietypublishing.org/content/281/1797/20141541

Cointet, J. P., & Roth, C. (2009, August). Socio-semantic Dynamics in a Blog Network. In *International Conference on Computational Science and Engineering, 2009. CSE '09* (Vol. 4, pp. 114–121).

Cornish, H., Smith, K., & Kirby, S. (2013). Systems from Sequences: an Iterated Learning Account of the Emergence of Systematic Structure in a Non-Linguistic Task. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J., & Lee, L. (2012, March). You had me at hello: How phrasing affects memorability. *arXiv:1203.6360 [physics]*. Retrieved 2016-02-12, from http://arxiv.org/abs/1203.6360 (arXiv: 1203.6360)

Dawkins, R. (2006). *The selfish gene*. Oxford; New York: Oxford University Press.

Durkheim, E. (1912). *Les formes élémentaires de la vie religieuse le système totémique en Australie*. Paris: F. Alcan. (OCLC: 489968385)

Giddens, A. (1984). *The constitution of society: outline of the theory of structuration*. (OCLC: 11029282)

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information Diffusion Through Blogspace. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 491–501). New York, NY, USA: ACM. Retrieved 2016-06-22, from http://doi.acm.org/10.1145/988672.988739

Kirby, S., Cornish, H., & Smith, K. (2008, August). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. Retrieved 2016-06-18, from http://www.pnas.org/content/105/31/10681

Kroeber, A. L. (1952). *The nature of culture.* Chicago: University of Chicago Press. (OCLC: 487751)

Kuper, A. (2000). If memes are the answer, what is the question? In *Darwinizing culture: the status of memetics as a science* (pp. 175–188).

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 497–506). New York, NY, USA: ACM. Retrieved 2016-02-21, from http://doi.acm.org/10.1145/1557019.1557077

Liben-Nowell, D., & Kleinberg, J. (2008, March). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, *105*(12), 4633–4638. Retrieved 2016-06-22, from http://www.pnas.org/content/105/12/4633

Mauss, M. (1936). Les techniques du corps. *Journal de Psychologie*, *32*(3-4).

Mesoudi, A., & Whiten, A. (2008, November). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1509), 3489–3501. Retrieved 2016-06-18, from http://rstb.royalsocietypublishing.org/content/363/15

Miton, H., Claidière, N., & Mercier, H. (2015, February). *Universal Cognitive Mechanisms Explain the Cultural Success of Bloodletting* (SSRN Scholarly Paper No. ID 2560786). Rochester, NY: Social Science Research Network. Retrieved 2015-03-26, from http://papers.ssrn.com/abstract=2560786

Morin, O. (2013, May). How portraits turned their eyes upon us: Visual preferences and demographic change in cultural evolution. *Evolution and Human Behavior*, *34*(3), 222–229. Retrieved 2016-05-18, from http://www.ehbonline.org/article/S1090513813000056/ab

Simmons, M. P., Adamic, L. A., & Adar, E. (2011, July). Memes Online: Extracted, Subtracted, Injected, and Recollected. In *Fifth International AAAI Conference on Weblogs and Social Media.* Retrieved 2016-02-12, from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper

Sperber, D. (1996). *Explaining culture: a naturalistic approach*. Oxford, UK; Cambridge, Mass.: Blackwell.