PhD Dissertation

Epidemiology of Representations:
An Empirical Approach
—original title may change—

Sébastien Lerique[1]

Advisor: Jean-Pierre Nadal[2]
Co-advisor: Camille Roth[3]

[1]Centre d'Analyse et de Mathématique Sociales (CAMS, UMR 8557, CNRS-EHESS, Paris). Email: sebastien.lerique@normalesup.org.

[2]CAMS, and Laboratoire de Physique Statistique (LPS, UMR 8550, CNRS-ENS-UPMC-Univ. Paris Diderot, Paris). Email: nadal@lps.ens.fr

[3]CAMS, Centre Marc Bloch (CMB, UMIFRE 14, CNRS-MAEE-HU, Berlin), and Sciences Po, médialab (Paris). Email: camille.roth@sciencespo.fr

Contents

Chapter 1

Introduction

Current scientific knowledge describes the extraordinary complexity of life, and of human life in particular, through a wide array of theoretical and empirical approaches. Given the heterogeneity of the phenomena we aim to understand, we consider it no surprise that biology, psychology, cognitive science, linguistics, anthropology, sociology, or philosophy claim such varied problems and explanatory programmes. But a strong diversity also begs for questions: how can we bring together programmes which, at times, seem to talk past one another in spite of taking humanity as the same core object? Do different programmes always correspond to different explanatory levels, or should we rather combine them as interlocking aspects of the same unique level? Which of those programmes build on incompatible ontologies or world views, making their perspectives on life irreconcilable? These questions have interested a great number of researchers throughout the 20th century and up to now. Indeed, if we aim to carve life at its joints, the way we combine the diverse theoretical and empirical programmes that describe it will constrain the resulting bones and muscles carved out. In other words, choosing an assembly of fields goes hand in hand with a view of what it means to be human, of what exactly nature and culture correspond to, and of how we can best approach the emergence, complexity, and evolution of human life through time.

LOCATE

Over the last 40 years the field of cognitive anthropology, alongside with the approach to cultural evolution it suggests, has emerged as the strongest view of how science could combine the findings of anthropology, cognitive science, and biology. This approach, now "a booming cottage industry on the borders of evolutionary biology, archaeology, and biological anthropology" (following Sterelny 2017, and the labels he suggests), has two central traditions: the Californian, initiated by parallel works of Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985), and the Parisian, federated around the seminal work of Sperber (1996). The two traditions agree on the combination of fields they defend.

- it has a number of variants, of which DIT and CAT, which seem to agree on the philosophical framing of things, but have open debates and differing focuses

- it also has stark critiques

- at stake in those debates is the overarching vision/paradigm of how facets fit together, which defines the set of obvious and difficult things, guides the design of new questions, experiments, and implementations.

- recently new generations have developed in several fields: cognitive science is at a crossroads, evolutionary biology also. 4E / DST / Eco-evo-devo

- the particular question here is: observe actual phenomena, explain them. From there, what are the ways of accounting for all these phenomena in a unique framework, i.e. with a small number of guiding principles, that will give intelligibility at high level, and guide further questions. What are the differences between the various approaches, and which approaches are viable. Oyama 2001: "We believe that the heuristic value of the idea of developmental information in certain contexts is more than outweighed by its misleading connotations"

- How could there be any problems, and how is this a real question, since it's all a matter of empiricism? How is this a scientific question? Well it's a question at the level of scientific programmes

- First, the breadth of things you can look at is such that a theory will let you define what things are worth looking at. It also makes some cases more natural to explain than others. So a first question is, how much of the phenomena does it capture easily.

- Inside those things, some phenomena are not explained and we have a good idea of what could be the explanation

- Some other phenomena are not explained, and we have no idea how to tackle them, like the hard problem of content

- And the approach you take on those unsolved problems can be somewhat foundational for the rest of your theory, and makes other phenomena more or less natural to account for

- I don't claim to help with anything in the HPC. But I do think that existing theories can be moved forward by improving current empirical methods, to try and see how they pan out in relation to these hard problems.

FOCUS

- I will focus on CAT

- CAT, or ER, was developed by Sperber in the 90's, in a series of innovative articles

- explicitly not a reductionism, and not a Grand Unified Theory

- he proposes a general ontology where we think of things as representations, public/private that circulate, and that global dynamical system has attractors (that are more or less contingent on a situation in space and time)

- that lets us give an ontology for a framework, where we can rephrase long-standing anthropological questions in terms of spread of representations

- the hypothesis of attractors is the main added value of the theory: it claims that because of the interaction of biology, cognition, environment, and current cultural state, the space of how things evolve is somewhat skewed

- as can be understood by the recent name change, it's not clear that representations are the focus (e.g. distance when talking, or even the Claidière & Sperber 2007 example of smoking). I believe the central intuition was that psychology, in the form of at-the-time-emerging cognitive science, has a rich part to play in any synthesis, and is a core ingredient in the appearance of attractors.

- indeed, that is what is emphasised in debates with DIT

- also, it is not an explanatory theory in the sense of quantum physics (as it's not a GUT): Sperber reminds us that finding an attractor is only a cue to looking for its underlying causes. Instead it's a way of thinking at a high level, like a philosophy maybe

- so in a way, it can't be proved or disproved. But it can be more or less fruitful in generating situations and ideas, and providing a global assembly of facts

ARGUE/EXPAND

- up to now, empirical approaches fall into three categories: experimental transmission chains on simple content, compilation of historical works or data, social network data analysis

- each has its problems: transmission chains are on excessively simple content, historical compilations miss the variety of situations (lose detail) and are hard-put to distinguish explanations, social network analysis doesn't look at cognitive factors

- my focus will be on short utterances, bringing case studies of two types, one of each

- in-vivo, where there is huge complexity, but it's ecological

- in-vitro, where one can control the complexity a little, but you have to set a task and anything you do is subject to that choice

- I choose this for two main reasons: first, the availability of data, or relative ease in collecting it (though we'll discuss that). Second, language is at the core of the criticisms on CAT, so it's a good way to go to tease approaches apart.

- discuss the usefulness of those cases for CAT (and vice versa): does it help capture and account for the phenomena? And above all, do we observe attraction/convergence in these case studies? If not why not? Do we transform meaning in a particular fashion given a particular situation, and if such effects are observed can they be somehow generalised?

- show how developing these experiments helps:

- (1) formulate existing questions in more detail, and possibly ask new questions: convergence becomes a concrete question with real measures in these experiments, obstacles to or inappropriateness of convergence can become evident also

- (2) show the pain-points and problems in CAT. Notably:

    – the question of what information is, i.e. the information dualism of representations
    – -> level of description for convergence/transformation (can't decide on a principled level)
    – -> interpretation and meaning
    – taking context & environment into account, which is tackled by RT and Niche Construction, although that may not be enough, as it can maintain the dualism of nature/culture (might follow from the information dualism), even if you believe that it started early in biological evolution

OUTLINE

- I start with a detailed review of the literature on current approaches to cultural evolution, neighbouring areas, and parallels, alternatives, or downright criticisms
- I go on to detail the open problems of the field, seen both from its advocates and from its critiques
- I then explain how I contribute to those problems
- The second part details the first case study, an in-vivo "experiment" led on quotes on the internet
- The third part details the second case study, an in-vitro transmission chain experiment of short utterances of various types
- The final part revisits the contributions such experiments can bring to highlight (1) what has been unexplored, (2) the main challenge further case studies in this area are faced with and how it relates to criticisms, (3) how it would be possible to move forward.

1.1   Relevant works

1.1.1   20th century traditions

1.1.1.1   Social science

Social science has concerned itself with the stability, temporal evolution, and spatial variations and regularities of cultures since the start of its discipline. Émile Durkheim already, in his seminal study ([1897] 2012), was looking at the regularities of suicide rates over the years, and the correlation of those rates with a partition of society into religion-related groups. The continuity of suicide over time, and the links between suicide rate and social group, he argued, suggest we should study suicide as a *social fact*: a phenomenon which, in spite of manifesting itself individually, has emerging properties in large groups, with a causal life at the level of other emergent social phenomena. In particular, this emergent level has effects on individual psychology (one of Durkheim's aims was to establish the autonomy of sociology as a natural science of society). By what mechanisms does such an effect operate, and what role does it play in shaping the stability and evolution of cultures? Contemporaries of Durkheim, as well as later researchers, made this matter one of their central preoccupations.

But as those works also acknowledged, such a question raises issues in need of prior clarification: what is the exact status of culture in relation to psychology, or even biology, and how separable are they? Correspondingly, the question of how culture and psychology are part of one another and, if at all separable, how are the two related, or how best to describe the intermingling of these possible levels, has generated much debate throughout the 20th century.

Mauss (1936) initiated one approach to that issue, in his studies of the ways in which people of different societies, throughout history, use their bodies differently. He noticed and began documenting the resting postures, the attitudes, the ways of walking, of swimming, or of sleeping, that different communities adopt, pass on to their offspring, and evolve through time. His endeavour focused precisely on describing (parts of) culture as an embodied and physical property of life, incorporated through the everyday practices of a community, into which children grow by imitation, teaching, or other kinds of learning. In this sense, he argued, there is no normal way of walking, there are only *ways* of walking: by living with different bodily practices, different communities develop different bodies, none of which is standard. Indeed most of these *techniques of the body* documented by Mauss play a role in the physiological development of people growing into them, with noticeable effects: they will make one able to crouch for long periods of time or sleep while standing or horse-riding, and will also, Mauss argued for instance, change the silhouette of an adult body by influencing the way bones develop. While such practices are undeniably linked to the physiology and to the history of communities, and in that sense are both biological and social, Mauss also asked what part psychology plays in their development, and what influence they have in turn on psychology: he considered these practices complete *physio-psycho-sociological* phenomena.

The question of the incorporation of culture, and behind it the inseparability of culture, physiology, and psychology, has remained central in social science. Works closer to us have contributed to the matter: for Bourdieu (1980), societal norms are also incorporated in perception, and shape our everyday unreflective interpretation of events, our way of navigating life (our *sens pratique*). According to him, norms become embedded in the fullest sense, through life, into our low-level perception and exploration of the environment. Together, incorporated norms form what he calls a person's *habitus*, a concept he puts at the centre of his theory of social reproduction, building on the idea that members of a society grow into a *habitus* leading them to perceive events in ways that reinforce existing power structures.

Social scientists often criticise Bourdieu's approach for not providing a satisfactory account of individual agency, as in this view it still seems opposed to structure acting as a constraint on action. Another prominent approach, the *Theory of Structuration* from Giddens (1984), offers a more balanced way out of the

tension between agency and structure (or norms), and in doing so reflects yet a different view of the relation between psychology and culture. Acknowledging that approaches that conceive of structure as an external constraint on individual agency cannot resolve that opposition, Giddens (closer to the notion used in structuralist works) sees structure itself as a set of properties of social systems that bear an inherent duality, and cannot be meaningfully isolated from agency: on one side, action arises by using existing structured resources, and its reliance on, or referring to, such structure is what makes it action (vs. noise); on the other side, the production of action is a new reinforcement of the structure on which it relies. This notion of structure is akin to Saussure's notion of linguistic structure, which in turn inspired the structuralist tradition in social science: on one side, Giddens recalls, in producing an utterance we rely on incorporated syntactic rules; on the other side, the production of a meaningful utterance contributes to maintaining language as a structured totality.

ADD: According to Sperber (1996 : 31), there's also the diffusionist program, which is a precursor to his approach

Authors from Anthropology and Sociology have produced countless other insights about these problems; the miniature sample I exposed here represents the works which have most influenced the initial questions of this thesis. Common behind the works of Mauss, Bourdieu, Giddens, and authors contemporary to each of them, lies a certain interest in eliminating biology-culture, nature-nurture, or substance-form dualisms which we routinely rely on in our conception of life. This concern has remained central in contemporary social anthropology, and offers a valuable critique to the cultural evolution approach that I will focus on in what follows. I will come back to this line of work in Section 1.1.4, and discuss it in more detail in chapter 4.

TODO: Maybe add dropped authors? Simmel, Garfinkel, Geertz, Kroeber, Bateson

1.1.1.2 Psychology

TODO: ? - Gibson

1.1.1.3 Darwinism

TODO: ? - Plotkin, Murdoch, Sociobiology, Evolutionary Psychology, Memetics

1.1.2 Evolution of culture

Today's discussion of the ties between culture and psychology is more influenced by proponents from cognitive science. Inspired by the same initial analogy as darwinist authors from 20th century traditions, a new wave of analyses developed throughout the eighties and nineties and proposed a blend of (a) the gene-centred level of evolutionary theory, (b) insights that the application of generalised darwinian principles may (or may not, depending on the case) bring to the study of cultural change, and (c) a construal of the two processes of change, genetic and cultural, as parallel evolutions that actively interact, notably but not only through psychology.

Cavalli-Sforza and Feldman (1981) articulated this view in detail, by building on the fact that an evolutionary process need not be mediated by genetic transmission to take place. Indeed (and following the synthesis provided by Lewontin 1982), the combination of (a) any transmission process leading to some nonrandom dependence of offspring phenotype on parent phenotype, (b) at least some variation of the phenotypes (against which to select), and (c) some differential survival and spread depending on phenotype, will lead to evolution by natural selection. By looking at genetic models of evolution as a special case of those general principles of evolution, Cavalli-Sforza and Feldman (1981) developed the mathematical

analysis of purely phenotypic transmission, which can take different paths in a given population: vertical (from parent to offspring), oblique (from a non-parent member of the previous generation, to member of the next generation), or horizontal (inside one generation). Boyd and Richerson (1985; see also Boyd and Richerson 2005) further developed this line, leading to the formulation of *Dual Inheritance Theory* (or DIT) which offers a systematic analysis of some interactions between cultural and genetic evolution, an approach which is now vibrant with empirical work (see Acerbi and Mesoudi 2015 for a review of recent studies). A notable feature of this programme is that it does not constrain itself into a particular view of what culture is. While the main authors do define culture as "information that people acquire from others by teaching, imitation, and other forms of social learning" (Boyd and Richerson 2005, 3), a definition which at first sight might prove difficult to reconcile with a non-informational view of culture (such as the incorporated views from Section 1.1.1.1), the mathematical models do not constrain the concept of culture as much. The approach also has deep links with the analysis of *Niche Construction* (e.g. Odling-Smee, Laland, and Feldman 2003), which offers promising steps towards a reconciliation with non-informational views of culture stressing the importance of the development process in evolution. I return to this subject in more detail in Section 1.1.3.

In the mid-nineties Dan Sperber formalised a second influential approach to the question of the evolution of culture: in a series of innovative articles gathered in Sperber (1996), the author puts forward a research programme called *Epidemiology of Representations* (better known today as *Cultural Attraction Theory*, or CAT), and seeks to provide the cognitive and social sciences with a common framework with which to address interdisciplinary questions. One of the guiding questions of Sperber's work is the following: how can we explain both the diversity culture across regions, and its relative stability through time, knowing that all human beings are more or less made of the same ingredients? In developing an answer, Sperber commits himself to presenting a coherent ontology where the status of each object he refers to is well defined.

The framework he suggests then starts from an ontology made of "mental representations", which correspond to those defined and studied by classical cognitive science, and "public representations", which are the expressions of mental representations in diverse cultural artefacts such as pieces of text, utterances, pictures, myths, built structures, etc.. New mental representations are constantly formed in people's minds whenever they perceive or interpret public representations. For instance, say I am thinking of a tune (mental representation), and I whistle it (public representation); someone else hears it, and forms their own mental representation. Most of the time, the new representation in that person's head is different from my original representation. This last point is a defining feature of what Sperber proposes, in contrast to memetics of course, but also to DIT, outlined above (Sperber 1996, 25–26, 31).

On this basis, Sperber proposes to model human societies as large dynamical systems of people continuously interpreting public representations into mental representations, and producing new public representations through their situated actions (in which mental representations play a role). To explain culture then, in this framework, is to analyse the processes by which representations circulate through a society, with different levels of change along the way. Those processes are many and heterogeneous, which corresponds to the diversity of cultural domains that exist in societies, but the basic ontology remains grounded in the same mental representations of cognitive science.

By developing such an ontology to connect disciplines, Sperber proposes a credible bridge between the notions of representation in social science and that of mental representation in cognitive science, without reducing one area to the other or making simplistic assumptions about the phenomena encountered. In this sense, his proposal is that of a naturalistic ontology for the study of culture which builds on cognitive science principles. It is amenable to anthropology, and encourages the combination of the two bodies of knowledge in a well defined way. Sperber can then rephrase interdisciplinary questions in terms of spread and transformation of representations. For instance: what types of representations are less transformed than others as people integrate them or perceive them, and produce them anew, making them circulate in a society? Such representations, spreading wider than others, become *cultural representations* and will characterise a given society (for instance the habit of clothing, eating customs or values, or technological

knowledge).

Why are those representations so stable, and how do they evolve? Sperber introduces an additional concept to analyse this evolution: the dynamical system which represents a society exhibits attractors, called *cultural attractors*, that depend on the complex interaction of psychological and ecological factors, and on the distribution of representations at a given moment in time (Sperber 1996, 106–18). Cultural attractors are one of the core concepts in CAT providing intelligibility to the evolution of culture and to the reciprocal influence of psychology, culture, and environment. As such, a key goal in the CAT research stream has been that of identifying existing attractors, and explaining their emergence based on the psychological and ecological factors that create them.

The most important intuition in Sperber's proposal, and what differentiates it from previous works, is the centrality of psychology for the evolution of culture (Sperber 1996, 31) and its role in the emergence of attractors. He substantiates this by relying heavily on contemporary cognitive science, and in particular by adopting and extending the view of the modularity of mind initially defended by Fodor (1983) (his view goes further than Fodor's, as he argues for a *massive modularity of mind* applying not only to perception but also to conceptual processes). Combining this view with epidemiology of representations, Sperber argues, results in a theory avoiding both the blank slate approach to psychology (Sperber 1996, 63–66) and a naive application of neo-darwinist formalism to the specific case of culture (Sperber 1996, 101), while still being able to account for the diversity of cultures (Sperber 1996, 120). This specificity of CAT is often highlighted as one of the main differences with DIT (Sterelny 2017, 47); CAT agrees nonetheless with the idea of gene-culture coevolution developed by DIT (Sperber 1996, 114), and most authors consider that both theories are compatible but have different core focuses (Sterelny 2017; Acerbi and Mesoudi 2015).

The space opened by the development of DIT and CAT has generated much debate (see e.g. the peer commentary to Mesoudi, Whiten, and Laland 2006) and some heated criticisms (Ingold 2007; Mesoudi, Whiten, and Laland 2007). Section 1.1.4 comes back to these debates and criticisms, which I group into two broad categories. The first concerns the specifics of the evolutionary approach adopted by DIT and CAT: what evolutionary mechanisms should they take into account (for instance niche construction, or epigenetic or extended inheritance) and, correspondingly, how long the restriction of inheritance to two isolated and parallel channels will remain the best approximation. The second critique concerns what Thompson (2007) has termed the "information dualism" of the notion of representation in cognitive science: at the concrete level, it questions whether a representiationist approach is able to naturalise meaning, a core aspect in the analysis of culture and life; at the programmatic level it suggests that, compared to a focus on dynamic couplings, a focus on representations in cognitive science has less heuristic value in guiding the exploration of the interactions between culture and psychology. Regardless of whether proponents can reconcile DIT and CAT with the second critique, I believe both approaches can benefit greatly from these debates, and that empirical study has a crucial role to play in exploring the implications of each side.

The empirical exploration of CAT has proven a difficult task, for two main reasons. First, Sperber clarifies that CAT is not, and should not be, a "grand unitary theory":

> "An epidemiological approach … should not hope for one grand unitary theory. It should, rather, try to provide interesting questions and useful conceptual tools, and to develop the different models needed to explain the existence and fate of the various families of cultural representations." (Sperber 1996, 83)

Indeed, CAT accomplishes this:

> "What the epidemiological analogy suggests is a general approach, types of questions to ask, ways of constructing concepts, and a plurality of not too grand theoretical aims." (Sperber 1996, 61)

As a consequence, testing CAT means evaluating the fruitfulness of its paradigm. No one study, or col-

lection of studies for that matter, will reach a yes/no answer to the validity of CAT. But we can reach a collective consensus about the usefulness of approaching the psychology-culture link with the tools and questions developed by that theory.

As the name CAT indicates, an important part of that toolbox is the notion of cultural attractor, which is amenable to empirical study: the presence of a well-defined attractor in a given situation is a testable hypothesis, and approaching concrete domains from that angle can be more or less fruitful in inspiring studies and questions. Now quantitative data on out-of-laboratory cultural artefacts is not easy to collect, which is the second reason empirical exploration is a challenge. Although theoretical models to guide that exploration are gradually appearing (Claidière and Sperber 2007; Claidière, Scott-Phillips, and Sperber 2014), developing methods for the quantitative study of attractors is still an open problem. Studies that can be explicitly linked to CAT involve an array of methods. The meta-analysis of large numbers of anthropological or historical works has shown interesting results. Morin (2013), for instance, shows how painted portraits change over the centuries to increasingly use direct-gaze (vs. oblique) portraying. Miton, Claidière, and Mercier (2015) also used this technique to explain the historical success of the bloodletting practice, in spite of its medical ineffectiveness. Baumard et al. (2015) fruitfully exploited similar records to tie the evolution of religious values to the affluence of societies in which they develop. Morin and Miton (EHBEA 2017? [Citation needed]) model the evolution of heraldic coats of arms to test which hypotheses account best for the distribution of patterns and colours through time. A second approach simulates rapid evolution of content in artificial transmission experiments. Claidière et al. (2014) explored the evolution of abstract visual patterns transmitted by apes; Claidière, Trouche, and Mercier (2017) also investigated the effect of argumentation on the transmission of solutions to simple but counter-intuitive reasoning problems. A large number of disciplines are developing diverse methods to study questions relevant to CAT and DIT. I review these in greater detail in Section 1.1.3.

While authors have acknowledged there is no core incompatibility between CAT and DIT (Acerbi and Mesoudi 2015; Sterelny 2017) and a growing number of social scientists is discussing both approaches (see again the comments to Mesoudi, Whiten, and Laland 2006; also Slingerland 2008), the work I present in this thesis is mainly focused on CAT itself, and less so DIT. While the prevalence of the theory in the French academic landscape naturally had an impact on this choice, CAT is also interesting to focus on for more valuable reasons. First, it puts cognitive science squarely in the middle of the problem, a focus I believe to be unavoidable. Second, its initial goal is not so much to develop a mathematical theory of culture (which will surely continue developing), but to interest all disciplines studying human life in a common framework, by using *principled philosophical arguments*. Third, this approach in turn leads to a crystal-clear articulation of the philosophical principles defended, discussed, or tested by the theory. As such, it seems to me that CAT is in a good position to generate productive debate between alternative approaches which, for now, are still competing on the principled level (such as those discussed in Section 1.1.4). In particular, the challenges that arise when attempting to study CAT experimentally, as is the case here for linguistic utterances, will prove helpful in developing a path to the resolution of the critiques faced by gene-culture coevolution approaches.

1.1.3   Neighbouring areas

1.1.4   Criticisms and alternatives

1.2   Open Problems

1.3   Stakes

Chapter 2

Brains Copy Paste

2.1 Introduction

The reciprocal influence between cognition and culture has a long history in both social science and psychology. While this question has been the subject of intense debate in the social sciences in the 20th century, today's discussion is mostly structured by proponents from cognitive science, who construe culture as an evolutionary process analogous and parallel to biological evolution. That analogy can be traced a long way back, with milestones such as Kroeber's works (1952), Dawkins' *Memetics* ([1976] 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 memetics cannot account for the levels of transformation culture undergoes as it is transmitted. Mesoudi and Whiten (2008) discuss 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, demonstrate how cognitive biases have influenced the evolution of cultural artifacts over several centuries. Kirby, Cornish, and Smith (2008) and Cornish, Smith, and Kirby (2013) 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 et al. (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 ("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 above-cited studies 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 cognitive effects in transformations bring 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 as 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, Claidière, and Mercier 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 et al. 2004) and in emails (Liben-Nowell and Kleinberg 2008); J. P. Cointet and Roth (2009) 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 et al. (2012) studied the characteristics of particularly memorable quotes that circulate in those networks. We believe these works can connect the field of cultural evolution with psycholinguistics to advance the testing of cultural attractors.

To show this we analyze how 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, and Adar 2011): authors spontaneously transform quotes, not only cropping them but also replacing words. For instance the quote "we will not be scared of these cowards" (a substring of a quote from former Pakistani President Asif Ali Zardari) is also found as "we will not be **afraid** of these cowards". More meaningful changes often happen too, such as the transformation of McCain's "I admire Senator Obama and his accomplishments" during the 2008 US presidential campaign, into "I **respect** Senator Obama and his accomplishments". Since authors are implicitly required to copy quotes exactly, we can assume that most transformations, especially simple ones such as those shown above, are the result of automatic (i.e. hard to control) low-level cognitive biases of the authors.

We thus ask the following question: given such representations that seem to evolve precisely because of the kind of automatic cognitive biases evoked in the theory of epidemiology of representations, do cultural attractors appear, and if so how do cognitive biases participate in them? We chose to restrict our analysis to substitutions (i.e., one word being replaced by another), both to keep the analysis tractable and because of missing information in our data set.[2] While this limits the scope of our results to the particular data set we use, the methodological point we also make is left intact. By characterizing words using 6 well-studied 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, Cornish, and Smith (2008), yet with natural language out of the laboratory.

---

[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 confined to) 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.

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 introduce the measures we built to observe 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.

2.2 Related work

The study of cultural evolution on the part of cognitive science emerged only recently. While formal models of cultural transmission appeared with the development of dual inheritance theory (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985) and have included the notion of cultural attractor since then (Claidière and Sperber 2007; Claidière, Scott-Phillips, and Sperber 2014), collecting data to test and iterate over such models has been more challenging. The first above-mentioned method consists in rebuilding the history of a given type of representation by compiling anthropological or historical works on the subject (as for instance Morin 2013; and Miton, Claidière, and Mercier 2015, have done). A second approach uses cultural evolution experiments in the laboratory, with an array of methods reviewed by Mesoudi and Whiten (2008). Transmission chains, in particular, have been used extensively to study the evolution of human language (see Tamariz and Kirby 2016 for a review). Other recent examples include studies of the evolution of simple audio loops through consumer preference (MacCallum et al. 2012), the emergence of structure in visual patterns transmitted by baboons (Claidière et al. 2014), and the amplification of risk perception through chains of casual conversation (Moussaïd, Brighton, and Gaissmaier 2015).

Research on online content points to a third approach to this question. By investigating the transformations of quotations in a large corpus of US blog posts and online news stories initially collected and studied by Leskovec, Backstrom, and Kleinberg (2009), Simmons, Adamic, and Adar (2011) and later Omodei, Poibeau, and Cointet (2012) show that even for quotations, a type of public representation that should change the least when transmitted on the Web, it is still possible to witness significant transformations. These studies focus on the influence of the quotation source (e.g. news outlet vs. blog) or of the surrounding public space (e.g. quotation frequency in the corpus), and suggest diffusion-transformation models to capture the dynamics of the population of quotations. But the cognitive features which may determine or, at least, influence these transformations, are overlooked. On the other hand cognitive and linguistic features have been used in diffusion studies not involving transformation: Danescu-Niculescu-Mizil et al. (2012), for instance, show that particularly memorable quotations (taken from movie scripts in this case) use more distinctive words and have more common syntax than less memorable quotations; they are also the quotes that adapt best to new contexts of use. One source of ideas to study the transformations of such quotes, then, might be the psycholinguistic literature studying word and sentence recall.

Mary C Potter and Lombardi (1990) suggest that immediate recall of sentences is based on the retention of an unordered list of words which is then regenerated as a sentence at the moment of production. Priming recall with other words can lead to replacement in the recalled sentence if the primed words are consistent with the overall meaning of the sentence. Regenerated syntax can also be influenced by priming recall with another syntactic structure (Mary C. Potter and Lombardi 1998), or with verbs whose category constraints call for a different structure (Lombardi and Potter 1992).

Compared to full sentences, recall of word lists provides a situation that is easier to fully explore and has been extensively studied. In particular, the Deese, Roediger, and McDermott paradigm (introduced by Deese 1959; and later popularized by Roediger and McDermott 1995) has shown that it is possible to construct lists of words which reliably create the false memory of an external word related to those in the list. This is done by using lists of words produced by free association from the target intrusion word; the intruding recall then happens with probability nearly proportional to the average semantic association strength between the intruding word and the words in the list. A sizable literature studies this type of task

with varying complexities in the design of the lists, a good review of which is given by Zaromb et al. (2006). One notable effect is that the semantic relations between words greatly influence, and correlate to, the order in which words are recalled (Tulving 1962; M. W. Howard and Kahana 2002), and that this reordering of items improves subjects' repeated recalls (Tulving 1966). The frequency and type of intrusions in lists of random words are also influenced by associations created by the presentation of previous lists (Zaromb et al. 2006). Indeed, the question of how such temporal associations (contributing to contextual information retrieval in recall) interact with the prior semantic associations of subjects (contributing to associative information retrieval) is at the core of many of these studies.

These effects do not transpose simply to sentence recall however, as not only syntax but also effects of attention come into play for both retrieval and encoding. Jefferies, Lambon Ralph, and Baddeley (2004), for instance, show that attention is central to the encoding and retention of unrelated propositions, on top of more automatic syntactic and semantic processes. This involvement of executive resources also seems to contribute to the much greater memory span subjects exhibit for sentences compared to word lists (see Jefferies, Lambon Ralph, and Baddeley 2004 again, for more details).

Given this complexity we decided to focus on more aggregate measures, where variations of the conditions in which sentences are read and produced have a chance of being statistically smoothed out.[3] If a cognitive bias in the substitution of words manifests itself with simple measures, then it will be worth applying predictive models of the substitution process in further research.

Lexical features, then, are obvious well-studied word measures that can be analyzed in aggregate. Indeed word frequency (see Yonelinas 2002 for a review), age-of-acquisition (Zevin and Seidenberg 2002), number of phonemes (see for instance Rey et al. 1998; Nickels and Howard 2004), and phonological neighborhood density (Garlock, Walley, and Metsala 2001) to name a few, all have known effects on word recognition or production. More complex features based on word networks built from free association or phonological data have also been analyzed: Nelson et al. (2013) for instance, show the importance of clustering coefficient in such a semantic network by studying the role it plays in a variety of recall and recognition tasks (extralist and intralist cuing, single item recognition, and primed free association). Chan and Vitevitch (2010) show that pictures are named faster and with fewer mistakes when they have a lower clustering coefficient in an underlying phonological network. 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 is a link between the ease of recall of words and their authority position (pagerank) in a language-wide semantic network built from external word association data (Austerweil, Abbott, and Griffiths 2012 further develop this tool to give a parsimonious account of the fact that related words are often retrieved together from memory).

On the whole, research on lexical features hints towards two antagonistic types of effects (also known as the 'word-frequency paradox', Mandler, Goodman, and Wilkes-Gibbs 1982). On one hand, part of the literature shows 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 retrieval, that is 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 rewrite a given utterance, is rather of the former or latter sort. We also aim to shed some light on this debate, considering oddness as a dimension of the purported fitness of utterances.

---

[3] Aside from our lack of control on the precise conditions of encoding and recall in our data set, the analysis techniques mentioned above are better suited to data consisting of a high number of measures over a smaller number of lists (in which case it makes sense to ask e.g. what proportion of intrusions come from prior lists). As is explained further down however, our data set is shaped the opposite way: a great number of sentences, with only very few to no measures at all on each sentence.

2.3 Methods

We rely on a text corpus made of quotations extracted from online blog posts, and focus on their evolution. Quotations appeared to be a perfect candidate to propose a first measure of automatic cognitive bias in cultural transmission. 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 are ideally 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. Since our approach is exploratory however, we do not know at the outset which precise effect of cognitive bias we are looking for. Indeed, the data we use does not come from a controlled experiment in the laboratory, designed to elicit a particular effect: they are recordings of real life interactions, with all the complexity and uncertainty of conditions this entails. Our goal, therefore, is to show that cognitive biases have measurable effects in this setting even if they are part of a larger complexity (the detailed prediction and deconstruction of the cognitive processes responsible for them being left to further research). If this is confirmed, we will have successfully tested fundamental cognitive biases with out-of-laboratory data, opening a path to explanations of actual (vs. simulated) cultural evolution with tools from cognitive science. Aiming to exhibit such subtle biases in complex data is the main reason we chose to use aggregate measures that have a chance of smoothing out the possible variations of experimental conditions in the data set.

To keep the analysis tractable, we focused on quotation transformations consisting of the *substitution* of a word by another word (and only those cases) in order to unambiguously discuss single word replacements. This restriction also allows us to more reliably infer the information that is missing in our data set, as explained in the "Substitution model" section. To quantify these 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 data set and the measures we used to assess this cognitive bias.

2.3.1 Corpus-based utterances

We used a quotation data set collected by Leskovec, Backstrom, and Kleinberg (2009), large enough to lend itself to statistical analysis. This data set 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, and Kleinberg 2009).[4] The authors automatically extracted quotations 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."

Then, the authors gathered quotations in a graph and connected each pair that differed by no more than one word or that shared at least ten consecutive words (they tested this procedure with a number of different parameters, see Leskovec, Backstrom, and Kleinberg 2009, for more details). We find for example the following variation of the above quote:

> "these operations are **intended** to address funding pressures over quarter-end."

---

[4]The original article (Leskovec, Backstrom, and Kleinberg 2009) does not provide further details on the source selection methodology.

Next, they applied a community detection algorithm to that quotation graph to detect aggregates of tightly connected, that is sufficiently similar, groups of quotations (see again Leskovec, Backstrom, and Kleinberg 2009, 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 ideally contains all variations of a same parent utterance, along with their respective publication URLs and timestamps (since the procedure cannot be perfect, sets of quotations contain occasional rogue unrelated variations that should have been discarded or assigned to another set).

Manual inspection of this data set 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 quotes with less than 5 words or whose occurrences span more than 80 days (indicating causally unrelated occurrences), as well as quotes not written in English. Clusters that are emptied by this procedure are therefore excluded. If, after this screening, a cluster's occurrences still span more than 80 days (because of short-lived but unrelated quotes far apart in time), we also exclude it. We eventually keep 50,427 clusters (out of 71,568; i.e. 70.5%), containing a total of 141,324 unique quotes (out of 310,457; i.e. 45.5%) making up about 2.60m occurrences (out of 7.67m; i.e. 33.9%).[5] Even if we lose some real event-related utterances which are present in clusters lasting more than 80 days (one such lost quote, for instance, is "the city is tired of me and the organization and I have run our course together"), we check that our filtering approach fulfills its goals by coding a random sub-sample of 100 clusters: 35 of them are rejected by the filter, with 15 false negatives (rejected clusters that should have been kept) and 9 false positives (clusters kept when they should have been rejected), giving a precision score of 0.862 and a recall score of 0.789. Furthermore, all but one of the 9 false positives are left with a single non-rejected quote, meaning those clusters are ignored by our substitution analysis; this brings the effective precision of our filter to 0.982.[6]

2.3.2 Word-level measures

2.3.2.1 Lexical features

We first introduce some lexical measures on words.

- **Word frequency**: the frequency at which words appear in our data set, 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, and Brysbaert 2012), known to have different effects than word frequency (Morrison and Ellis 1995; Dewhurst, Hitch, and Barry 1998),
- **Phonological** and **Orthographic Neighborhood Density** (obtained from Marian et al. 2012), also known to be relevant for word production (Garlock, Walley, and Metsala 2001),
- 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)[7], as well as **Number of Letters**, 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.

---

[5]The significantly larger loss in occurrences indicates that, on average, the clusters we lose contain more occurrences than those we keep, which is to be expected for everyday language utterances.

[6]A similar analysis was made for language detection, which is part of the cluster filtering: out of 100 randomly sampled quotes, 17 are rejected because their detected language is not English, with no false positives and 6 false negatives, giving a precision score of 1 and a recall score of 0.933. Of the 6 false negatives, 4 had less than 5 tokens and would have been excluded by the cluster filter anyway.

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

We also consider grammatical types within quotations by detecting *Part-of-Speech* (POS) categories with TreeTagger (Schmid 1994); we distinguish between verbs, nouns, adjectives, adverbs, and closed class-like words.

Aside from these raw features, the systemic dimension of vocabulary (Cornish, Smith, and Kirby 2013) has led authors to develop measures based on the full topology of networks built from free association data or phonological similarity. Several such measures have been shown to be involved in recall, recognition, and naming tasks (Nelson et al. 2013; Chan and Vitevitch 2010; Griffiths, Steyvers, and Firl 2007).

To compute these features we relied on the free association (FA) norms collected by Nelson, McEvoy, and Schreiber (2004), which record the words that come to mind when someone is presented with a given cue. As Nelson, McEvoy, and Schreiber (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." Similar to what Griffiths, Steyvers, and Firl (2007) did, we first considered the directed weighted network formed by association norms, where nodes are words and edges are directed from cue to target word, with a weight equal to the association strength (that is the probability of that target word being produced when this particular cue is presented). This network is of particular interest since it lets us define features that reflect the associations driving false memories in word lists (Deese 1959), a phenomenon which may be involved in the transformation of quotations.

We used three standard measures on the FA network:

- **Incoming degree centrality**, measured by the number of cues for which a given word is triggered as a target, and a corresponding generalized measure, node **Pagerank** (Page et al. 1999), which has already been used on the FA network by Griffiths, Steyvers, and Firl (2007). In the present case these two polysemy-related measures are quasi-perfectly correlated.[8]
- **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 as unavoidable waypoints on association chains connecting one word to another.[9]
- **Clustering coefficient**, which measures the extent to which a node belongs to a local aggregate of tightly connected nodes (Watts and Strogatz 1998), computed on the undirected weighted version of the FA network.[10] This tells us if a word belongs more or less to a group of equivalent words (from a free association point of view).

2.3.2.2 Variable correlations

Several of these features are strongly related and can be grouped together. To make correlation values as well as future comparisons more reliable, we log-transformed features that have marked exponential distributions (i.e. a few words valued orders of magnitude higher than the vast majority of other words).

The pairwise correlations in the initial set of features appears in Fig. 2.1. By looking at absolute values, three subsets of highly correlated features can be easily identified: (a) number of letters, phonemes, and syllables with pairwise correlations greater than .75; (b) orthographic and phonological neighborhood densities, with a correlation of .8; (c) age of acquisition, betweenness, degree, and pagerank centralities, with absolute pairwise correlations at .41, .59, .6, .61, .63 and .85. Applying a feature agglomeration algorithm targeted at 6 groups refined this observation by producing identical (a) and (b) groups, a (c) group

---

[8]Note that in-degree does not take the weights of links into account, as it counts 1 for each incoming link. Pagerank on the other hand, does take the weights into account.

[9]For this measure, weights are interpreted as inverse cost: the stronger a link, the easier it is to travel across it. A stronger link will be favored over weaker links in the computation of the shortest path between two words.

[10]he Clustering coefficient is formally defined as the ratio between the number of actual versus possible edges between a node's neighbors; this is poorly defined in the case of directed networks, which led us to ignore the direction of links in the network for this measure (if two words are connected in both directions, the weights of both links are added to make the final undirected link's weight).
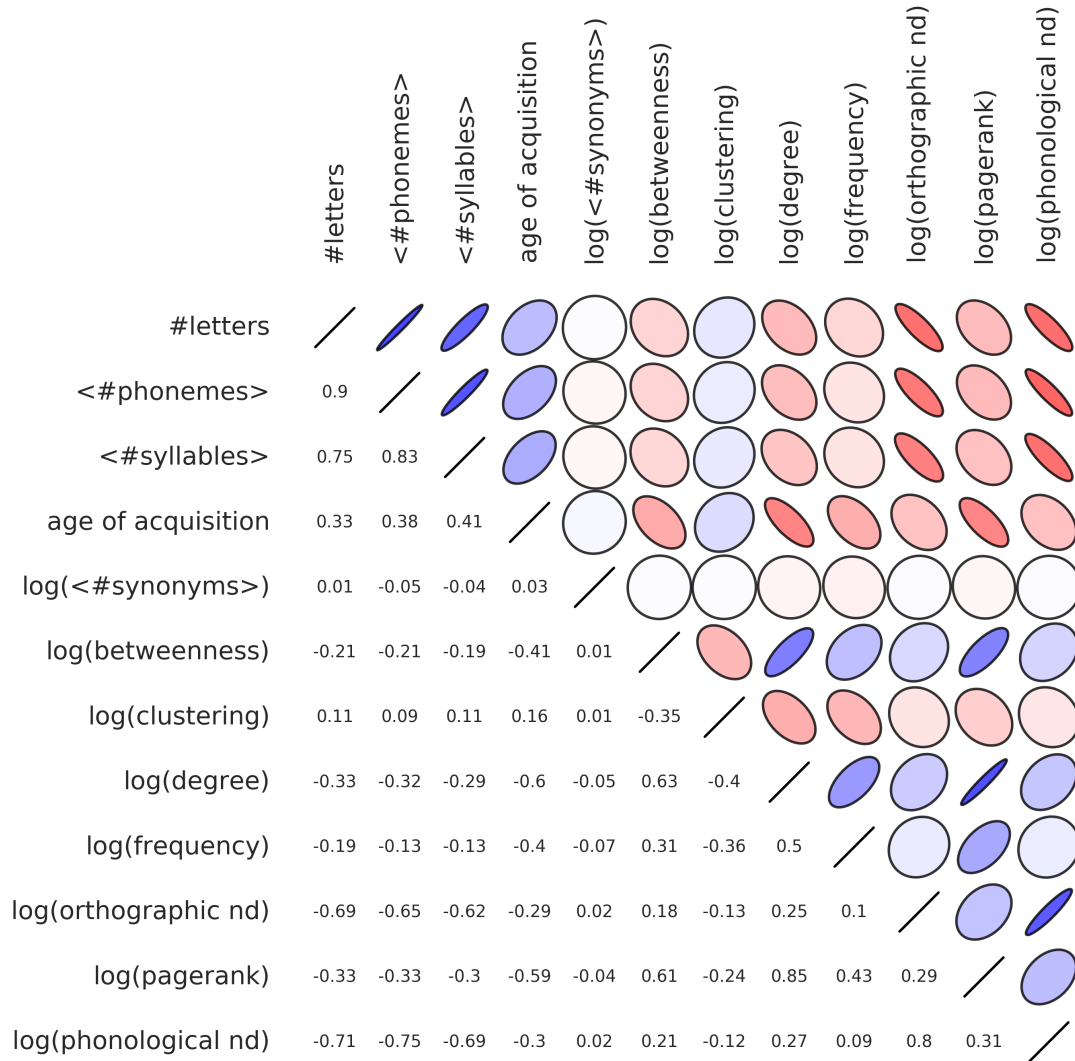
	#letters	<#phonemes>	<#syllables>	age of acquisition	log(<#synonyms>)	log(betweenness)	log(clustering)	log(degree)	log(frequency)	log(orthographic nd)	log(pagerank)	log(phonological nd)
#letters												
<#phonemes>	0.9											
<#syllables>	0.75	0.83										
age of acquisition	0.33	0.38	0.41									
log(<#synonyms>)	0.01	-0.05	-0.04	0.03								
log(betweenness)	-0.21	-0.21	-0.19	-0.41	0.01							
log(clustering)	0.11	0.09	0.11	0.16	0.01	-0.35						
log(degree)	-0.33	-0.32	-0.29	-0.6	-0.05	0.63	-0.4					
log(frequency)	-0.19	-0.13	-0.13	-0.4	-0.07	0.31	-0.36	0.5				
log(orthographic nd)	-0.69	-0.65	-0.62	-0.29	0.02	0.18	-0.13	0.25	0.1			
log(pagerank)	-0.33	-0.33	-0.3	-0.59	-0.04	0.61	-0.24	0.85	0.43	0.29		
log(phonological nd)	-0.71	-0.75	-0.69	-0.3	0.02	0.21	-0.12	0.27	0.09	0.8	0.31	

Figure 2.1: Spearman correlations in the initial set of features

without betweenness centrality which was instead assigned to a group (d) with clustering coefficient, and the remaining features (frequency and number of synonyms) as singletons.[11]

Since our data is about written transformations, number of letters and orthographic neighborhood density are the natural representatives of groups (a) and (b) respectively. Given the importance of age of acquisition in the lexical feature literature, we chose it to represent group (c). Finally we used clustering coefficient to represent group (d) since it has already been used in previous studies. The final set of features we will discuss in the rest of the paper, as well as their cross-correlations, can be seen in Fig. 2.2.[12]
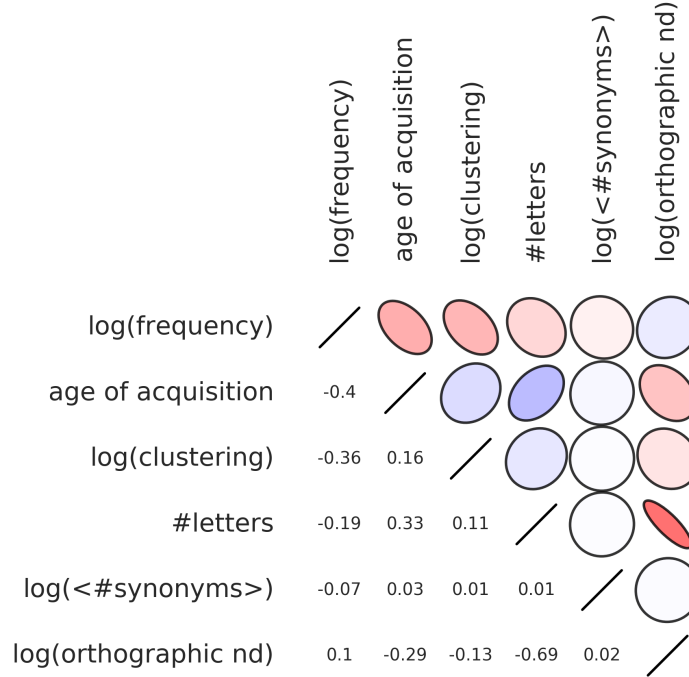
Figure 2.2: Spearman correlations in the filtered set of features

2.3.3 Substitution model

We finally need a substitution detection model, for the quotation data we use presents a challenge: quote-to-quote transformations and substitutions are not explicitly encoded in the data set. 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. In other words we face an inference problem where, given all quotations and their

---

[11]Agglomerating into less than 6 groups merged groups (a) and (b), which we excluded to keep neighborhood densities in their own group; agglomerating into more than 6 groups separated age of acquisition from group (c), which we excluded given its high correlation values to the rest of group (c). We used scikit-learn's FeatureAgglomeration class for this procedure (Pedregosa et al. 2011).

[12]Note that feature values stem from different data sets which do not always encode the same words. Indeed, we have data on frequency for about 33.5k words, on age of acquisition for 30.1k words, on clustering coefficient for 5.7k words, number of synonyms 111.2k, and orthographic density 17.8k words. Quite often then, not all features are available for a given word in our data set; 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.

occurrence timestamps, we must 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. Given a particular occurrence of a quotation, the first issue is deciding whether that occurrence is a strict copy of an earlier occurrence, or a substitution from another quotation, or maybe a substitution or copy from quotes appearing outside the data set, that is from a source external to the data collection perimeter. The second issue is deciding which source originated such a substitution when several candidate sources are available.

Let us give an example: say the quotation "These accusations are false and **absurd**" (q) appears in two different blogs on January 19, and the slightly different quotation "These accusations are false and **incoherent**" (q') appears in another blog on the 20th of January. The second occurrence of q can safely be assumed to be a faithful copy of the first one the same day. And since q is fairly prominent when q' first appears, we could assume that the author of q' on the 20th based herself on q, as is shown with a dashed line in Fig. 2.3. Now say a third version, "These **allegations** are false and **incoherent**" (q'') also appears once on January 19 and once on January 20 after q'. Here, q and q'' differ by two substitutions, so we discard the possibility that one was written based on the other (see below for further details). q'' is only one substitution away from q' however, so we could also consider the first occurrence of q'' as a potential source for q' on the 20th. Conversely, the occurrence of q'' on the 20th could be considered as a substitution from q', or as a faithful copy from its initial occurrence on January 19. (Options shown in Fig. 2.3.)
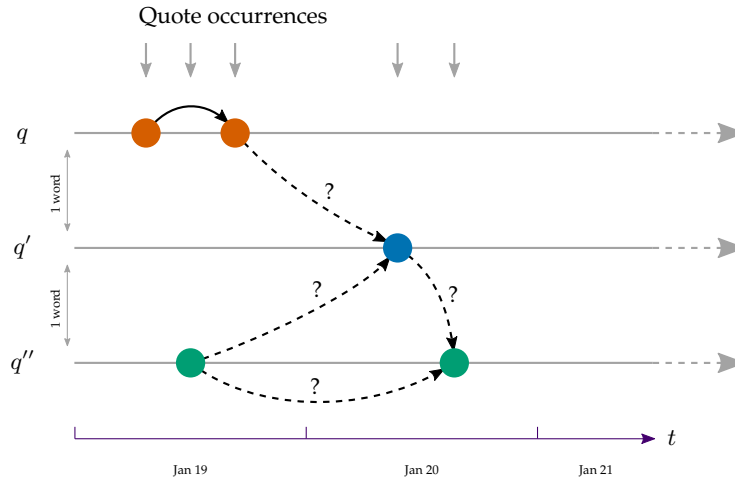

Figure 2.3: **Possible paths from occurrence to occurrence.** q, q' and q'' are three quotation variants belonging to the same cluster. q and q'' differ by two words, but q' differs from both q and q'' by one word. The second occurrence of q can safely be considered a faithful copy of the first, but the occurrences of q' and q'' are uncertain: while the first occurrence of q' is most likely a substitution from q, it could also stem from q''; conversely, the second occurrence of q'' could also be a substitution from q' instead of being a faithful copy of its first occurrence.

One way to settle these questions is the following: 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 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. Fig. 2.4a shows the inferences made by such a model. The assumptions it embeds, however, are a subset of a much wider set of possibilities, each leading to alternative inferences.

(a) Source must be majority in preceding bin, destination can be anything

(b) Source must be majority in preceding bin, destination must not appear in preceding bin

(c) Source can be anything, destination can be anything

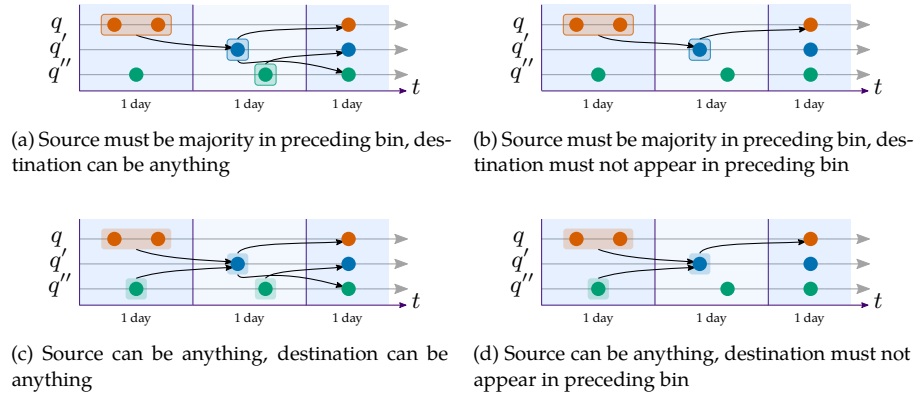(d) Source can be anything, destination must not appear in preceding bin

Figure 2.4: **Substitution models.** Substitutions inferred by four models in the situation introduced by Fig. 2.3. Each of these models uses bins spanning 1 day aligned to midnight (see the main text for a complete description of parameters). In the top left panel (a), q holds the majority in the first bin and is considered the unique basis for q' in bin 2. q' and q'' have equal maximum frequency in bin 2 however, so both are sources of substitutions towards bin 3. In the top right panel (b), quotes that appear in the preceding bin cannot be the target of a substitution; this removes two substitutions compared to panel (a). In the bottom left panel (c), the majority constraint is lifted compared to panel (a), making q'' in bin 1 a candidate source for q' in bin 2. In the bottom right panel (d), the majority constraint is also lifted compared to panel (a) (adding the same $q'' \rightarrow q'$ substitution as in panel (c)), and the excluded-past constraint is added as in panel (b) (removing the two same substitutions from bin 2 to bin 3 as in panel (b)). If the bins were extended to the beginning of the quotation family, the excluded-past constraint would also remove the $q' \rightarrow q$ substitution from bin 2 to bin 3. In all four panels, a background rectangle or square indicates the quotation is the source of a substitution. A thick border on that rectangle or square indicates the quotation was selected because it has maximum frequency.

We identified four binary parameters that differentiate potential models, such that the resulting 16 combinations cover most of the reasonable answers to inference uncertainties. The first two parameters define the preceding time bin from which authors could have drawn a source when producing a new occurrence: (1) **bin positions**, which can be aligned to midnight (as in the model presented above) or kept sliding (for each occurrence, use a bin that ends precisely at that occurrence); (2) **bin span**, which can be $\Delta t = 1$ day (as in the model above) or can be extended up to the very first occurrence in the quotation family. The other two parameters configure rules on the selection of source and destination quotes of a substitution: (3) **candidate sources** can be restricted to the most frequent quotations in the preceding time bin (as in the model above), or not (in which case all quotations in the preceding bin are candidate sources); (4) **candidate destinations** can be restricted to quotations that do not appear in the preceding bin, or without restriction (as in the model above). A substitution model, then, is the given of a value for each of those parameters; it considers valid all the substitutions (and only those) where the source and destination follow the rules set out by the parameters. If a destination has substitutions from multiple sources we count a single effective substitution where, for each feature, the value for the effective source word is the average of the values of the candidate source words.

Put shortly a model defines how many times, and under what source and destination conditions, quote occurrences can be counted as substitutions. Fig. 2.4 shows the inferences made by the four models that use bins spanning 1 day aligned to midnight: later occurrences of q' and q'' are counted as substitutions in Fig. 2.4a and Fig. 2.4c, whereas in Fig. 2.4b and Fig. 2.4d they are not.

The results reported and discussed in the following sections are valid for all 16 models, and the graphics we present were produced by the model first introduced above. Finally, note that this inference proced-

ure is one of the reasons we restricted our analysis to single-substitutions: looking for more complex transformations would (a) exponentially increase the number of candidate sources for a destination occurrence, which correspondingly reduces the confidence in inferences made, and (b) greatly increase the complexity of the transformation models used to make these inferences.[13]

In practice for the model first introduced above, from the 2.60m initial occurrences spread over 50,427 quotation families, with significant redundancy (many quotes are indeed simple duplicates), we mine 40,868 substitutions. From these substitutions we remove those featuring stopwords, minor spelling changes (e.g. center/centre), abbreviations (e.g. November/Nov or Senator/Sen), spelled out numbers, words unknown to WordNet, and deletions in substrings (which can appear as substitutions of non-deleted words); this eventually yields 6,177 valid substitutions.[14]

2.4   Results

Substitutions usually replace a word with another semantically related word: manual observation of a random subset of 100 substitutions shows that, compared to the word it replaces, the new word often achieves a similar meaning in the context of its sentence while still slightly changing the implications or the attitude expressed by the author.[15] The following examples illustrate this phenomenon:

- "This is {socialism → welfare} for the rich,
- [The] "perverse logic of {clashes → confrontation} and violence,
- "This {crisis → problem} did not develop overnight and it will not be solved overnight.

Our question concerns the low-level properties of these substitutions: we ask (a) which words are targets of the substitutions and (b) what change these words are subjected to. To this end, we build the following two observables for each word feature. First, we measure which word features are more or less substituted compared to how often they would be if the process were random, in order to capture the susceptibility for words to be the target of a substitution in a quote. Second, we measure the change in word feature upon substitution, looking at the variation of a given feature between start and arrival words. Since sentence context is also central to this process, we extend these two observables by applying them to feature values relative to the distribution of feature values in the sentence in which a word appears.

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*, that is 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 we cannot know whether there has been perfect human reformulation or simply digital copy-pasting of a source ("Ctrl-C/Ctrl-V"). Moreover, perfect human reformulation possibly involves different practices than those involved in alteration — for instance drafting before publishing, double-checking sources, proofreading — 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.

---

[13]We checked that this restriction does not bias the results discussed below by extending our protocol to two-substitution transformations; the results were unchanged.

[14]Manually coding a random subset of 100 substitutions to evaluate this last filter showed that 84 were true negatives, 5 were false positives, and 11 true positives, giving a recall score of .688. Precision was evaluated over a random subset of 100 *kept* substitutions, showing a score of .87. Finally, note that excluding minor spelling changes does not bias our use of orthographic neighborhood density as a feature: out of the first 100 substitutions coded for recall, those with Levenshtein distance equal to 1 (which is what orthographic neighborhood density codes, Marian et al. 2012) were all typos or UK/US spelling changes, neither of which are relevant for this study.

[15]However, the substituted and substituting words are not so often direct synonyms: only a third of all substitutions travel less than 3 hops on the hyponym-hypernym network defined by WordNet (direct synonyms count as 0 hops on this network), meaning that at least two thirds involve non-synonyms. A similar phenomenon is observed on the FA network, where about 104 clusters have substitutions traveling only 1 hop, 110 traveling 2 hops, 137 traveling 3, 72 traveling 4, and 13 traveling 5.

2.4.1 Susceptibility

We say that a word is *substitutable* if it appears in a quote which undergoes a substitution, whether the substitution operates on that word or on another one. For a given group of words g, say all nouns, or all words in a particular range of values for a feature (e.g. words 2 to 4 letters long), susceptibility is computed as the ratio of s_g, the number of times words of that group are substituted, to s_g^0, the number of times words of that group would be substituted if substitutions fell randomly on substitutable words.[16] That is:

$$\sigma_g = \frac{s_g}{s_g^0}$$

In other words, susceptibility measures how much more or less a group of words g actually gets substituted compared to picking targets at random in quotes undergoing substitutions. By applying this measure to Part-of-Speech (POS) categories and feature bins (e.g. for a feature ϕ and a bin $[a; b]$, $g = \{w | \phi(w) \in [a; b]\}$), susceptibility measures the bias in the selection of start words involved in substitutions, i.e. it measures the preferential selection of some word properties for substitution.

Fig. 2.5 gathers the results for POS groups. A Goodman-based multinomial goodness-of-fit test (L. A. Goodman 1965) shows that these categories have a significant effect on susceptibility ($p < .05$ in all substitution models), but this seems mostly due to the *Closed class-like*[17] and *Adverb* categories. Indeed, detailing which categories are out of their confidence region under $\mathcal{H}_0$ shows that susceptibility for closed class-likes is significantly below 1 (confirmed in all substitution models), as is that for adverbs in 13 of the 16 substitution models; none of the other categories are significantly different from $\mathcal{H}_0$ (except nouns which appear significantly above 1 in a single substitution model). While we acknowledge the low susceptibilities of adverbs and closed class-likes, these categories concern less than 7% of all substitutions under $\mathcal{H}_0$ (and even less in the actual data); it seems, then, that POS categories do not capture any strong bias in the selection of substitution targets.

The results for word features presented in Fig. 2.6, on the other hand, show marked effects for several features. Word frequency, Age of acquisition, and Number of letters each exhibit significant susceptibility variations (Goodman goodness-of-fit with $p < .05$ in all substitution models, $p < .001$ in most) consistent with known effects of those features on recall. High-frequency words, much easier to recall, are substituted about half as much as they would be at random; conversely low-frequency words, harder to recall, are substituted about 50% more than random. Age of acquisition and Number of letters show the opposite pattern, consistent with their negative correlation to word frequency ($-.4$ and $-.19$): words learned before 5 or 6 years old, or made of less than 5 letters, are substituted less than random, whereas words learned after 10 years old, or made of more than 8 letters, are substituted far more than random. Orthographic neighborhood density also shows a slight effect (significant at $p < .05$ in 15 of the 16 substitution models): words with very sparse neighborhoods are more substituted than random (which may seem counter-intuitive, but is probably because over 70% of those words have 7 letters or more). Clustering

---

[16] s_g^0 is computed by summing, over all quotes undergoing a substitution, the ratio of the number of words in a quote that are from group g to the number of words in the same quote that could have been the substitution's target. If, for instance, half the content words of a given quote are nouns, such a quote contributes .5 to the total s_{nouns}^0. Further, 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 scale s_g and s_g^0 to count one for each cluster. That is, each quote cluster has a maximum contribution of 1, computed as the average contribution of all substitutions in that cluster.

[17] The *Closed class-like* category gathers all the POS groups representing closed class words (coordinating conjunctions, prepositions, subordinating conjunctions, modals and possessive endings). These groups, essentially made of stopwords, feature very low counts for both s (substitutions falling on stopwords are filtered out) and s^0 (stopwords are never counted as substitutable). While the susceptibility reported for the remaining words is left unbiased (as s and s^0 are equally affected), they represent a very small portion of all substitutions, which led us to group them together. Finally, we added to this meta-category the few POS groups that cover words entirely excluded from the analysis (foreign words, punctuation symbols and interjections), only sporadically present because of tagging fluctuations; hence the name *Closed class-like*.
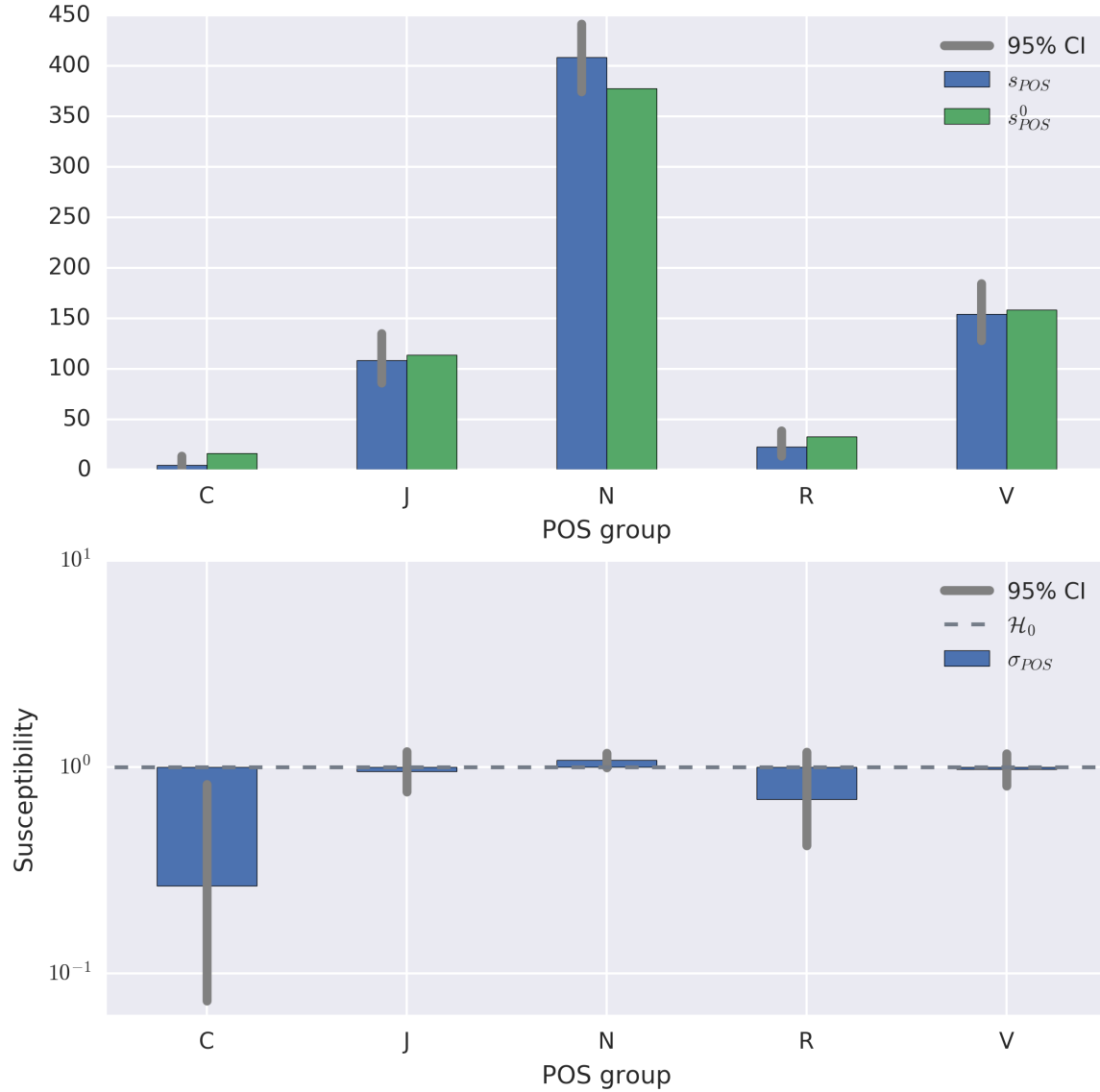
Figure 2.5: **POS-related results:** categories are simplified from the TreeTagger tag set: *C* means *Closed class-like* (see main text for details), *J* means adjective, *N* noun, *R* adverb, and *V* means verb. The top panel shows the actual s_{POS} and s_{POS}^0 counts. The bottom panel shows the substitution susceptibility σ_{POS}, which is the ratio between the two previous counts. Confidence intervals are computed with the L. A. Goodman (1965) method for multinomial proportions.
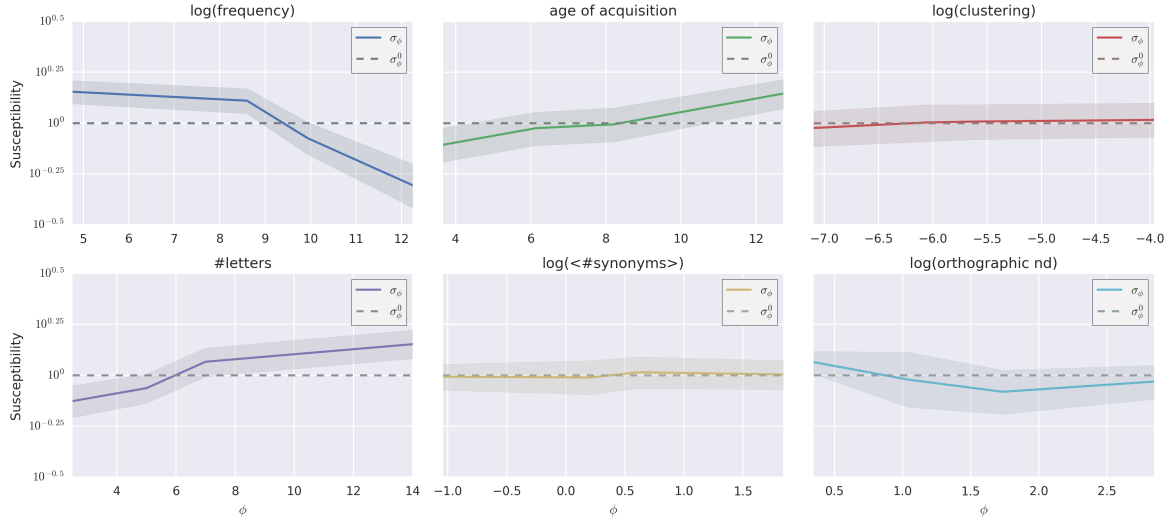
Figure 2.6: **Substitution susceptibility for feature values:** susceptibility to substitution versus feature value of a candidate word for substitution (binned by quartiles), with 95% asymptotic confidence intervals (Goodman-based multinomial).

coefficient shows no effect on susceptibility, and neither does Number of synonyms: in particular, words with many synonyms do not attract substitutions more than random (in fact, half the substitution models show they have a slight tendency to be substituted less than random).

On the whole, the trends observed are consistent with known effects of word frequency, age of acquisition, and number of letters, indicating that the triggering of a substitution could behave quite similarly to word recall in standard tasks.

2.4.2   Variation

We now examine how words are modified when they are substituted, that is how their features change upon substitution. Considering a word w substituted for w', we measure how a feature ϕ 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:[18]

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

We are interested in comparing the value of $\nu_\phi(f)$ to f itself, as this 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 attract words towards some typical feature value or not — a standard procedure in the study of dynamical systems.

We also introduce two null hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_{00}$, to compare the actual variation of a word's feature to expected variations under unbiased transformations. $\mathcal{H}_0$ models the situation where the arrival word w' is randomly chosen from the whole pool of words available in the data set for that feature.[19] In this

---

[18]Similarly to what we do for susceptibility, we avoid possible autocorrelation effects by averaging start- and arrival-word features of substitutions from the same cluster into a single aggregate substitution per cluster.

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

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:

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

$\mathcal{H}_{00}$ models the situation where the arrival word w' is chosen *among immediate synonyms of the start word* w, i.e. an arrival word chosen among semantically plausible though still random words. In this case ν_ϕ^{00} does depend on f:[20]

$$\nu_\phi^{00}(f) = \Big\langle \langle \phi(w') \rangle_{w' \in syn(w)} \Big\rangle_{\{w | \phi(w) = f\}}$$

This approach yields a fine-grained view of how word features evolve upon substitution, on average, with respect to (a) the original feature (vs. $y = x$), (b) a random arrival (vs. ν_ϕ^0), and (c) an unbiased semantically plausible arrival (vs. ν_ϕ^{00}).
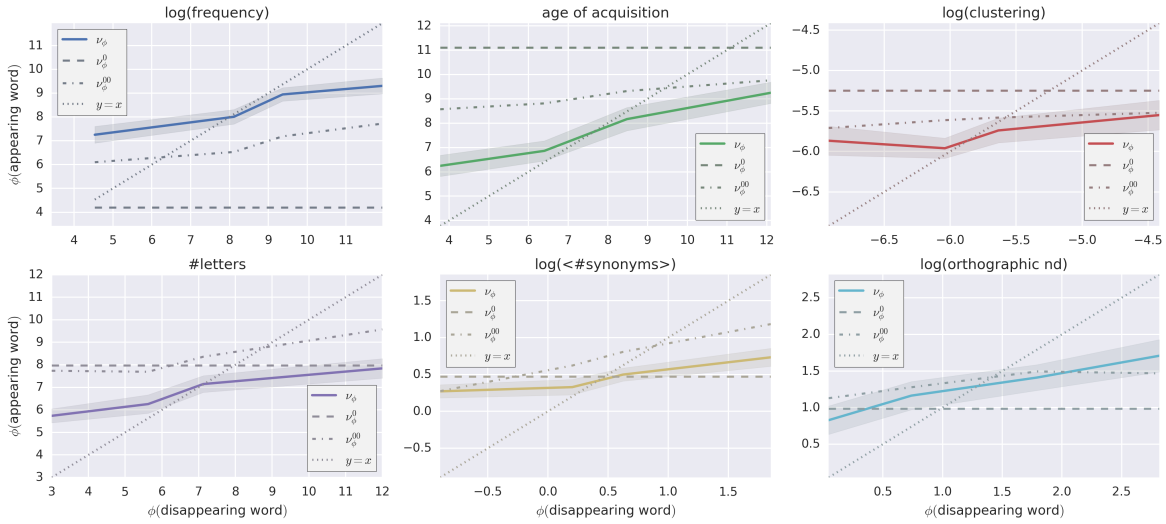
Figure 2.7: **Feature variation upon substitution:** ν_ϕ, average feature value of the appearing word as a function of the feature value of the disappearing word in a substitution (binned by quartiles), with 95% asymptotic confidence intervals based on Student's t-distribution. The overall position of the curve with respect to the dashed line representing $\mathcal{H}_0$ (constant ν_ϕ^0) indicates the direction of the cognitive bias compared to a purely random variation. The position with respect to the dash-dotted line representing $\mathcal{H}_{00}$ (ν_ϕ^{00}) indicates the bias compared to a semantically plausible random variation obtained by choosing a random synonym of the disappearing word. The intersection with $y = x$ marks the attractor value. The fact that all curves have slopes smaller than 1 in absolute value means that the substitution operation is contractile on average: it brings each feature closer to its own specific asymptotic range.

Results are gathered in Fig. 2.7. A first observation is that all graphs show the existence of a unique intersection of ν_ϕ with $y = x$, and the slope of ν_ϕ is smaller than 1 in absolute value, independently of the feature considered. This means that for each feature ϕ, whichever the value $\phi(w)$ of the disappearing word, the appearing word's feature value $\phi(w')$ will, on average, be closer to that feature's intersection of

---

[20]The actual implementation has an additional level of averaging since WordNet, used to get a word's synonyms, defines several meanings for a single given word, which we have no means of disambiguating. Therefore:

$$\nu_\phi^{00}(f) = \Bigg\langle \Big\langle \langle \phi(w') \rangle_{w' \in syn(m)} \Big\rangle_{m \in meanings(w)} \Bigg\rangle_{\{w | \phi(w) = f\}}$$

ν_ϕ with $y = x$.[21] In other words, beyond individual variation patterns, the substitution process exhibits a unique attractor for each feature. Note that this is also true under $\mathcal{H}_0$ or $\mathcal{H}_{00}$ (both null hypothesis curves have single intersections with $y = x$ with slopes smaller than 1): the substitution process naturally leads to an attraction even under reasonable random conditions.
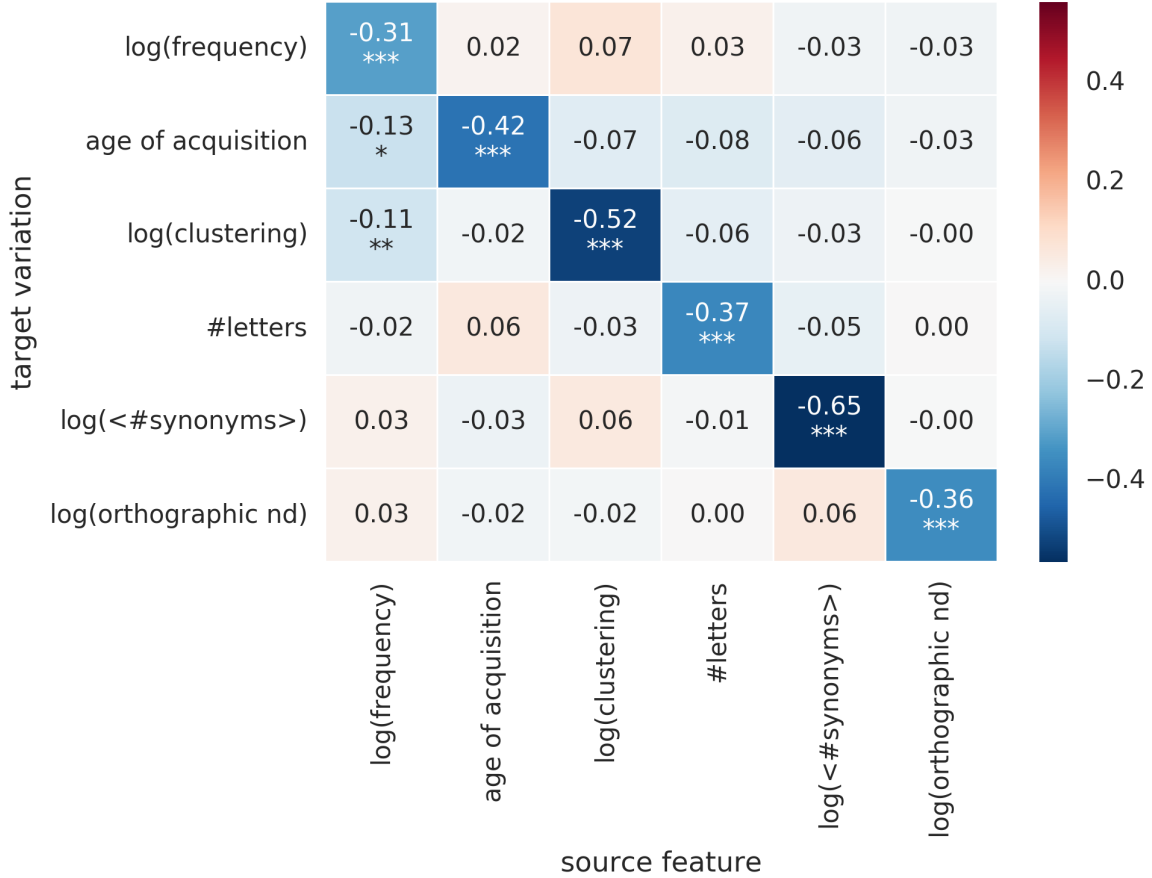

Figure 2.8: **Feature variations regression coefficients**: source feature values (columns) and feature variations (rows) were normalized to $[0; 1]$ to ensure the coefficients are comparable. Significance levels for individual t-tests against the hypothesis of a null coefficient are denoted by stars below the corresponding coefficient (*** for $p \leq .001$, ** for $p \leq .01$, * for $p \leq .05$, and nothing when $p > .05$). Frequency has a slight effect on Age of acquisition and Clustering coefficient, with small coefficients compared to the respective diagonal ones. Aside from those two, only diagonal values are significantly non null.

Second, the comparison with ν_ϕ^0 and ν_ϕ^{00} shows that there are two classes of attractors, depending on whether: * there is a triple intersection (of $y = x$, ν_ϕ, and ν_ϕ^0 or ν_ϕ^{00}); * or ν_ϕ always remains above or below ν_ϕ^0 and ν_ϕ^{00}.

The first class (Number of synonyms and Orthographic neighborhood density) are features for which

---

[21] This reasoning is standard in the analysis of dynamical systems (where the same transformation is applied to the whole system over and over), and becomes obvious when one manually simulates a substitution on the graph by picking a start value, using the ν_ϕ curve to obtain the corresponding arrival value, and comparing it to the start value: the arrival value is always closer to the intersection with $y = x$, meaning that that intersection is an attractor for the substitution process. If the slope of ν_ϕ were greater than one (in absolute value), the arrival value would always be *farther* from the intersection than the start value was, making the intersection with $y = x$ a *repulsor* point. This is how the number of intersections with $y = x$ and the slope of ν_ϕ at those intersections characterize the behavior of substitutions.

the substitution process only brings words slightly closer to ν_ϕ^0 (for Number of synonyms) or ν_ϕ^{00} (for Orthographic neighborhood density), and no uniform bias can be observed. The second class (comprising Word frequency, Age of acquisition, Clustering coefficient, and Number of letters) are features for which the substitution process has a clear bias, positive or negative, with respect to both the purely random situation ($\mathcal{H}_0$) and the semantically plausible random situation ($\mathcal{H}_{00}$).

Word frequency, with ν_ϕ always significantly above ν_ϕ^0 and ν_ϕ^{00}, 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. Age of acquisition, Clustering coefficient and Number of letters, on the other hand, exhibit a clear negative bias for the substitution process (except for high clustering values or very high number of letters). The three curves are significantly below their respective ν_ϕ^0 and ν_ϕ^{00} curves for most start values, which is consistent with the literature on recall: words learned earlier, with lower clustering coefficient or with fewer letters are easier to produce than average (Nelson et al. 2013; Zevin and Seidenberg 2002; Baddeley, Thomson, and Buchanan 1975). All these effects are significant with two-tailed t-tests at $p < .05$ (and more often $p < .001$) and were verified across the 16 substitution models.

To make sure our observations are not the product of correlations or interactions, we model the variations of the 6 features as a linear function of the start word's feature values:

$$\phi(w') - \phi(w) = \boldsymbol{A} + \boldsymbol{B} \cdot \phi(w)$$

where ϕ is the vector of all 6 features of a word, $\boldsymbol{A}$ is an intercept vector, and $\boldsymbol{B}$ is a 6×6 coefficients matrix. This regression achieves an overall R^2 of .33. The corresponding matrix of coefficients $\boldsymbol{B}$ is shown in Fig. 2.8: aside from Age of acquisition and Clustering coefficient on which word frequency has a slight effect, the variation of all other features depends solely on the disappearing word's same feature. In other words there is little to no interaction between a disappearing word's features in determining the variations that that word will undergo when substituted.

To make things concrete, here is an example substitution taking place in the data set. Around mid-November 2008 several media websites reported the following quote from Burmese poet Saw Wai (arrested for one of his poems),

> "Senior general Than Shwe is foolish with power.

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

> "Senior general Than Shwe is **crazy** with power.

The word *foolish* is acquired at an average of 8.94 years old, appears 675 times in the data set, has a Clustering coefficient of 8.2×10^{-3} and is 7 letters long. The word it was replaced with, *crazy*, is acquired on average at 5.22 years old, appears about 4.1k times in the data set, has a Clustering coefficient of 1.7×10^{-3}, and is 5 letters long. Such a change, though minor in appearance, is a typical example of alteration along the lines shown by our results.

2.4.3   Sentence context

The alterations we study are always made in a context, that is while the author is writing. We wish to ask, therefore, if taking that context into account can provide more insight into the substitution process. To do so we adapt the two observables presented above to capture some of the relationships between a word and the sentence it appears in.

Let us start with the first one: given a feature ϕ, we define the context-relative susceptibility to substitution with the following three steps. (1) For each quote in which a substitution appears, compute the distribution of ϕ values in that quote (excluding stopwords) and divide it into quartiles. (2) Count how many
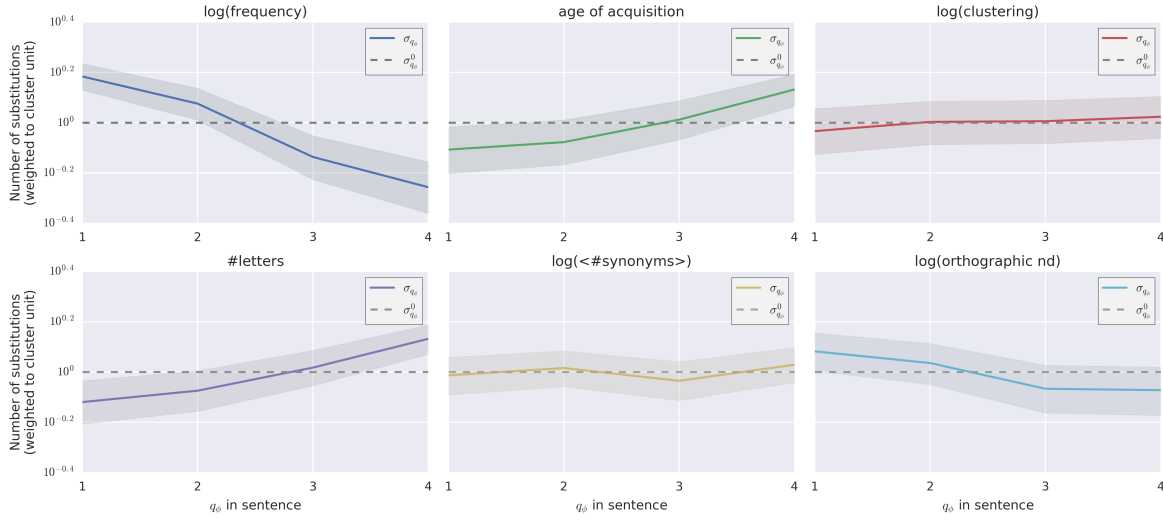
Figure 2.9: **Substitution susceptibility for in-quote feature quartiles:** susceptibility to substitution versus quartile of the feature distribution in the originating quote, with 95% asymptotic confidence intervals (Goodman-based multinomial).
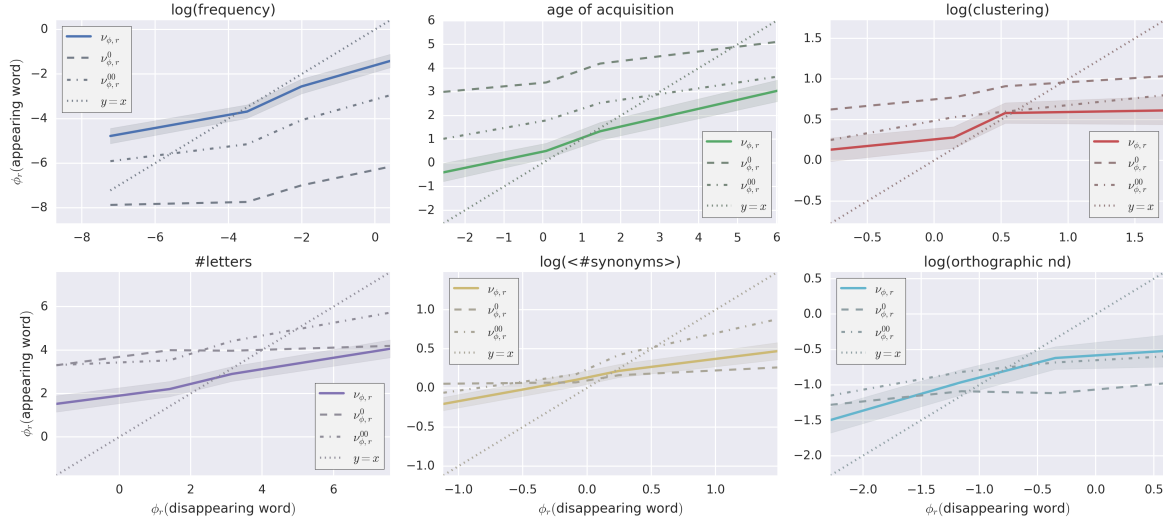

Figure 2.10: **Sentence-relative feature variation:** ν_{ϕ_r}, average sentence-relative feature value of the appearing word as a function of the sentence-relative value of the disappearing word (binned by quartiles), with 95% asymptotic confidence intervals based on Student's t-distribution. $\nu_{\phi_r}^0$ and $\nu_{\phi_r}^{00}$ are similarly converted to be sentence-relative. Attraction, magnitude and direction of bias with respect to null hypotheses are similar to Fig. 2.7. However, attractors are always positioned between sentence median ($y = 0$) on one side and $\nu_{\phi_r}^0$ and $\nu_{\phi_r}^{00}$ on the other side. Clustering coefficient, Number of synonyms and Orthographic neighborhood density are limit cases, with triple intersections with one of the null hypothesis curves.

times each quartile (first, second, third or fourth) contains a word that is substituted. This procedure tells us, for $i \in \{1; 2; 3; 4\}$, how many times substitutions fall in the i-th quartile of each in-quote distribution of ϕ; in other words it gives us the numerator s_{q_i} for the computation of susceptibility, where q_i represents the i-th quartile of the distributions of ϕ in the quotes. (3) Finally divide each quartile count by its corresponding $s_{q_i}^0$, that is the number of times the i-th quartile would receive substitutions if targets were picked at random; since the random situation would equally distribute a fourth of all substitutions to each quartile, we divide by the number of substitutions divided by 4. Taking for instance word frequency, this measure tells us if words that have high- or low-frequency *compared to the quote they appear in* are more or less substituted than at random.

Surprisingly the results for this measure are no different from the context-free measure, as can be seen in Fig. 2.9: low-frequency words compared to the sentence are substituted much more than higher-frequency words, words learned earlier than the rest of the sentence are substituted less than words learned later, shorter words less than longer words, and words with scarce neighborhoods slightly more than words with denser neighborhoods. Clustering coefficient and Number of synonyms are, here again and across all substitution models, not significantly different from $\mathcal{H}_0$: with or without context, they do not seem relevant to the selection of substitution targets.

Feature variation is more easily extended to the context-relative case. To do so we consider all feature values relative to the median word feature in the sentence. That is, in all the equations of the previous subsection we replace $\phi(w)$ with:

$$\phi_r(w) = \phi(w) - \text{median}\,\{\phi(w)|w \in \text{sentence}\}$$

ν_ϕ, ν_ϕ^0 and ν_ϕ^{00} each transpose to ν_{ϕ_r}, $\nu_{\phi_r}^0$ and $\nu_{\phi_r}^{00}$ (note that $\nu_{\phi_r}^0$ now depends on w since it is sentence-relative, whereas ν_ϕ^0 did not).

The results for sentence-relative feature variations are gathered in Fig. 2.10. Here too, the behavior is strikingly similar to the context-free measure: a single attractor is visible for each feature, and the magnitude and direction of biases are near-identical to those for the previous measure. The values of the appearing words give an additional insight into the process, however: the attractor value of a feature, at the intersection of ν_{ϕ_r} and $y = x$, is always between the sentence median, on one side, and $\nu_{\phi_r}^0$ and $\nu_{\phi_r}^{00}$ on the other side (for Number of synonyms it is a triple intersection with $\nu_{\phi_r}^0$; for Clustering coefficient and Orthographic neighborhood density, a triple intersection with $\nu_{\phi_r}^{00}$). Substitutions, therefore, seem to attract words closer to the sentence median than what a random process would do. This is true with respect to both null hypotheses (semantically plausible or not) for Frequency, Age of Acquisition and Number of letters, and true with respect to at least one of the two null hypotheses for the remaining features.

On the whole, we observe a clear attraction pattern for each feature, with two different classes corresponding to the psychological relevance of each feature for the substitution process. More awkward words along relevant features (less frequent, learned later, or made of more letters), both globally and with respect to the sentence they appear in, are substituted more often than what would happen if targets were picked randomly in the sentences; conversely, more common words are substituted less. Finally, across all features, substituted words are attracted towards a point closer to the sentence median than what a random process, semantically plausible or not, would do.

2.5 Discussion

We initially aimed to connect the field of cultural evolution with psycholinguistics by asking if cultural attractors appear in a corpus of online news-related quotes gradually transformed by low-level biases. The data set we used imposed a few constraints on our analysis: first, it was necessary to infer source-destination links, an operation made more reliable when restricting the scope of transformations to very

simple cases, which we did by focusing on single word substitutions. Second, contrary to laboratory experiments which produce data made of many repeated measures on a small number of cases (e.g. a given list of words), we have a great number of different cases (one case per cluster in which substitutions are found, i.e. 698 cases), with very few measures on each of them (average 9, median 5). This rendered the prediction of individual words impractical: if we cannot compute a percentage of explained data for a given case, any approximate prediction will be heavily underestimated. This last factor, added to the potential for variation of external conditions when authors wrote the quotes, led us to use word features to analyze the transformations by aggregating over individual cases.

By characterizing substitutions with 6 features on the disappearing word, we show that authors preferentially substitute words known for being harder to recall: most prominently words with low frequency (Gregg 1976), learned later (Dewhurst, Hitch, and Barry 1998), or made up of more letters (Nickels and Howard 2004), both globally and in comparison to the sentence they appear in. Further characterizing the substitutions by examining the variation of word features from disappearing to appearing words, we show: (a) that the operation is contractile on average, that is words are brought closer to an attractor point on each feature; (b) that authors produce words that are easier to remember than the average of synonyms of the disappearing word (a fact that is reflected in the position of the attraction point).

We do not actually observe quotes converging on a global scale towards attractors in their various dimensions. Indeed the limits of the data set do not allow us to infer chains of substitutions, and substitutions themselves are not the only type of transformation at work in the data set. Nonetheless, these findings (a) bring light to this simple type of transformation, and (b) are consistent with known psycholinguistic effects, with the hypothesis of cultural attractors in representations from everyday life, and with the lineage specificity discussed in the iterated learning literature (Claidière et al. 2014; Cornish, Smith, and Kirby 2013). They are obtained by successfully applying knowledge from cognitive science to real-life complex data, a task that remains a challenge in the study of cultural evolution. More broadly, we believe that applying such data mining tools to manage the complexity of real-life data is a promising approach for the joint analysis of cognitive science and culture.

In the simple case presented here, however, much remains to be explored. Since it is clear that observing cognitive biases in such data is now possible, questions addressed in controlled laboratory situations could be opened by further research. One question concerns the influence of the context surrounding a quote, be it in terms of other quotes preceding it temporally or of text surrounding it in a post. A first step could be the application of results from Zaromb et al. (2006) who have shown, in the simpler task of recall of random word lists, that the source of prior-list intrusions can be predicted based on the associations those preceding lists have formed: in our case, a substitution could be triggered and directed by associations formed by preceding context. A further step would be to follow what Cornish, Smith, and Kirby (2013) have shown about reciprocal influences between context and transformations (in their case, with transmission chains of artificial content). Indeed substitutions, and more generally all transformations, also participate in creating the context for later quotes. One can ask, therefore, what are the reciprocal effects between, on one side, the corpus-level evolution of quotes through iterated transformations, and on the other side, a gradual change in the properties of transformations operated because of the evolution of surrounding context. Such interactions have been shown to underlie the lineage specificity observed in transmission chains (Claidière et al. 2014). Exploring how similar loop interactions happen in real-life data could indeed be the next step in understanding the coevolution of cultural content and the ways in which it is transformed. In our particular case, such insight could shed some light on how the feature attractors examined in this paper actually emerge, and help assess their potential role on this coevolution.

2.6   Concluding remarks

The theory of Epidemiology of Representations proposes a unifying framework for the study of cultural evolution. One of its core claims, the existence of cultural attractors, has been both a challenge to test empirically and a fruitful line to pursue in the study of cultural evolution. We aimed to contribute to testing this hypothesis by studying a simple everyday-life task where individuals are implicitly trying to reproduce quotations. To some extent, our work amounts to an out-of-laboratory experiment where we examine the influence of well-known word features on the accuracy of reproduction of short sentences. Our analysis of substitutions shows that words are attracted, in each dimension, to feature-specific values. Furthermore, the features' known effects in psycholinguistic experiments are reflected in the biases of these attraction points, meaning that the evolution of such quotations can be partially explained by known low-level cognitive biases. We believe that such an approach, which combines psycholinguistic knowledge and data mining tools, can be fruitfully developed to improve the study of cultural attractors and explore the reciprocal influences of cognition and culture.

Let us conclude by noting that the question of short- and long-term cultural evolution, and the approaches to study them, are becoming increasingly relevant to other fields. In biology in particular, work on evo-devo and non-genetic inheritance has accumulated evidence that is poorly accounted for by the modern synthesis of biological evolution, and is creating a demand for new or extended approaches to joint cultural and biological evolution (see Gilbert, Bosch, and Ledón-Rettig 2015 for instance). Such an approach has long been called upon by anthropologists like Ingold (2004; 1998), in line with Mauss' initial works (Mauss 1936), and the question is not entirely foreign to the enactive-representational debate in cognitive science. The study of cultural evolution will most likely benefit greatly from the growing interactions between these disciplines.

Acknowledgements

Software colophon

Finally, this paper was developed using Python's scientific computing ecosystem (Millman and Aivazis 2011). In particular, we directly used NumPy and SciPy (van der Walt, Colbert, and Varoquaux 2011), Matplotlib (Hunter 2007), Pandas (McKinney 2010), scikit-learn (Pedregosa et al. 2011), NetworkX (Hagberg, Schult, and Swart 2008), NLTK (Bird, Klein, and Loper 2009), IPython (Pérez and Granger 2007), and many other libraries from the Python ecosystem. The software and analyses written for the paper are documented and published under a Free Software license. They can be found at github.com/wehlutyk/brainscopypaste.

Chapter 3

Gistr

Chapter 4

Discussion

Chapter 5

Conclusion

References

Acerbi, Alberto, and Alex Mesoudi. 2015. 'If We Are All Cultural Darwinians What's the Fuss About? Clarifying Recent Disagreements in the Field of Cultural Evolution'. *Biology & Philosophy* 30 (4): 481–503. doi:10.1007/s10539-015-9490-2.

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

Austerweil, Joseph L., Joshua T Abbott, and Thomas L. Griffiths. 2012. 'Human Memory Search as a Random Walk in a Semantic Network'. In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 3041–9. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/4761-human-memory-search-as-a-random-walk-in-a-semantic-network.pdf.

Baddeley, Alan D., Neil Thomson, and Mary Buchanan. 1975. 'Word Length and the Structure of Short-Term Memory'. *Journal of Verbal Learning and Verbal Behavior* 14 (6): 575–89. doi:10.1016/S0022-5371(75)80045-4.

Baumard, Nicolas, Alexandre Hyafil, Ian Morris, and Pascal Boyer. 2015. 'Increased Affluence Explains the Emergence of Ascetic Wisdoms and Moralizing Religions'. *Current Biology* 25 (1): 10–15. doi:10.1016/j.cub.2014.10.063.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media, Inc.

Bloch, Maurice. 2000. 'A Well-Disposed Social Anthropologist's Problems with Memes'. In *Darwinizing Culture: The Status of Memetics as a Science*, edited by Robert Aunger, 189–204. Oxford, NY: Oxford University Press.

Bourdieu, Pierre. 1980. *Le sens pratique*. Paris: Editions de Minuit.

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

———. 2005. *The Origin and Evolution of Cultures*. New York, N.Y. ; Oxford: Oxford university press.

Cavalli-Sforza, L. L, and Marcus W Feldman. 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, N.J.: Princeton University Press.

Chan, Kit Ying, and Michael S. Vitevitch. 2010. 'Network Structure Influences Speech Production'. *Cognitive Science* 34 (4): 685–97. doi:10.1111/j.1551-6709.2010.01100.x.

Claidière, Nicolas, and Dan Sperber. 2007. 'The Role of Attraction in Cultural Evolution'. *Journal of Cognition and Culture* 7 (1): 89–111. doi:10.1163/156853707X171829.

Claidière, Nicolas, Thomas C. Scott-Phillips, and Dan Sperber. 2014. 'How Darwinian Is Cultural Evolution?' *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1642): 20130368.

doi:10.1098/rstb.2013.0368.

Claidière, Nicolas, Kenny Smith, Simon Kirby, and Joël Fagot. 2014. 'Cultural Evolution of Systematic-ally Structured Behaviour in a Non-Human Primate'. *Proceedings of the Royal Society of London B: Biological Sciences* 281 (1797): 20141541. doi:10.1098/rspb.2014.1541.

Claidière, Nicolas, Emmanuel Trouche, and Hugo Mercier. 2017. 'Argumentation and the Diffu-sion of Counter-Intuitive Beliefs'. *Manuscript Submitted for Publication*. https://sites.google.com/site/hugomercier/Argumentation%20and%20counter-intuitive%20beliefs.pdf.

Cointet, J. P., and C. Roth. 2009. 'Socio-Semantic Dynamics in a Blog Network'. In *International Conference on Computational Science and Engineering, 2009. CSE '09*, edited by Alex Pentland, Justin Zahn, and Daniel Zeng, 4:114–21. Washington, DC: IEEE Computer Society. doi:10.1109/CSE.2009.105.

Cornish, Hannah, Kenny Smith, and Simon Kirby. 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*, edited by Markus Knauff, Michael Pauen, Natalie Sebanz, and Ipke Wachsmuth. Austin, TX: Cognitive Science Society.

Danescu-Niculescu-Mizil, Cristian, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. 'You Had Me at Hello: How Phrasing Affects Memorability'. In *ACL '12 Proceedings of the 50th Annual Meeting of the Asso-ciation for Computational Linguistics: Long Papers*, edited by Haizhou Li, Chin-Yew Lin, and Miles Osborne, 1:892–901. Stroudsburg, PA: ACM. http://arxiv.org/abs/1203.6360.

Dawkins, Richard. (1976) 2006. *The Selfish Gene*. Oxford; New York: Oxford University Press.

Deese, James. 1959. 'On the Prediction of Occurrence of Particular Verbal Intrusions in Immediate Recall'. *Journal of Experimental Psychology* 58 (1): 17–22. doi:10.1037/h0046671.

Dewhurst, Stephen A., Graham J. Hitch, and Christopher Barry. 1998. 'Separate Effects of Word Frequency and Age of Acquisition in Recognition and Recall'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24 (2): 284–98. doi:10.1037/0278-7393.24.2.284.

Durkheim, Emile. (1897) 2012. *Le Suicide: Étude de Sociologie*. Project Gutenberg. http://www.gutenberg.org/ebooks/40489.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press.

Freeman, Linton C. 1977. 'A Set of Measures of Centrality Based on Betweenness'. *Sociometry* 40 (1): 35–41. doi:10.2307/3033543.

Garlock, Victoria M., Amanda C. Walley, and Jamie L. Metsala. 2001. 'Age-of-Acquisition, Word Fre-quency, and Neighborhood Density Effects on Spoken Word Recognition by Children and Adults'. *Journal of Memory and Language* 45 (3): 468–92. doi:10.1006/jmla.2000.2784.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge, UK; Oxford, UK: Polity Press & Blackwell.

Gilbert, Scott F., Thomas C. G. Bosch, and Cristina Ledón-Rettig. 2015. 'Eco-Evo-Devo: Developmental Symbiosis and Developmental Plasticity as Evolutionary Agents'. *Nature Reviews Genetics* 16 (10): 611–22. doi:10.1038/nrg3982.

Goodman, Leo A. 1965. 'On Simultaneous Confidence Intervals for Multinomial Proportions'. *Technomet-rics* 7 (2): 247–54. doi:10.1080/00401706.1965.10490252.

Gregg, Vernon. 1976. 'Word Frequency, Recognition and Recall'. In *Recall and Recognition*, edited by John Brown, 183–216. Oxford, UK: John Wiley & Sons.

Griffiths, Thomas L., Mark Steyvers, and Alana Firl. 2007. 'Google and the Mind: Predicting Fluency with

PageRank'. *Psychological Science* 18 (12): 1069–76. doi:10.1111/j.1467-9280.2007.02027.x.

Gruhl, Daniel, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. 'Information Diffusion Through Blogspace'. In *Proceedings of the 13th International Conference on World Wide Web*, edited by Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, 491–501. New York, NY: ACM. doi:10.1145/988672.988739.

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart. 2008. 'Exploring Network Structure, Dynamics, and Function Using NetworkX'. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, edited by Gaël Varoquaux, Travis Vaught, and Jarrod Millman, 11–15. Pasadena, CA.

Howard, Marc W, and Michael J Kahana. 2002. 'When Does Semantic Similarity Help Episodic Retrieval?' *Journal of Memory and Language* 46 (1): 85–98. doi:10.1006/jmla.2001.2798.

Hunter, John D. 2007. 'Matplotlib: A 2d Graphics Environment'. *Computing in Science & Engineering* 9 (3): 90–95. doi:10.1109/MCSE.2007.55.

Ingold, Tim. 1998. 'From Complementarity to Obviation: On Dissolving the Boundaries Between Social and Biological Anthropology, Archaeology and Psychology'. *Zeitschrift Für Ethnologie* 123 (1): 21–52.

———. 2004. 'Beyond Biology and Culture. the Meaning of Evolution in a Relational World'. *Social Anthropology* 12 (2): 209–21. doi:10.1111/j.1469-8676.2004.tb00102.x.

———. 2007. 'The Trouble with "Evolutionary Biology"'. *Anthropology Today* 23 (2): 13–17. doi:10.1111/j.1467-8322.2007.00497.x.

Jefferies, Elizabeth, Matthew A. Lambon Ralph, and Alan D. Baddeley. 2004. 'Automatic and Controlled Processing in Sentence Recall: The Role of Long-Term and Working Memory'. *Journal of Memory and Language* 51 (4): 623–43. doi:10.1016/j.jml.2004.07.005.

Kirby, Simon, Hannah Cornish, and Kenny Smith. 2008. '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): 10681–6. doi:10.1073/pnas.0707835105.

Kroeber, A. L. 1952. *The Nature of Culture*. Chicago, IL: University of Chicago Press.

Kuper, Adam. 2000. 'If Memes Are the Answer, What Is the Question?' In *Darwinizing Culture: The Status of Memetics as a Science*, edited by Robert Aunger, 175–88. Oxford, NY: Oxford University Press.

Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. 'Age-of-Acquisition Ratings for 30,000 English Words'. *Behavior Research Methods* 44 (4): 978–90. doi:10.3758/s13428-012-0210-4.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 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*, edited by John Elder, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed J. Zaki, 497–506. New York, NY: ACM. doi:10.1145/1557019.1557077.

Lewontin, R. C. 1982. 'Cultural Transmission and Evolution: A Quantitative Approach'. *American Journal of Human Genetics* 34 (5): 831–32. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1685425/.

Liben-Nowell, David, and Jon Kleinberg. 2008. 'Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data'. *Proceedings of the National Academy of Sciences* 105 (12): 4633–8. doi:10.1073/pnas.0708471105.

Lombardi, Linda, and Mary C Potter. 1992. 'The Regeneration of Syntax in Short Term Memory'. *Journal of Memory and Language* 31 (6): 713–33. doi:10.1016/0749-596X(92)90036-W.

MacCallum, Robert M., Matthias Mauch, Austin Burt, and Armand M. Leroi. 2012. 'Evolution of Music by Public Choice'. *Proceedings of the National Academy of Sciences* 109 (30): 12081–6.

doi:10.1073/pnas.1203182109.

Mandler, George, George O. Goodman, and Deanna L. Wilkes-Gibbs. 1982. 'The Word-Frequency Paradox in Recognition'. *Memory & Cognition* 10 (1): 33–42. doi:10.3758/BF03197623.

Marian, Viorica, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. 'CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities'. *PLOS ONE* 7 (8): e43230. doi:10.1371/journal.pone.0043230.

Mauss, Marcel. 1936. 'Les Techniques Du Corps'. *Journal de Psychologie* 32 (3-4): 271–93. http://classiques. uqac.ca/classiques/mauss_marcel/socio_et_anthropo/6_Techniques_corps/Techniques_corps.html.

McKinney, Wes. 2010. 'Data Structures for Statistical Computing in Python'. In *Proceedings of the 9th Python in Science Conference (SciPy 2009)*, edited by Stéfan van der Walt and Jarrod Millman, 445:51–56. Pasadena, CA. http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Mesoudi, Alex, and Andrew Whiten. 2008. '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. doi:10.1098/rstb.2008.0129.

Mesoudi, Alex, Andrew Whiten, and Kevin N. Laland. 2006. 'Towards a Unified Science of Cultural Evolution'. *Behavioral and Brain Sciences* 29 (04): 329–47. doi:10.1017/S0140525X06009083.

———. 2007. 'Science, Evolution, and Cultural Anthropology. A Response to Ingold (This Issue)'. *Anthropology Today* 23 (2): 18–18. doi:10.1111/j.1467-8322.2007.00498.x.

Millman, K. Jarrod, and Michael Aivazis. 2011. 'Python for Scientists and Engineers'. *Computing in Science & Engineering* 13 (2): 9–12. doi:10.1109/MCSE.2011.36.

Miton, Helena, Nicolas Claidière, and Hugo Mercier. 2015. 'Universal Cognitive Mechanisms Explain the Cultural Success of Bloodletting'. *Evolution and Human Behavior* 36 (4): 303–12. doi:10.1016/j.evolhumbehav.2015.01.003.

Morin, Olivier. 2013. 'How Portraits Turned Their Eyes Upon Us: Visual Preferences and Demographic Change in Cultural Evolution'. *Evolution and Human Behavior* 34 (3): 222–29. doi:10.1016/j.evolhumbehav.2013.01.004.

Morrison, Catriona M., and Andrew W. Ellis. 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): 116–33. doi:10.1037/0278-7393.21.1.116.

Moussaïd, Mehdi, Henry Brighton, and Wolfgang Gaissmaier. 2015. 'The Amplification of Risk in Experimental Diffusion Chains'. *Proceedings of the National Academy of Sciences* 112 (18): 5631–6. doi:10.1073/pnas.1421883112.

Nelson, Douglas L., Kirsty Kitto, David Galea, Cathy L. McEvoy, and Peter D. Bruza. 2013. 'How Activation, Entanglement, and Searching a Semantic Network Contribute to Event Memory'. *Memory & Cognition* 41 (6): 797–819. doi:10.3758/s13421-013-0312-y.

Nelson, Douglas L., Cathy L. McEvoy, and Thomas A. Schreiber. 2004. 'The University of South Florida Free Association, Rhyme, and Word Fragment Norms'. *Behavior Research Methods, Instruments, & Computers* 36 (3): 402–7. doi:10.3758/BF03195588.

Nickels, Lyndsey, and David Howard. 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. doi:10.1080/02643290342000122.

Odling-Smee, F. John, Kevin N Laland, and Marcus W Feldman. 2003. *Niche Construction: The Neglected Process in Evolution*. Princeton: Princeton University Press.

Omodei, Elisa, Thierry Poibeau, and Jean-Philippe Cointet. 2012. 'Multi-Level Modeling of Quotation Families Morphogenesis'. In *Proceedings of the 4th ASE/IEEE International Conference on Social Computing,*

edited by Anton Nijholt, Alessandro Vinciarelli, and Dirk Heylen, 392–401. Washington, DC: IEEE Computer Society. http://arxiv.org/abs/1209.4277.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. 'The PageRank Citation Ranking: Bringing Order to the Web'. Stanford InfoLab.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. 'Scikit-Learn: Machine Learning in Python'. *Journal of Machine Learning Research* 12 (Oct): 2825–30. http://www.jmlr.org/papers/v12/pedregosa11a.html.

Pérez, Fernando, and Brian E. Granger. 2007. 'IPython: A System for Interactive Scientific Computing'. *Computing in Science & Engineering* 9 (3): 21–29. doi:10.1109/MCSE.2007.53.

Potter, Mary C, and Linda Lombardi. 1990. 'Regeneration in the Short-Term Recall of Sentences'. *Journal of Memory and Language* 29 (6): 633–54. doi:10.1016/0749-596X(90)90042-X.

Potter, Mary C., and Linda Lombardi. 1998. 'Syntactic Priming in Immediate Recall of Sentences'. *Journal of Memory and Language* 38 (3): 265–82. doi:10.1006/jmla.1997.2546.

Rey, Arnaud, Arthur M Jacobs, Florian Schmidt-Weigand, and Johannes C Ziegler. 1998. 'A Phoneme Effect in Visual Word Recognition'. *Cognition* 68 (3): B71–B80. doi:10.1016/S0010-0277(98)00051-1.

Roediger, Henry L., and Kathleen B. McDermott. 1995. 'Creating False Memories: Remembering Words Not Presented in Lists'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (4): 803–14. doi:10.1037/0278-7393.21.4.803.

Schmid, Helmut. 1994. 'Probabilistic Part-of-Speech Tagging Using Decision Trees'. In *Proceedings of International Conference on New Methods in Language Processing*, edited by Daniel B. Jones, 12:44–49. Manchester, UK: UMIST.

Simmons, Matthew P., Lada A. Adamic, and Eytan Adar. 2011. 'Memes Online: Extracted, Subtracted, Injected, and Recollected'. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, edited by Nicolas Nicolov, James G. Shanahan, Lada A. Adamic, Ricardo Baeza-Yates, and Scott Counts, 353–60. Menlo Park, CA: The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2836.

Slingerland, Edward G. 2008. *What Science Offers the Humanities: Integrating Body and Culture*. Cambridge; New York: Cambridge University Press.

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach*. Oxford, UK; Cambridge, Mass.: Blackwell.

Sterelny, Kim. 2017. 'Cultural Evolution in California and Paris'. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 62 (April): 42–50. doi:10.1016/j.shpsc.2016.12.005.

Tamariz, Monica, and Simon Kirby. 2016. 'The Cultural Evolution of Language'. *Current Opinion in Psychology* 8 (April): 37–43. doi:10.1016/j.copsyc.2015.09.003.

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Tulving, Endel. 1962. 'Subjective Organization in Free Recall of "Unrelated" Words'. *Psychological Review* 69 (4): 344–54. doi:10.1037/h0043150.

———. 1966. 'Subjective Organization and Effects of Repetition in Multi-Trial Free-Recall Learning'. *Journal of Verbal Learning and Verbal Behavior* 5 (2): 193–97. doi:10.1016/S0022-5371(66)80016-6.

van der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux. 2011. 'The NumPy Array: A Structure for Effi-

cient Numerical Computation'. *Computing in Science & Engineering* 13 (2): 22–30. doi:10.1109/MCSE.2011.37.

Watts, Duncan J., and Steven H. Strogatz. 1998. 'Collective Dynamics of "Small-World" Networks'. *Nature* 393 (6684): 440–42. doi:10.1038/30918.

Weide, Robert. 1998. 'The CMU Pronouncing Dictionary'. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

WordNet. 2010. 'Princeton University "About WordNet"'. https://wordnet.princeton.edu/wordnet/.

Yonelinas, Andrew P. 2002. 'The Nature of Recollection and Familiarity: A Review of 30 Years of Research'. *Journal of Memory and Language* 46 (3): 441–517. doi:10.1006/jmla.2002.2864.

Zaromb, Franklin M., Marc W. Howard, Emily D. Dolan, Yevgeniy B. Sirotin, Michele Tully, Arthur Wingfield, and Michael J. Kahana. 2006. 'Temporal Associations and Prior-List Intrusions in Free Recall'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32 (4): 792–804. doi:10.1037/0278-7393.32.4.792.

Zevin, Jason D, and Mark S Seidenberg. 2002. 'Age of Acquisition Effects in Word Reading and Other Tasks'. *Journal of Memory and Language* 47 (1): 1–29. doi:10.1006/jmla.2001.2834.