PhD Dissertation

Epidemiology of Representations: An Empirical Approach

Sébastien Lerique[1]

Supervisor: Jean-Pierre Nadal[2]
Co-supervisor: 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

General introduction

Chapter 1

Introduction

TODO: This might be too high-level, so make it more focused. Also add an early hint of what I'll do, which is not that broad. Remember to start easy (but rich) before going to more complex things.

Current scientific knowledge describes the 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 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, along 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 SCE 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". (This problem, the cog-soc link, is also sometimes viewed as the question of what human nature is, because an answer to that question would define the unit of analysis for culture and everything else.)
- 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 SCE
- 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 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 (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), in his studies of the ways in which people of different societies and throughout history use their bodies differently, represents one approach to that question. 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* developed by 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 structure, or referring to it, is what makes it action and not 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: steiner-autonomy-2009, who connects Giddens (and others) with the questions of social cognition

Anthropology and sociology have produced considerable amounts of work concerned with the nature of culture in relation to psychology, the most prominent parts of which are reviewed by Risjord (2012). The sample I exposed here represents the approaches which have most influenced the initial questions of the present 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 is worth keeping in mind when studying the cultural evolution approach that I will focus on in what follows.

While current writings often argue that such a separation is not "sharp" (Acerbi 2016, 2; relying on Morin 2016), the dualism it comes from permeates most of our theories about non-human animals, human beings, and life in general. This thesis is no exception, as I am myself an apprentice of the traditions I present and discuss here. However I see no reason to believe it is the best conceptual dicing one could achieve, and am (for now) agnostic about whether or not it should be maintained; whenever such distinctions appear in what follows, for instance between cultural and biological evolution, I will thus be referring to the conception of the works discussed, and will attempt to clarify when otherwise. I will come back to the consequences of this dualism in Section 1.1.5, and

will propose a more detailed discussion in Chapter 4.

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 generality of the populational approach underlying contemporary evolutionary theory, a new wave of analyses developed throughout the eighties and nineties by proposing a blend of (a) the gene-centred level of evolutionary theory, (b) insights that the application of generalised evolutionary principles may, or may not, 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 through psychology, learning, and environmental engineering. The combination of (a) and (b) is not new, as a similar inspiration underlay the development of memetics. Tim Lewens clarifies the position of this new wave regarding previous evolutionary approaches to culture:

> "Cultural evolutionists frequently begin their theorizing from a starting point that does not invite use of the meme concept, and may therefore appear neutral regarding its propriety. Their project is to integrate various forms of learning into evolutionary theory, in a way that leaves open the degree to which learning has anything in common with genetic inheritance." (Lewens 2012, 466)

(c), which differentiates current cultural evolutionists from previous approaches, has been developed into two parallel approaches, which Sterelny (2017) terms the Californian and the Parisian traditions.

ADD: scott-phillips-simple-2017, mesoudi-cultural-2016

Californian

Cavalli-Sforza and Feldman (1981) first articulated the combination of these three ideas in detail, by building on the fact that an evolutionary process need not be mediated by genetic transmission to take place. Indeed, following the synthesis provided by Lewontin (1982): for any type of item, the combination of (a) a transmission process leading to nonrandom dependence of subsequent instances on preceding instances (e.g. offspring and parent phenotype), (b) at least some variation of the properties of items (against which to select), and (c) some level of differential survival and spread depending on those properties, 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), horizontal (inside one generation), or any combination thereof. Boyd and Richerson (1985; 2005) further developed this line, leading to the formulation of *Standard Cultural Evolution* (hereafter SCE) which offers a systematic analysis of part of the 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), 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 1.1.4.

ADD: mesoudi-cultural-2011?

Parisian

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*, hereafter noted 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 of 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, while at the same time connecting with the many ontologies he identifies in anthropology.

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 the theory Sperber proposes, in contrast to not only memetics, but also to SCE as 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 notion of mental representations from 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 of representations which models a society's culture 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 central goal in the CAT research stream has been to identify existing attractors, and explain their emergence based on the interaction of psychological and ecological factors.

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 SCE (Sterelny 2017, 47); CAT is nonetheless compatible with the framework of gene-culture co-evolution developed by SCE (Sperber 1996, 114), and most authors consider that both theories are compatible but focus on slightly different questions (Sterelny 2017; Acerbi and Mesoudi 2015).

The space opened by the development of SCE and CAT has generated much debate (see e.g. the peer commentary to Mesoudi, Whiten, and Laland 2006) and some heated criticisms (Ingold 2007; answered in Mesoudi, Whiten, and Laland 2007; Lewens 2012, 461–2 provides further discussion and context). Sections 1.1.4 and 1.1.5 come back to these debates and criticisms, which I group into two broad categories. The first concerns the specifics of the evolutionary approach adopted by SCE 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 "informational dualism" of the notion of representation in cognitive science: at the concrete level, it questions whether a representationalist approach can naturalise meaning, a core aspect in the analysis of culture and life; at the programmatic level it suggests that, compared to the focus on dynamic couplings that Thompson and his colleagues develop under the "enactive approach" banner, 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 SCE 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 single study, or collection of studies for that matter, will reach a yes or 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 or absence of a well-defined attractor in a given situation can be turned into a testable hypothesis. In addition at the meta-analytical level, one can readily evaluate the usefulness of a cultural attraction approach in concrete domains, by analysing whether or not the questions it encourages result in fruitful studies, on average, for that domain. So while testing CAT empirically is not a simple matter of yes or no, the theory provides a clear set of tools and lines of thought that are well worth evaluating in a wider, programmatic appraisal.

The second challenge to empirical exploration resides in the fact that quantitative data on out-of-laboratory cultural artefacts is not easy to collect. 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. The literature on CAT shows at least two important methods that have been used up to now. First, the meta-analysis of large numbers of anthropological or historical works, which has uncovered several relevant effects. The way portraits are painted over the centuries, for instance, has been show to increasingly favour direct- versus oblique-gaze portraying (Morin 2013). Miton, Claidière, and Mercier (2015) also used this technique to propose an explanation for the historical stability of the bloodletting practice, in spite of its medical ineffectiveness. Exploiting similar records, Baumard et al. (2015) showed a link between the evolution of religious values and the affluence of societies in which they develop. Morin and Miton (EHBEA 2017? [Citation needed]), using the same approach, 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 is to reproduce the evolution of content with human or animal participants in the laboratory, where subjects repeatedly transmit or interact around pieces of content initially chosen by the experimenters. The resulting data is an approximation of the rapid evolution of artefacts in real life, yet in a controlled situation that can then be statistically analysed with access to all the parameters. This technique has been used in studies ranging from the evolution of visual patterns transmitted in a group of apes (Claidière et al. 2014), to the role of argumentation in the transmission of solutions to simple but counter-intuitive reasoning problems (Claidière, Trouche, and Mercier 2017). # ADD: baumard-mutualistic-2013. A number of other disciplines have developed methods that can be relevant to the study of CAT and SCE, and I review these in Section 1.1.3.

While authors have acknowledged that there is no core incompatibility between CAT and SCE (Acerbi and Mesoudi 2015; Sterelny 2017) and a growing number of social scientists is discussing both approaches positively (see the comments to Mesoudi, Whiten, and Laland 2006; and Slingerland 2008), the work I present in this thesis is mainly focused on CAT itself, and less so SCE. While the prevalence of the cultural attraction approach in the French academic landscape naturally had an impact on this choice, I see three other arguments for focusing on CAT. First, it puts cognitive science squarely in the middle of the problem because of the importance of psychology for the study culture. I admit to sharing that opinion (and given current evidence, it is not more than an opinion). Second, its initial goal is not so much to develop a mathematical theory of culture (which

will surely flourish in due time), 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 clear articulation of the philosophical principles defended, discussed, and tested by the theory at the cultural and cognitive levels.[1] # ADD: smaldino-let-2014. As such, it seems that CAT is in a good position to generate productive debate between alternative approaches (such as the propositions coming from social anthropology, and those discussed in Section 1.1.5) which, at the moment, are still competing on the principled level. In particular, the challenges that arise when attempting to study CAT experimentally, as will be the case here for linguistic utterances, should prove helpful in developing a path to the resolution of the critiques faced by both CAT and SCE.

1.1.3 Neighbouring empirical areas

A number of adjacent disciplines have developed interest in and contributed to CAT. Simulations and experimental studies of the evolution of morality (Baumard, André, and Sperber 2013), religion (Boyer 2001), or reasoning (Claidière, Trouche, and Mercier 2017; Mercier and Sperber 2011) have applied the theory; studies on the evolution of language, and of online content through digital media, are also part of this cluster. The latter two areas are particularly relevant for applying CAT to the evolution of linguistic material, as they have already explored factors in this process for a variety of domains, as well as conditions under which gradual changes can build up cumulatively, and the role that such accumulation plays in long term evolution. Essential to these studies are the serial reproduction paradigm and its derivatives, better known today as transmission chains and micro-societies, which have all been used extensively in recent works.

The versatile serial reproduction paradigm

The serial reproduction paradigm was first applied in a series of influential studies on memorisation and recall by Bartlett (1995), who was experimenting with ways of reproducing the evolution of content through its iterated production by and transmission to different people, under somewhat more controlled conditions than what can be achieved in field work. The paradigm was only one among several other techniques introduced by Bartlett, but it came to have a lasting impact on later studies linking memory to culture. Similar to a game of Chinese Whispers, people participate in a chain along which content is transmitted; the experimenter gives a first participant initial material, typically a picture or a short piece of text, with instructions to read or memorise it; that participant is later asked to recall or reproduce the material, and the experimenter uses their output as input for the next participant, thus constructing a chain of successive memorisation (or perception) and recollection (or reproduction) of the initial material. Participants may or may not know that they are part of a chain. The setup approximates the transmission and change process that happens in everyday life, and while it is quite idealised and imperfect it allows experimenters to explore effects of different factors on the process, and examine the evolution of the artefacts produced as a trace or signal of the overall phenomenon at work.

Important independent variables that may vary include the chain structure (e.g. the number and sources of preceding recollections a participant is exposed to), the reading or memorising instructions and context, the interval between exposition and recall, the possible task given during that interval (or, conversely, a possible overlap of exposition and recall, turning the task into copying, or

[1] In the words of Tim Ingold, an outspoken critic of both CAT and SCE: "[Sperber's work] has the virtue of rendering unusually explicit the assumptions built into much contemporary theorising about culture and cognition, and of driving them through to their logical conclusions" (Ingold 2001, 113).

online interaction between participants), the recall instructions and context, the ordering and organisation of participants in the chain, or the ordering of the content itself in the chain (e.g. is a given participant exposed to material from a single generation all in one go, or to material successively sampled from random generations). Typical studies will then analyse the trends in the transformation of artefacts that were produced, possibly comparing the behaviour across different chains. When faced with the complexities of quantitatively analysing the content itself, studies often opt to contrast a simpler measure (such as rate of change) across two different populations, two conditions in the independent variables mentioned above, or two minimally different types of content. In CAT parlance, this method corresponds to the simplest instantiation of the causal chains of public representation to mental representation to public representation (and so on) described by Sperber (1996, 99), and has thus garnered much attention in the literature related to cultural attraction.

Early work contemporary to Bartlett's applied the technique in a variety of settings. A first stream of research explored ways in which verbal memory and social or cultural background can interact. Maxwell (1936), for instance, contrasted the evolution of a story containing deliberate inconsistencies in groups of different social status and age such as soldiers, priests, educated men or women, students, or boy scouts. He found indications that different groups shortened the story at different rates, and conserved or transformed different pieces of the initial story. Northway (1936) also contrasted the conservation of story parts across groups of children from schools with different social backgrounds, relating the proportion of good recollections of a given item to the everyday activities of the children, or the types of transformations (addition, recasting, modification) to the different age groups and to the diversity of backgrounds in a given group. A later stream of research focused more on pictorial content, such as Ward (1949) who related a record of European coin types from 4th to 1st century B.C. to the trends obtained by Bartlett in serial reproduction of pictures. He confirmed that what Bartlett called "representative detached details", that is parts of a picture that represent a distinctive pattern or object even when isolated from the whole, are well preserved both in artificial and historical serial reproductions. Ward thus suggested that comparing historical data to results of serial reproduction experiments can be a useful method to investigate the influence of universal-psychological versus local-cultural factors in theories of historical evolution of artefacts. Hall (1950) further explored hypotheses made by Bartlett on the effect of titles for pictorial and verbal serial reproduction: in those new experiments, it appeared that titles had a considerable influence on the reproduction of both pictures and texts, by acting as an interpretive frame that guides (or confuses) the participants in interpreting the material.[2] More recently the reliability of Bartlett's studies has been discussed, shifting the focus towards a better control of experimental conditions: Gauld and Stephenson (1967), while still praising Bartlett for the progress that his work represented compared to that of his predecessors, showed that changing the exposition instructions, for instance by giving strict memorisation instructions, or by adding a simple sentence asking participants to be "accurate", would considerably reduce the transformation rate in the resulting chains, a fact that should question the effects measured by a setup that lacks explicit memorisation instructions. The authors also examined the level of conscientiousness of participants in accomplishing their task, showing that the measure correlated negatively to the participants' transformation rate. Noting that the effect was not explained away by a measure of intelligence of participants, they finally suggested that "errors could, it seemed, be avoided, if the subject was so inclined" (1967, 45). Today, one overcomes such an issue by designing experiments that create an intrinsic motivation for participants to adopt the behaviour under study (see Claidière, Trouche, and Mercier 2017 for an example). In contemporary work, Kashima (2000b) offered a reappraisal of the social aspect of Bartlett's contri-

[2] Hall, noting that "the function of the headline or title to some story or article is that of giving a particular emphasis to certain aspects of the text, and is one of the main methods of distorting and biasing what is remembered" (Hall 1950, 120), later reflects on the fact that such results have crucial political implications given the development of mass media.

butions. Kashima argues that while Bartlett's legacy is mostly seen in psychological studies that adopt methodological individualism, Bartlett fiercely opposed that approach; indeed, his view of the interaction of culture and psychology had much in common with views from social science such as those evoked in Section 1.1.1, and with today's social psychology view of a deeper integration between culture and biology.

One recurrent problem, found across most of the above studies, is the difficulty of quantitatively analysing meaning in the material transmitted along chains. Northway (1936), through her focus on meaningfulness, Hall (1950, 120) and Gauld and Stephenson (1967, 42) all indicate that analysing content itself is a laudable but yet unreachable goal. In facing that difficulty, studies recur to analyses of form, survival, or other measurable aspects of the participants' productions. For linguistic material for instance: the length of a recollection, the number of words, concepts, or propositions accurately recalled, or a contrast of the concepts conserved at the end of the chains. Today the problem remains, and contemporary studies of meaningful material use the same techniques to index or approximate changes in the content of artefacts produced (indeed, the work presented in this thesis will be no exception). While much progress can be made by using these approaches to quantifying aspects of the material, or by instead focusing on artefacts that bear no content, it seems a consensual account of meaning, and a corresponding means of analysis, will be necessary for a full-fledged theory of cultural change to establish itself. I will return to this issue in Section 1.2.

Contemporary revival

Owing to the development of SCE and CAT, the last two decades have seen a regain of interest for the serial reproduction paradigm (now known as a transmission chain) and its derivatives (generally known as cultural transmission experiments), resulting in the development of new case studies and methodologies. A number of effects have been catalogued by recent works. Bangerter (2000), for instance, showed that in transmitting a scientifically styled account of human sexual reproduction, participants tended to personify ovum and sperm, and attribute stereotypical gender roles to them. Mesoudi and Whiten (2004) have argued that the loss of detail that is repeatedly observed in transmission chains, and in particular in the transmission of reports of everyday events, is due to a hierarchical encoding of memories that biases participants' recollections in favour of higher-level descriptions. Works studying the evolution of religion have focused on the effect of counter-intuitive information: by studying transmission chains of stories made of elements with varying degrees of counter-intuitiveness, Barrett and Nyhof (2001) and Norenzayan et al. (2006) observed a conservation advantage for minimally counter-intuitive elements which supports the eponymous Minimal Counter-Intuitiveness account of religion (Boyer 2001; and see Purzycki and Willard 2016 for a critical discussion). Other similar transmission advantages have been identified in relation to stereotypes: for instance, elements of a story that are consistent with gender stereotypes are less degraded than elements that are inconsistent with such stereotypes, but only when relevant to the story plot (Kashima 2000a); stories made of social information, that is featuring human interactions and plots, are also better transmitted than stories involving non-interacting people or than stories about physical nonhuman elements alone (Mesoudi, Whiten, and Dunbar 2006). The extensive reviews provided by Mesoudi and Whiten (2008) and Whiten, Caldwell, and Mesoudi (2016) give a broader idea of the effects studied and methods used in the literature.

ADD: bebbington-sky-2017, stubbersfield-serial-2015

A more recent stream of research explicitly focuses on the individual transmission step in a chain. For instance, setups with two parallel chains that cross-fertilise each other have been shown to improve transmission rates (Eriksson and Coultas 2012): if at each generation, the participants of two parallel

chains both receive two inputs, one from each chain at the previous generation, information loss is decreased compared to a single chain where participants read twice the same piece of content (of the previous generation). Acerbi and Tennie (2016) further modelled such error-correcting redundancy, simulating minimal scenarios that could favour its evolution given fixed cost-reward constraints. Eriksson and Coultas (2014) also decomposed real world transmission into three phases: choose-to-receive, encode-and-retrieve, and choose-to-transmit. Focusing on emotional selection studies showing that participants are more willing to pass on stimuli that elicit disgust (Heath, Bell, and Sternberg 2001), the authors showed that any of those phases can be the target of a selection pressure. Such pressures can additionally counteract each other, showing that the transmission process is more complex than was initially assumed. Studies of digital media have adopted the increased level of detail, as digital communication lets users copy-paste and bypass the encode-and-retrieve phase of transmission (Acerbi 2016).

Methodological exploration has also led to more free-form interaction setups: in a study of the transmission of risk perception, Moussaïd, Brighton, and Gaissmaier (2015) chose to make the interaction and underlying transmission of information an open-ended process. After initiating a chain by providing the first participant with a set of documents to read on their own, later participants were left to talk freely in successive dyads, and the whole session was recorded on film for subsequent analysis. An analogous change of setup was operated by C. A. Caldwell and Millen (2008b; 2008a) who investigated the cumulative aspect of the evolution of building techniques in a series of experiments asking participants to construct spaghetti towers or paper planes, later evaluated by their height and flight distance. A crucial point in those setups, Sterelny notes (2017, n. 12), is that participants could observe the preceding generation during their experimenting and building, meaning they had access to rich context for the learning phase of transmission. Taking another step to embrace interaction with a setup where participants had to repeatedly build Lego cars in pairs (though without transmission), McGraw et al. (2014) take the complete joint interaction to be their object of study and propose to use the constructions resulting from interacting dyads as a trace of the processes that took place in the interaction. By looking at the characteristics of the cars built by participants, the authors claim to present "methods for discerning, and quantifying, schema-like intersubjective understandings in material form" (McGraw et al. 2014, 4; see also Mitkidis et al. 2015; and Wallot et al. 2016), effectively integrating critiques that I discuss below in Section 1.1.5.

The works reviewed here show that there is much room for exploration, on one side, of the methodological choices in transmission chain experiments (some of which will turn out to be more important than others in the trends observed), and on the other, of the theoretical background that sustains a given study. Such experiments have therefore much to bring to underlying theoretical debates.

Language change: Experimental semiotics and Iterated learning

ADD: tamariz-experimental-2017

A related strand of research has developed similar methods to study the emergence and evolution of language. A central experimental paradigm is that of iterated learning, which resembles a transmission chain for artificial languages where participants learn a communication system at each step: a first group of participants must learn to use a simple artificial language (e.g. an artificial vocabulary for naming a range of objects), after which a second group of participants must learn that language through some interaction with or transmission from the first group. The process is then iterated over successive generations, leading to the evolution of the artificial language initially introduced. A related setup used in experimental semiotics surfaces the evolution of interaction without transmission across generations. It consists in pairing participants and assigning them a task that

they must cooperatively solve over repeated iterations, without changing partners. Most often, participants can use a communication channel (or must create it) to help in coordination, such that the setup exposes the way participants iteratively develop conventions over the channel. Studies using these two paradigms have shown that simple biases in participants' interactions and learning capabilities can lead the final evolved communication system to exhibit non-trivial structure. Several factors have been shown to influence the process, including the structure of the objects referred to by the artificial language, the transmission or interaction task (and its surrounding context) for iterated learning, and the reinforcement rules favouring expressivity of the language that will interact with learnability pressures.

A first major goal in this stream is to provide an non-nativist account of the emergence an structure of communication systems (Kirby, Dowman, and Griffiths 2007). Galantucci (2005), for instance, studied pairs of participants facing a cooperative task that required them to develop ad hoc communication conventions. The communication channel, a form of shared whiteboard, distorted participants' input by a constant drift such that they could not enter letters or pictorial drawings. The study highlighted the variety of strategies used by participants to develop conventional sign systems, and in particular the way those systems were adapted to the interaction history of dyads. The authors showed that sign systems were deeply meshed with the situated and time-dependent information that was available, allowing the participants to solve their task by dynamically coordinating (Galantucci 2005, 748–49). Further stripping down the set of assumptions built into the experimental setup, Scott-Phillips, Kirby, and Ritchie (2009) created a cooperative task without providing any communication channel parallel to the task itself. To solve the challenge, participants had to develop a communication system where their own behaviour in the task could become an embodied communicative signal, aside from accomplishing their action for the task either simultaneously or at other times. Here too, the authors underline the importance of developing a dialogue, based on common ground provided by the history of interactions, to bootstrap the creation of a general communication system that can solve the task in all situations. These studies show that communication systems are inherently embodied and situated, but still involve some degree of intentional design by the participants.

Iterated learning setups, on the other side, have focused on the unintentional emergence of structure. Kirby, Cornish, and Smith (2008), for instance, studied the evolution of an initially random artificial vocabulary set which participants had to learn, use, and extrapolate to name events (events were a combination of a shape, a colour, and a movement). By filtering ambiguities in the output vocabulary produced by one participant and given to the next, the authors were able to create a pressure for expressivity of the complete vocabulary; combined with the learnability pressure inherent to the task, and the fact that participants had to extrapolate to unknown events, the vocabulary gradually evolved to regularise variation with the emergence of compositionality corresponding to the three dimensions of the events to name. Crucially, the participants were not aware of the goals of the experiment, nor that they were part of a chain: the emergent structure in the vocabulary thus appeared without intentional design. Using a comparable setup where participants had to extrapolate a colour-naming vocabulary, a subset of which was then transmitted to the next generation, Xu, Dowman, and Griffiths (2013) observed that (probably culture-specific) biases in colour grouping played an important role in the convergence of vocabulary terms. Combinatorial structure and distinctiveness has also been shown to emerge in a set of acoustic signals devoid of meaning that participants had to learn and reproduce (Verhoef, Kirby, and Boer 2014). Cornish, Smith, and Kirby (2013) encountered similar results for sequences of categorical items that had no inherent meaning: since participants must reproduce a whole set of sequences, or acoustic signals, at each generation, the set behaves as a interconnected system for which learning pressures gradually increase the combinatorial structure.

Experimental semiotics and iterated learning also cross-fertilise each other: after Garrod et al. (2007) used a Pictionary-like collaborative task, without transmission, to study the emergence of symbolism in a lexicon, a process they called "interactive grounding", Fay et al. (2010) extended the findings to a micro-society. Fay and colleagues showed that globally shared symbols can emerge through the gradual alignment of such interactive groundings, leading to an increasingly refined and streamlined symbol system. The authors thus proposed "symbolisation" as an additional mechanism in the emergence of communication systems, based on intra-generational collaborative coordination through interaction, and parallel to the inter-generational learning biases and bottlenecks studied by iterated learning studies. Taking another leaf from experimental semiotics, Winters, Kirby, and Smith (2015) exploited the observation that a large part of the meaning of an utterance comes from its situational context: they studied the influence of the situations in which participants use vocabulary items on the structure of the language that evolves from interaction and iterated learning (thus extending Silvey, Kirby, and Smith 2015). The authors show that, if the situations in which participants communicate do not contrast items on all the dimensions on which they differ, then the vocabulary set often evolves to not encode those dimensions. In other words, if the usage situations shield the participants from certain contrasts between items, the final vocabulary is often under-specified with respect to the full item space: it does not encode the dimensions that discriminate the unobserved contrasts, instead adapting to be useful only for the contrasts that users observed. This stream of research is active and gradually relaxing the constraining hypotheses made by initial studies. Carr et al. (2017), for instance, recently made the space of items more realistic by exploring the emergence of vocabulary sets where participants communicate about a continuous unbounded set of items.

A second, closely related question concerns how an already structured communication system evolves given a set of external, learning, or interaction pressures. Croft (2013) provides a general framework for this question, inspired by principles from biological evolution and by recent debates on the nature of the evolutionary process, and crucially focused on identifying an adequate unit of analysis for the evolutionary study of language change. Indeed the author combines two key insights. On one side, Hull's General Analysis of Selection (Croft 2013, 16) lets him abstract out the principles of evolutionary processes and distinguish between replicator, interactor, and selection, three core components that provide "a model for disentangling different cultural evolutionary processes and identifying their interconnections" (Croft 2013, 18). On the other side, he draws on the critique that Developmental Systems Theory opposes to a gene-centred view of biological evolution (Oyama, Griffiths, and Gray 2001), and insists that a theory of language change should consider utterances to be full life cycles made of pronunciation, meaning, and interpretation in context. He thus proposes the Theory of Utterance Selection, which takes linguemes (i.e. the linguistic structure of sounds, words, constructions and utterances) to be replicators, but always part of a larger cycle; language speakers are the interactors (2013, 16), and the theory defines language as the "population of utterances in a speech community" (2013, 35).

A number of existing studies fit well in this framework. Tamariz et al. (2014), for instance, used a common setup requiring participants to develop a pictorial vocabulary for a pre-given set of words, and modelled the trends participants exhibit in adopting new signs as they go through the interactions of the experiment. The authors found that participants do not select new signs neutrally; rather, they tend to favour signs they have used in the past, even if their partners use different ones, unless they encounter a sign they find obviously superior in representative power. Similarly, in a picture-description transmission chain using an artificial minimal language, Smith and Wonnacott (2010) showed that the accumulation of individual participant biases will regularise the marking of plurals in the evolved language. Kirby et al. (2015), while studying the emergence of structure as the result of combined pressures of expressivity and compressibility (the latter often attributed to learning, Tamariz and Kirby 2015), hint to the fact that the way structure emerges and evolves is

highly dependent on the combination of such pressures. They note, in particular, that "there is some suggestive evidence that structure in language can be modulated by the composition of populations" (Kirby et al. 2015, 99): different communication patterns at the population level, or a different fabric in the population responsible for the transmission and evolution of a language (e.g. more second-language learners, or more children learners), should lead to differences in the evolution of language structure. Regarding symbolism, Caldwell and Smith (2012) extended the micro-society Pictionary-like task studied by Fay et al. (2010) to one where participants were gradually replaced, inducing increased symbolism and successful transmission of the evolved symbols at the same time. Initial members of the micro-society constructed highly iconic representations of the meanings to convey, but as the experiment introduced newcomers to those signs through observation and consequent use, the drawings gradually lost the iconic link to their referent and became simpler.

Large portions of the iterated learning literature draws on and contributes to a parallel theoretical track which laid down the first analytical predictions for models of Bayesian agents learning and producing languages in chains. Griffiths and Kalish (2007) were the first to show that the analytical structure of iterated learning with uniform Bayesian agents can correspond, depending on the way the agents produce new iterations, to well-known statistical inference methods (Gibbs sampling and a flavour of the EM algorithm). In such a setup, iterated learning predictably converges towards distributions determined by the internal prior distributions of agents (i.e. their inference bias). As a consequence, in those analytically derived situations one can straightforwardly predict the final distributions that should evolve under iterated learning, a fact that Kalish, Griffiths, and Lewandowsky (2007) verified with humans in a function learning task. Griffiths, Christian, and Kalish (2008) further exploit this result by using it in the reverse direction: since the outcome of iterated learning, for specific setups, is predictable on the basis of participants' priors, one can use such experiments to investigate the inductive biases of participants. The authors confirmed this, showing that the method infers well-known participant biases in category learning tasks. Griffiths, Kalish, and Lewandowsky (2008) then explored the relevance of these findings for the study of cultural evolution, showing in particular that individual cognitive biases can have significant effects on long-term cultural evolution. Reali and Griffiths (2009) further related those results to the evolution of vocabulary, showing that they are consistent with experimental cases of word-meaning mapping regularisation. Finally Perfors and Navarro (2014), through analytical derivation and experimental confirmation, reintroduced the impact of the external world in those results; the authors showed that the structure of referents (i.e., the external world) will also play a role in the final evolved language, provided it has an effect on the choice of items people actually talk about (versus, the choice of items talked about only depends on the language itself).

Scott-Phillips and Kirby (2010) and Tamariz and Kirby (2016) provide further reviews of the iterated learning literature, and Galantucci, Garrod, and Roberts (2012) and Roberts and Galantucci (2017) offer reviews of experimental semiotics. An interesting and important development in recent works is the reintroduction of pragmatics into theoretical questions. Scott-Phillips (2017), in particular, reaffirms the central role of pragmatics in the creation and understanding of meaning in context, and argues for a much stronger focus on the evolution of pragmatics itself, that is, as he envisions it, on the evolution of *ostensive communication*. Let me close this review by noting that there is an increasing convergence both in the literature and in empirical questions, of the cultural evolution and language evolution fields; the intersection of questions from the two fields is likely to push theoretical issues forward.

ADD: sperber-beyond-2015

Digital media

Acerbi (2016) defines digital media as "media encoded in digital format, typically to be transmitted and consumed on electronic devices, such as computers and smartphones". The ubiquity of this medium, which created the ongoing avalanche of available digital traces, has opened both questions and possibilities for the study of cultural evolution over the past 15 years. Indeed digital media is both a measurement tool and an object of study, as it has become embedded in everyday life in many societies, with its own practices of interaction, mediation, or transmission, possibly impacting cultural evolution. While digital practices are different from those in physical encounters, the digital transition remains an addition to the possible range of interaction media, and the cultural evolution framework can study it as such, with increased access to the artefacts those interactions produce. Acerbi (2016) argues precisely for such an approach to digital media, and reviews relevant works that have explored that space. In what follows I present three areas of focus that have received particular attention in the literature.

A core—and somewhat canonical—challenge for digital media has been to describe the behaviours of diffusion and change of artefacts in social networks, and if possible predict their macroscopic spread and evolution. The question is far from new (see Rogers 2005) and works have historically tackled this question through analytical models, simulations and empirical studies, but the recent increase in access to digital traces and computing power to make sense of such data has boosted empirical developments. Gathering data from blogspace, for instance, has allowed studying the propagation of information topics, as Gruhl et al. (2004) did by separating topics into "chatter" and externally-triggered ("spike") subjects to model their spread over the social network formed by users. The email network is another source of digital traces, with patterns specific to it; indeed Liben-Nowell and Kleinberg (2008) showed that information diffusion along email chains has an unexpected deep tree-like structure, which they suggest is because of the asynchronous nature of email. Such studies focus on *socio-semantic systems*, that is systems made of, on one side, a collection of users whose interactions or links form a social network, and on the other side, a set of topics or subjects around which the users interact, which also features a network-like structure. The two levels of structure reciprocally influence each other, as Cointet and Roth (2009) show for blogspace (see also Cointet and Roth 2007 who explore the relative roles of social network topology and transmission rules, related to the structure of topics in the spread of information).

The scale of the study of social networks has grown considerably over the past decade, and linguistic memes in particular have received much attention. In a landmark endeavour, Leskovec, Backstrom, and Kleinberg (2009) gathered and published a data set of quotes extracted from a million blogs and news outlets over a nine month period, and developed a method to group minimally different occurrences into quotation families in order to quantify the popularity of news topics over time. The technique allowed the authors to study the evolution of the online news cycle, measuring differences in publication timings across blogs and news outlets. Simmons, Adamic, and Adar (2011) further analysed that data set, showing that transformations of quotes upon copy are frequent (contrary to what one would expect for such memes), work that Omodei, Poibeau, and Cointet (2012) then extended with a more accurate multi-level transformation model. Adamic et al. (2016) developed a similar study for the evolution of explicit memes (that include instructions asking the reader to copy and pass on the contents of the meme) in a Facebook data set of hundreds of millions of occurrences; by using a biological evolutionary model of mutation and replication where genotype corresponds to the meme's content and phenotype to the copying instructions, the authors explore the implications of pushing the biological analogy to its limits in such a paradigmatic case. The range of empirical questions, and the technical challenges involved in tackling them, are such, that the focus has moved towards developing methods for the collection and study of similar data sets. For instance, the

MemeTracker project initiated by Leskovec, Backstrom, and Kleinberg (2009) has now evolved into a fully-fledged network collection and analysis platform (Leskovec and Krevl 2014) with associated data sets (Leskovec and Sosič 2016). Another noteworthy example of this is the development by Moritz et al. (2016) of text re-use detection methods for historical works, a technique that could open the application of the above studies to digitised historical corpora.

A second research stream isolates the different processes involved in the spread and change of artefacts. In particular for transformation, separating effects of content from effects of context is a necessary step to understand the processes responsible for the changes of artefacts. Danescu-Niculescu-Mizil et al. (2012) thus studied the memorability of movie quotes by identifying features that can predict quotes marked as memorable by users of the Internet Movie Database (call these IMDb-memorable): from about 1000 movie scripts, the authors extracted around 2200 pairs of quotes, each consisting of one IMDb-memorable quote paired with the closest quote in the movie script that has the same length, is spoken by the same character, and is not IMDb-memorable. By contrasting these pairs, the authors surface the content-related features of a quote that make it memorable, and factor out the context in which the quotes appear, context which otherwise plays an important role in the memorability rating. After checking that human subjects can identify which quote in the pair is memorable (they do so with an average 78% success rate), the authors show that memorable quotes, on average, use less frequent vocabulary, more frequent grammatical categories (POS tags), and more general constructions (fewer 3rd person pronouns, more indefinite articles, etc.) that make them more adaptable to changing contexts (each of these measures, taken individually, partitions the quote pairs into two subsets containing about 40% vs. 60% of the whole set). Cancelling out context effects to develop content-related features has become a widely used approach, with adaptations ranging from the identification of linguistic markers of politeness in online content (Danescu-Niculescu-Mizil et al. 2013) to the measurement of attractiveness of famous quotes (Acerbi and Tehrani 2017). In a study reminiscent of Hall (1950), Acerbi and Tehrani (2017) compared the relative strength of content and presentational context in a sample of famous quotes that participants had to rate for attractiveness. The authors compared conditions where quotes were presented alone, versus presented with random attribution to more or less famous personalities, or versus presented with a random popularity score. They found that such minimal context has little effect if any at all: attribution, famous or not, bears no effect on the attractiveness of a quote, and popularity has little. Althoff, Danescu-Niculescu-Mizil, and Jurafsky (2014) also opened the study of context versus content to relational variables, by showing how social status and presentational features (such as showing a strong need) can affect the success of requests on Reddit.

A third related stream of research focuses more specifically on influence in social networks, and its links with attention: what network effects trigger the diffusion of a particular meme or piece of information? Among the micro-processes involved in the spread of information in networks, what is the role of influence across connected nodes? Bakshy, Karrer, and Adamic (2009) investigated the question of social influence by examining information cascades in Second Life. Information cascades, where a comparatively small initial event triggers large scale diffusion, are a well-known phenomenon in social networks, and their size distribution is well modelled by peer-pressure threshold models which link the cascade behaviours to the topology of the network in which they occur (Watts 2002; Ruan et al. 2015). Bakshy, Karrer, and Adamic (2009) thus tracked the spread of *assets* in the virtual world provided by Second Life (that is pieces of content introduced and copied by players in the game); they find that a significant part of contagion happens along the friend network, instead of in avatar-to-avatar interactions, indicating that the adoption rate of (in-game) social circles has a strong impact on a person's adoption of an asset in Second Life (see Bakshy et al. 2011 for another example study, on Twitter, separating the strength of content from the strength of social influence).

Attention in social networks is another related factor. Considering the amount and constant flow of information available, filtering and attention management is a necessary component of the diffusion of artefacts; it is usually accounted for through competition among pieces of information. Weng et al. (2012), for instance, model the spread of Twitter hashtags through agents with bounded memory and attention, and show that such simple assumptions account well for the distribution of hashtag diffusion along the social network. The relationship between attention and strength of ties has also been explored by Weng et al. (2015) in data gathered from Twitter, cell phone, and email networks. In these data sets, the authors confirm that while strong ties transport the majority of events, users devote comparable attention levels to both strong and weak ties; they suggest that strong ties play a social communication role, while users use weak ties for seeking novel information, a distinction which could explain the different attentional patterns they measure across the different media.

The empirical study of information diffusion and spread has steadily grown since the advent of digital traces; the number of factors included in analyses is growing, and the influence of core processes such as attention is gradually becoming clearer (an interesting addition would be the role of power relationships, which are also detectable through markers of interactive behaviour, Danescu-Niculescu-Mizil et al. 2011). As mentioned above, Acerbi (2016) provides a useful overview of other works that are relevant for current questions of cultural evolution.

Conclusion

As I discuss in Chapter 2, the development of data set collection and analysis methods can bring insight, as well as refined questions, to the study of the reciprocal influences between cognition and culture. # TODO: "I" or "we", for BCP? Other empirical fields in psychology and linguistics are useful to the study of CAT as a framework for cognition-culture interactions: I further introduce works in psycholinguistics relevant to the study of quotes online in Chapter 2, and Chapter 4 will return to how future works could make deeper use of "Smartphone Psychology" to contribute to the more contentious issues. # TODO: actually do that. Let me now move on to the most debated developments of cultural evolution, the criticisms opposed to the approach, the alternatives emerging from these critiques, and the possibilities of reconciliation.

1.1.4 Developments

The exact nature of evolution is subject to debate in biology and philosophy of biology, and some of the recent developments have made their way into the core of cultural evolution theory. In a parallel movement, the nature of cognition is itself debated inside philosophy of cognitive science. The aim of the next two sections is to briefly discuss the relevant parts of those debates for the cultural evolution approach, delineating first the elements that could be—or are already partly—integrated into mainstream CAT, and second the critiques which, at least in current writings, seem to call for a partial rethink of the paradigm. The questions here are more theoretical than above, as they explore both the way different disciplines studying life are best meshed together, and what core components should be at the root of such a convergence. This is not to say the debates in biology and cognitive science concern theory alone, as each debated position is well supported by empirical work; rather, up to now those works have not translated to actionable contradictory predictions in the study of cultural evolution proper. Nonetheless, I will argue in Chapter 4 that these debates provide crucial context to understand a particular practical challenge in the empirical study of CAT, namely the definition of the meaning of representations and its impact on the dimensions of attraction. I begin with developments then, that is the elements that seem possible to integrate into CAT. These also

lay some of the groundwork for the subsequent criticisms, which I believe challenge CAT closer to its foundations.

Niche construction theory

A central question in the study of evolution is the definition of what counts as heritable material, for which two broad views are competing. The debate, agreements and disagreements between both views are well documented, and I base the following discussion on the recent reviews provided by Laland et al. (2014) and Scott-Phillips et al. (2014). The standard account of biological evolution, or Standard Evolutionary Theory (SET) as termed by Scott-Phillips et al. (2014), defines evolution as "change in the frequency of DNA sequences (i.e., genes and associated regulatory regions) in a population, from one generation to the next" (Scott-Phillips et al. 2014, 1232; referring to Futuyma 2005). Such change occurs through what is known as an evolutionary process:

> "Evolutionary processes are generally thought of as processes by which these changes occur. Four such processes are widely recognized: natural selection (in the broad sense, to include sexual selection), genetic drift, mutation, and migration (Fisher 1930; Haldane 1932). The latter two generate variation; the first two sort it." (Scott-Phillips et al. 2014, 1232)

In this view, DNA sequences constitute the principle heritable material transmitted from parent to offspring across generations, and their distribution and change should be the main focus of evolutionary theory. Furthermore:

> "There are many factors that can cause these four evolutionary processes to occur, and for the skeptics [of Niche Construction Theory], niche construction is one such factor." (Scott-Phillips et al. 2014, 1233)

Niche construction is the process by which organisms engineer their own and other organisms' environment in ways that are often beneficial to them. A classic example of such niches are the dams built by beavers along the rivers they inhabit; a beaver-built dam creates a local lake, and its presence actively changes the environment in which future generations of beavers—as well as neighbouring organisms—develop. The constructed niche is inherited across generations such that it can have a lasting impact on the selection pressures under which later generations evolve. SET recognises this phenomenon and defenders of the classical account are among those who actively study it (Laland et al. 2014). Niche Construction Theory (NCT, Odling-Smee, Laland, and Feldman 2003), however, contends that increasing amounts of evidence are unsatisfactorily accounted for by SET (though not in contradiction with it), and proposes "a broadened concept of inheritance, including 'ecological inheritance,' the modified environmental states that niche-constructing organisms bequeath to their descendants" (Scott-Phillips et al. 2014, 1233). Those constructed environmental states bias the natural selection of later generations (so-called "selective niche"), and also affect the social and ecological environment in which offspring develop (so-called "ontogenetic niche"), both being processes that can lead to evolutionary feedback loops. The evolution of dairying, as analysed by O'Brien and Laland (2012), is claimed as a paradigmatic case that is well accounted for by NCT.

ADD: gilbert-eco-evo-devo:-2015 if clearly aligned

NCT is part of a broader movement in evolutionary biology that seeks to integrate a strong view of such feedback dynamics into evolutionary theory, by combining evolutionary developmental biology ("Evo-Devo") with the evolution of the environment in which development take place. These

works argue for an Extended Evolutionary Synthesis (EES, also presented as Eco-Evo-Devo by Gilbert, Bosch, and Ledón-Rettig 2015), conceiving of evolution as the co-evolution of organism and environment, a system that inherits genetic material, but also constructed selective and ontogenetic niches. The crux of the disagreement with SET lies in the importance of the dynamics that this feedback generates: SET considers it more parsimonious to define evolution as change in frequency of DNA sequences, and thus frames niche construction and other ecological inheritance processes as a cause for changes in DNA. Conversely, EES considers it more *fruitful* to define evolution as change in the whole organism-environment system, for which niche construction is a core evolutionary process, like genetic mutation or natural selection. According to Scott-Phillips et al. (2014), the current evidence does not tease the two perspectives apart unequivocally: all known phenomena can still be explained by both approaches with varying degrees of shoehorning, and predictions from each theory can be rephrased into the other one (although such inseparability might not last). However, proponents of EES argue that the study of ecological processes in evolution, while present in SET, has become systematic only thanks to the change of focus brought by the development of NCT.

The Extended Evolutionary Synthesis offers a natural framework for the study of all aspects of evolution, be they cultural or biological, and indeed the gene-culture co-evolution framework fits well with this synthesis. The communities developing those approaches overlap partially (Marcus Feldman, notably, is a core contributor to both research streams), and Sterelny (2017) argues that NCT is a core—if sometimes implicit—component of both Californian and Parisian cultural evolution. Indeed, EES is capable of integrating a non-opinionated notion of culture as part of the organism-environment system under study, and the task at hand then joins up with that of dual inheritance theory, presented above: identifying the co-evolution dynamics of genetic and environmental inheritance channels. On this view, then, culture is accounted for by a blend of ecological and cognitive-epistemic niche construction processes.

"4E" cognitive science: the extended mind

TODO: use chemero-after-2008 for this summary

In a strikingly parallel movement in cognitive science and philosophy of mind, the nature of cognition and its units of analysis have been debated along two broad dimensions (see Chemero and Silberstein 2008 for a detailed review of questions and possible answers). (1) The boundaries of cognition: are cognitive processes brain-bound, do they extend to the body, or do they include the environment or the surrounding (cognitive) organisms, and if so in what sense (Clark and Chalmers 1998; Menary 2010). (2) The role of time-dependent dynamics, and the corresponding construal of the nature of cognition: are cognitive systems best described as digital computers processing information in the form of representations (i.e. symbol processing systems), where time can often be reduced to an ordering of events, or are they best described as dynamical systems where time is important to define rates of change, flows, or dynamic couplings (Gelder 1998; Beer and Williams 2015). In treating these questions, the extended, embedded, embodied, and enactive approaches to cognitive science (the so-called "4E") have argued to various degrees that cognition is not only (or not at all) an information-processing operation that can take place in the void, but also (or exclusively) a situated activity supported by (or a dynamic coupling with) its environment. The extended mind theory, among the less radical 4E approaches, is quite compatible with EES and thus with both Californian and Parisian cultural evolution. Indeed Sterelny (2010; 2012), building on NCT, has argued that the extended mind approach is a special case of epistemic niche construction. He suggests that the environments human beings grow in are the result of cumulative cognitive niche construction processes, that engineer the material and social environment of humans to support the

growth of everyday cognitive capacities, thus scaffolding cognition during development and life.

1.1.5 Criticisms

Developmental systems theory

A related and somewhat complementary extension to the standard evolutionary account has developed in parallel to NCT, with a more radical notion of extended inheritance: Developmental Systems Theory (DST, Oyama 2000; Oyama, Griffiths, and Gray 2001). Kim Sterelny characterises it, on one side, with three critical theses:

> "(1) We cannot simply assume that the organism/environment boundary is of theoretical significance for developmental and evolutionary biology … (2) It may be legitimate to foreground genetic structure and genetic change for specific explanatory or predictive purposes. But in general, the genes an organism carries are just one set of developmental resources among many. Genes and gene changes are important both to development and to evolution, but they are not of primary or privileged importance. (3) Developmental systems theorists are skeptical about the project of explaining intergenerational similarity by appealing to the transmission of phenotype-making information across generations." (Sterelny 2001, 335)

Crucially, DST claims that overlooked evidence in development indicates that there is a "causal parity" between genes and non-genetic development factors, such that evolutionary theories should not give greater (or smaller) importance to the former over the latter. This in turn sustains the third thesis: views of the genome as a bearer of biological information should be qualified in light of the complex interactions between developmental processes in which genes participate (for further detail, see Griffiths and Stotz 2013, who extensively review the ways in which genes can be, or historically have been, considered to bear information). Now the other side of Sterelny's characterisation of DST is its positive story:

> "The positive program of developmental systems theory is that the fundamental unit of evolution is the life cycle. In turn, the life cycle is the set of developmental resources that are packaged together and interact in such a way that the cycle is reconstructed. The most obvious life cycle is that of the organism plus its immediate environment, but developmental systems theorists are open to the idea that cycles will exist at both finer and coarser grains." (Sterelny 2001, 335)

A driving goal in DST is to recontextualise, explain, and if possible do without the conceptual divide between matter and form, that is between specification and realisation, which underlies most discussions of nature-nurture (Oyama 2000). This implies putting an endogenous account of the concept of information at the centre of its theory, by focusing on the way such information is generated through the dynamics of a system as it develops and as its resources interact.

In many aspects, NCT-EES and DST complement each other (see Griffiths and Gray 2005 on the complementarity of DST and Evo-Devo in particular). However for the current purposes I suggest we locate both approaches with respect to the two following questions:

- How far should models and theories of evolution integrate the detailed processes of development and environmental interaction to provide an accurate picture of evolution? This question is not about finding the right level of descriptive complexity, but about finding the relative importance of each level of complexity. In other words, it asks what is the shape of the cost

function representing the trade-off between parsimony and explanatory power (rather than where on that function should a theory operate). The classical account of evolution suggests development can be usefully abstracted away, such that analysing gene flow with the four recognised processes (natural selection, genetic drift, mutation, and migration) can account for all important evolutionary dynamics. NCT claims that organism-environment interactions can generate dynamics that do not fit into the standard account but have long term effects nonetheless, warranting an extension of the evolutionary processes considered. DST further claims that evolution's unit of analysis should be the full life cycle of a developmental system (organism or other), which makes it more difficult to abstract out absolute information items but guarantees an accurate view of the causal parity of developmental resources, such that developmental processes are not obviated.

- What notion of information should evolutionary theory rely on? Classical evolutionary theory and NCT rely on the idea that DNA bears the biological information that is used throughout development, in an interaction with the environment. Conversely, DST refuses to conceptually separate inherited from acquired traits, a separation it sees as unnecessary for a populational approach: instead it adopts a relational notion of information as the co-product of development, "not by special creation from nothingness, but always from the conditional transformation of prior structure—that is, by ontogenetic processes" (Oyama 2000, 4).

Notably, the stances adopted by DST put it at odds with a dual-channel account of cultural evolution. While the developmental systems approach is compatible with population thinking, and thus with darwinian approaches in general (Griffiths and Gray 2005), it sees no theoretical reason to conceptually isolate different genetic and cultural channels of inheritance. Aside from genetic material and ecological niches, developmental systems also inherit epigenetic material, behavioural patterns and communication systems (Jablonka 2001) and more generally the full matrix of resources involved in the development of the system at its next life cycle. It might be useful for practical or descriptive reasons to focus more on some resources than on others, as is done for instance in the study of phylogeny, but doing so does not change the developmental interactions and causal parity of those resources with the rest of the set. A similar point will be made, below, about representations for the study of cognition.

Convergence

The notion of developmental system, and the importance it gives to the ontogenetic niche in which organisms grow, is convergent with 4E cognitive science and Sterelny's scaffolded mind approach. Aside from providing a natural account of culture similar to that of EES, DST also integrates the role of the ontogenetic niche in development all the way up to the cognitive level (Stotz 2010). Wimsatt and Griesemer (2007) have argued that an appropriate focus on scaffolding and development could shed light on the mechanisms of inheritance across generations, a necessary step in an account of cultural evolution. In agreement with DST, however, the authors feel that fixed- or dual-channel accounts of inheritance based on a notion of information transferred to offspring[3] obviate the role of development. Conversely, grounding an account of constancy and change in developmental processes requires considering all the resources that a system inherits, and their interaction in development.

In spite of such differences, the literature indicates that many aspects of EES, DST, and 4E cognitive science seem possible to fruitfully integrate for the study of cultural evolution. Indeed, DST is

[3]The definition provided by Boyd and Richerson (2005), and adopted by Mesoudi, Whiten, and Laland (2006), is along these lines. The authors define culture as "information capable of affecting individuals' behavior that they acquire from other members of their species by teaching, imitation, and other forms of social transmission" (Boyd and Richerson 2005, 6).

compatible with the population approach necessary for a darwinist analysis of culture (Griffiths and Gray 2005; Lewens 2012, 477), which is one of the core elements on which both the Parisian and the Californian cultural evolution streams are built. Lewens (2012, 474) also remarks on the encouraging fact that Russell Gray is both one of the main authors of DST, and is now an outspoken proponent of cultural evolutionary theory. A final case in point is the successful application of these ideas in theoretical and empirical proposals: as noted above, Croft (2013) provides a compelling account of language evolution that integrates the main contentions of DST; McGraw et al. (2014) also merge an interactionist approach to mind with classical cultural evolutionary theory, relying on both research streams to analyse the products of the repeated interaction experiment they study.

The shift in focus, from intrinsic capabilities of information-processing systems to the dynamical properties of the coupling of organisms with their environment, is also aligned with parts of the criticisms addressed by anthropologists to the initial versions of Californian and Parisian cultural evolution (Fuentes 2006; 2009; Ingold 1998; 2001).

"4E" revisited: the enactive approach

TODO: use chemero-after-2008 for this summary

A second avatar of the debate on the nature of information in evolution presents itself with the nature of representations in cognitive science. There are two levels to this question. First, if cognitive systems are construed as information- or representation-processing systems, the naturalisation of the content of such representations is a non-trivial matter; indeed it has attracted much attention in philosophy of mind. Second, it is not clear that information processing, and thus an account centrally based on representations, is the best description of the nature of cognition itself (as noted above, this is one of the driving questions in the debate around 4E cognitive science). More radical streams of the 4E movement contend that a notion of representation is unnecessary to account for the vast majority—if not all—of human cognition, and are developing alternative proposals (the capacity of cognitive systems to represent, if maintained in the theory, may then be seen as a contingent property and not a constitutive part). The enactive approach, in particular, proposes such an account of cognition. As described by Varela, Thompson, and Rosch (1991), Di Paolo (2005), and Thompson (2007), the enactive account builds on the notion of autopoiesis and ties cognition to living systems, considered as networks of processes that depend on each other for continued operation and continually produce and reproduce both the boundaries separating them from their environment and the conditions of their operation. Given this definition, a living system dies, that is loses its identity, if its network of self-producing processes ceases from functioning, and as such every interaction with the environment bears intrinsic value in terms of its contribution to the maintenance of the system's identity. This value is the basis for the enactive notion of meaning, which displaces the focus on information, or content, in representationalist accounts. Since living systems use their environment to self-reproduce, they are continually coupled to it in order to maintain the organisation of their network of processes. Cognition, then, is the dynamic regulation that the system operates on its coupling with the environment, and is an intrinsically meaning-making activity.

The enactive approach leads to a notion of meaning as a property emerging from the dynamics of interaction with the environment, and can be relevant at several levels of a system's organisation. This notion, and the view of cognition as a *meaning-making* activity, are also central in enactive accounts of social cognition (De Jaegher and Di Paolo 2007) and of the basic processes underlying language (Cuffari, Di Paolo, and De Jaegher 2015). The rationale for the approach is quite similar to that of developmental systems theory: talk of representations (or of biological information, in the case of

DST) easily leads to what Thompson (2007), building on Oyama (2000), calls an "informational dualism". The conceptual separation between matter and content creates a gap, and reifies information in a way that makes it difficult to naturalise.

These works are not explicitly directed towards the study of cultural evolution, and are by no means the only proposals competing within the debate on representations. Yet I will argue in Chapters 3 and 4 that CAT's reliance on mental representations as a unit of analysis renders this matter especially relevant, both theoretically and empirically, to the study of cultural evolution. The enactive treatment, among the most radical of the positive accounts because of its non-representationalist commitment, can serve as a useful point of reference, at the far end of the spectrum, in assessing approaches to information and meaning in the context of cultural evolution.

ADD: Somewhere: "The epidemiological approach renders manageable the methodological problem raised by the fact that our access to the content of representations is unavoidably interpretive. The solution to this methodological problem of ethnography is not to devise some special hermeneutics giving us access to representations belonging to a culture, yet uninstantiated in the individual heads or the physical environment of its members. The solution is merely to render more reliable our ordinary ability to understand what people like you, Opote or me say and think. This is so because, in an epidemiological explanation, the explanatory mechanisms are individual mental mechanisms and inter-individual mechanisms of communication; the representations to be taken into account are those which are constructed and transformed by these micro-mechanisms. In other words, the relevant representations are at the same concrete level as those that daily social intercourse causes us to interpret." (p. 53) So it really isn't a reductionism, but a "bridging the gaps" between approaches. "The kind of naturalism I have in mind aims at bridging gaps between the sciences. not at universal reduction. Some important generalizations are likely to be missed when causal relationships are not accounted for in terms of lower-level mechanisms. Other valuable generalizations would be lost if we paid attention to lower-level mechanisms only. If we want bridges, it is so as to be able to move both ways." (p. 98)

ADD: Somewhere: "The psychological features pertinent to determining types of cultural things may well include features of their content. Of course, content features can be characterized only interpretively. To say that various representations share a content feature amounts to saying that they can all be interpreted, at a given level and from a given point of view, in the same way. Still, that property of common interpretability, with all its vagueness, may suffice, if not to describe, then at least to pick out, a class of phenomena all affected by some identical causal factors." (p. 54)

1.2 Open problems

Moving back into the core of the cultural evolution framework put forth by Californian and Parisian cultural evolution, we can now put the focus on a number of outstanding questions for current and future research. The following discussion far from exhausts the questions to tackle, but gives nonetheless an overview of what I consider to be the most actionable or central items in the field.

1.2.1 Attraction versus source selection

The Californian and Parisian research streams are built around two complementary processes: attraction and source selection. Attraction is the umbrella phenomenon studied mostly under the Parisian approach, and is a central concept to explain cultural constancy and change when there

is no clear copying behaviour (i.e. transmission is low-fidelity): if we find constancy and gradual change in a given cultural domain in spite of interpretation, rich effects of psychology, cognitive biases, interaction, and no clear copying behaviour of agents, then cultural attraction might be a good candidate to explain its evolution. In itself, finding a cultural attractor is not an explanation, but an indication that the transmission biases are interacting in such a way that the state of culture is maintained in spite of important changes in micro-level transmission events. Source selection is the umbrella process studied mostly under the Californian approach, and is an explanatory factor of evolution in domains that feature copying behaviours (i.e. transmission is high-fidelity): when agents copy traits, or attempt to copy them, with for instance conformity, prestige, or content biases, then evolution can be usefully explored by looking at the way agents select the sources from which they copy, and the differential spread that entails. Both source selection and attraction act on multiple scales, as cultural traits or elements can be copied and transformed on several dimensions that potentially interact.

As Acerbi and Mesoudi (2015) note, attraction and source selection are not incompatible, as both are part of the overall cultural evolution process. Given their multi-level nature, both processes can also be part of one another; for instance, transformation at the level of a sentence can be analysed as selection at the level of words or concepts. Sterelny (2017) also notes that the Parisian and Californian research streams differ mostly in what they aim to explain. For the Californians, the question is how humans managed to survive and develop successful practices in the face of an opaque environment (for instance poisonous plants that, only if processed properly, can become highly nutritive). They explain this by appealing to cumulative cultural learning, which allows agents to learn from the practices of their preceding generation by copying and slightly modifying them in the process. Some level of imitation or copy is crucial to this account, precisely because the opacity and the dangerousness of the environment makes the sort of self-confident experimenting one could observe without copy extremely risky. For the Parisians, the question is why cultural traditions with no clear utilitarian value can exist, and evolve, aside from survival-related practices such as cooking techniques. They argue that the evolution of such traditions crucially involves ostensive communication (where the recipient of the communication must recognise the intention behind the communicative act), which is fidelity-neutral but not content-neutral: not all pieces of content are communicated equally, and successful communication only rarely entails copy. The Parisians thus ask why, in the face of a low-fidelity process that introduces such variation at every step, some cultural elements maintain a level of constancy and keep being transmitted through time, thus becoming evolving traditions.

If the two approaches aim to explain different features of culture, in which domain is each approach most appropriate? Are there domains that involve both imitation and non-copying communication in important degrees? How do the two processes interact in such cases? These are the questions left open by recognising a complementarity between the Parisian and Californian approaches.

1.2.2 Interaction of cultural and genetic evolution

Opinions on the importance of gene-culture co-evolution dynamics vary widely. To what extent do genetic and cultural evolution interact, and in which cases does cultural change drive genetic change? Niche construction theorists, and more broadly the Extended Evolutionary Synthesis movement, have convincingly argued for the relevance of such interactions in some specific cases, such as lactose tolerance. Developmental systems theorists further argue that genetic material and culture in its broadest sense are part of a wider matrix of resources that participate in the growth of developmental systems, and are thus necessarily interacting. Proponents of the Californian approach to cultural evolution broadly agree with the niche construction perspective, considering that the cul-

tural processes they study (concerned more with norms than with traditions *per se*) can have a long-term impact on the selective niche in which later generations develop. In particular, one view has it that the centrality of cultural transmission in human beings created a selective pressure for transmission capacities themselves (Sterelny 2017), such that cultural evolution drives genetic changes that enhance transmission. While CAT is broadly compatible with a gene-culture co-evolution account, and also discusses cases of change in context with downstream effects, Morin (2016) is not convinced that an impact of culture on genetic change is necessary to explain the emergence of global human traditions. Indeed, he argues that faithful transmission itself is neither necessary nor sufficient to explain the diffusion of global traditions, and consequently that scenarios that do not involve genetic adaptation to imitation are just as plausible given the current evidence.

This question is intimately related to the relative importance of imitation and non-copying communication. But answers might also lie in the importance and exact nature of the cognitive niche described by Sterelny (2010): how does the construction of such niches contribute to both ecological and psychological (through scaffolding of cognitive development) factors of attraction as the Parisian stream views them? How can the process of niche construction be usefully modelled to test hypotheses on the feedback dynamics between different channels of inheritance? Here too much is open to explore, which leads us to the two next points: formalisation and further development of empirical studies.

1.2.3 Framework versus formal theory contrasted with alternatives

Sperber (1996, 83) argues that CAT should not aim to become a "grand unitary theory", a position which is well informed by the diversity of domains CAT seeks to explain. It also seems in accord with the variety of ways one can approach a single domain, as the Cinderella example reminds us. Instead, CAT proposes a framework, a way of thinking that generates certain questions for the explanation of culture. Nevertheless, the approach relies on two fundamental elements that provide some degree of unification. First, Sperber insists on providing a clear ontology to the study of culture: that of public and mental representations, the latter relying on cognitive science. Second, part of CAT's argument for cultural attraction rests on the idea that much human communication is *ostensive* communication (Morin 2016, chap. 2), that is it works through the inference of communicative intentions, a process which involves a great deal of reconstruction. Relevance Theory (Sperber and Wilson 1995; Wilson and Sperber 2004), an approach that integrates well with CAT and proposes a framework to understand how agents select salient dimensions and implications of behaviours in concrete situations, is one contender for the detailed explanation of ostensive modes of communication. Given these two fundamentals, what prevents CAT from developing an abstract but systematic mapping of communicative process to stylised phenomenon, in a similar manner to the modelling endeavours of Boyd and Richerson (1985)? Sterelny (2017, 49) notes indeed that no such systematic formal models have developed in the Parisian research stream (although some models have been developed, for instance Claidière and Sperber 2007; Claidière, Scott-Phillips, and Sperber 2014).

Part of the challenge, at least, is the formalisation of both representations and the context in which they are communicated and evolve. While models need not reach this level of detail to be useful (indeed Claidière and Sperber 2007; and Claidière, Scott-Phillips, and Sperber 2014 propose higher-level models based on Evolutionary Causal Matrices), the richness of CAT lies principally with the recognition of an important role of cognition in cultural evolution, that is, of transformative and reconstructive processes. Thus, constructing stylised, simple and tractable models for CAT without emptying the approach of its main contribution is a challenge that has yet to be overcome. Classical cultural evolution, to the contrary, rests on imitative processes, such that models can avoid

restrictions on the exact nature of a cultural trait and focus on its frequency or spread in a population, without having to worry too closely about mutation. Not so for CAT: a meaningful model of cultural attraction must account for representations and their transformations, up to a degree, but in a sufficiently general way to (1) elicit effects due to the added ingredient of transformation, and (2) be applicable to different contexts. Claidière and Sperber (2007) and Claidière, Scott-Phillips, and Sperber (2014) have developed the first steps of such an approach, with discrete or continuous one-dimensional representation spaces that already show that attraction can create behaviours not accounted for by imitation-only models. The next step, however, is much higher, and is likely to involve a simple version of Relevance Theory, that is on one side an account of the multi-level and multi-dimensional aspect of representations, and on the other side a mechanism for agents to sieve through that added complexity.

While such an endeavour seems quite ambitious given the current state of modelling, it would provide a well-defined playing field to confront accounts of cultural evolution and especially the theories of cognition on which they are based. Indeed, as we have seen in Section 1.1.5, defining the cognitive factor of cultural evolution in terms of representation processing (a view shared by CAT and Relevance Theory) is not the only option on the market; one could imagine, for instance, a modification of CAT where the cognitive-level processes have been swapped out in favour of an enactive account of interactions. Indeed, that approach is also developing a description of the way organisms make sense of their world, a description which can come to replace the information-processing account of organisms selecting and inferring relevant information in their communications (in the terms of information processing approaches).

Overall, modelling CAT has not yet been pushed to its limits, and for good reasons: the challenges involved are considerable. But such a line, if successful, is extremely promising for the confrontation of approaches that genuinely compete for accounts of the interaction of cultural evolution and cognition. In the meantime however, the empirical investigation of cultural attractors is a workable task has much to bring to these theoretical questions.

1.2.4 Empirical attractors

Reliably defining and observing phenomena that qualify as cultural attractors in real life can be challenging, first and foremost because of the multi-level nature of culture, representations, and attraction. As Acerbi and Mesoudi (2015, 494) note, the definition of what counts as a cultural trait is not settled. Should one focus on mental processes, on artefacts, or on both? At what descriptive level of artefacts or representations does one consider a transformation to be a meaningful cultural change, worthy of being included as an instance of cultural attraction? Take the much debated example of Cinderella: any telling of that story will differ from other instances on dozens of features, which could all be taken to be significant in a particular context. A change in prosody or in the choice of words might alter the overall feel of the story, but not its narrative structure. A change in narrative style or structure, or an instantiation of the characters in a modern setting, might completely change the face of the story while maintaining the "persecuted heroine" aspect that tale classification systems attribute to Cinderella. Telling the tale in a particular context or to certain people only, say as an intimate bed-time story for one's child versus as a political metaphor in public discourse, can dramatically change the way it is received by its listeners. Looking for the information encoded in a version of the tale, as it were, or the representation it can elicit, only serves to postpone the problem one step further. The nature of these effects is well recognised by most writings on CAT, and one way of tackling them could be using Relevance Theory. Nevertheless, treating this diversity in an empirical study, even if it is only by classifying levels to evaluate their isolated relevance, is

very much an open question (even though Acerbi and Mesoudi 2015 note that similar multi-level indetermination of the unit of analysis is also found in the study of genetics).

As we shall see, identifying a cultural attractor in a specific domain also opens the question of the feedback loops it emerges from or participates in. Indeed an attractor is expected to have effects on the ecological context from which it arises, since it changes the distribution of representations people are exposed to. When psychological and ecological factors are involved in the emergence of the attractor, feedback effects are to be expected. How does an attractor participate to or transform its own context? In which cases is that retroaction a factor in the emergence of the attractor, versus a side-effect? For long-lived attractors, is this process a form of niche construction?

Second, as explained in Section 1.1.2, the scope of the phenomena to measure makes practical data collection also quite challenging. Existing methods can be divided into the meta-analysis of large bodies of historical and anthropological works, laboratory transmission chains or micro-societies, and the analysis of digital traces, especially from social networks. The expertise necessary to uncover reliable patterns using each of those methods makes higher integration a demanding task. Nonetheless, there is much to gain by articulating these paradigms together, individually pushing them beyond their current limitations, and connecting them to studies in cognitive science. Developing empirical methods is an integral part of any theoretical endeavour, and creating better observation tools is a sure path to push theoretical issues forward and to open previously unavailable questions to exploration.

More generally, fleshing out the predictions that attraction makes in specific empirical cases and developing methods to test those situations will help bring the ideas underpinning CAT into sharper focus. Much more empirical application is needed to evaluate the framework, and only such extensive testing in varied domains will determine how fruitful that approach is for the study of cultural evolution.

The present thesis aims to contribute to this project. In the following chapters, I will present two empirical projects aimed at testing for the presence of attractors in a particular case: the evolution of short linguistic written utterances. The practical goals throughout this work have been (1) to combine disconnected but complementary disciplines to improve on current empirical techniques, and (2) to explore dimensions in which those techniques can be pushed beyond existing limitations. In doing so, we shall also face the tension between two general approaches. On one side, *in vivo* studies use passively collected, ecological data, and must face the full complexity of reality plagued with external factors. On the other side, *in vitro* (laboratory) studies control most of the situation in which their data is generated, but are hostage to those same conditions; in particular, it is often necessary to give a task—or something much like it—to participants of a laboratory study, making any results dependent on the implications of the task. The work I present has two additional final goals: (1) contribute to the empirical evaluation of CAT as a theory of the interactions of culture, evolution, and cognition, in the form of the data points that I was able to collect; (2) highlight outstanding questions that are in need of more attention to make progress in the understanding of those interactions, thanks to a discussion of the costs handled and opportunities navigated by empirical work in the linguistic domain.

1.3 *Notes on things to add*

These two parapraphs come from the removed discussion of Ingold, but can introduce the chat about NCT/EES/DST/Evo-devo. Simplify them.

Ingold argues that Cultural Attraction Theory, as well as other approaches inpired by Darwin, relies on a description of life in three broad layers: the biological layer, which serves as the substrate for the second layer, cognitive, which itself serves as the substrate for the third layer, cultural. Each layer identifies a conceptual domain taken to be fundamentally different from the other two. The biological level is described by the modern synthesis of evolutionary theory, centred on genetic change, and studies the *body*. The cognitive level is described by cognitive science using the computational metaphor, which allows for the study of *mind* without necessarily worrying about its biological manifestation. In turn, the cultural level is described by (for instance) Cultural Attraction Theory, which studies *culture* as a distribution of representations circulating in society. The biological and cognitive layers define constraints on the way representations evolve at the cultural level (in particular through cognitive biases), but knowledge about the cultural level is not crucial to study the essential, universal parts of the mind or the body.

Each layer thus corresponds to its own object of study, implemented by the lower layers but conceptually independent from them. Ingold calls this assembly the *complementarity thesis*, referring to the way body, mind and culture are construed as separate and complementary parts of reality (Ingold 1998). The criticism he opposes to this approach can then be grouped into three essential points: (1) Each layer of the complementarity thesis is rooted in a problematic dualism that separates form and substance. (2) Without such dualisms, it is not possible to conceptually distinguish the three layers: there concrete manifestations are inextricable from one another, as they appear interlocked in a single object of study: the organism. If it is necessary to conceptually dice reality into parts or layers, it is not clear that the separation between biology, cognition and culture is the best way of doing so. (3) The notion of organism is itself problematic.

- the question about the genetic/non-genetic divide is precisely what is discussed by NCT/EES/DST/Evo-devo, and is also well discussed by Griffiths and Stotz (2013) (read intro); we take it as an open question about how to cluster the inheritance lifecycles
- debates on the cognitive side of things are beyond the scope of this intro/review, though chapter 4 will touch on them

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* (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); 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.

[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.

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.

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.

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 (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; 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 al-

[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.

phabet. 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.

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."

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

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

[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.

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

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.

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:

[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.

- **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).

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 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]

[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).

[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

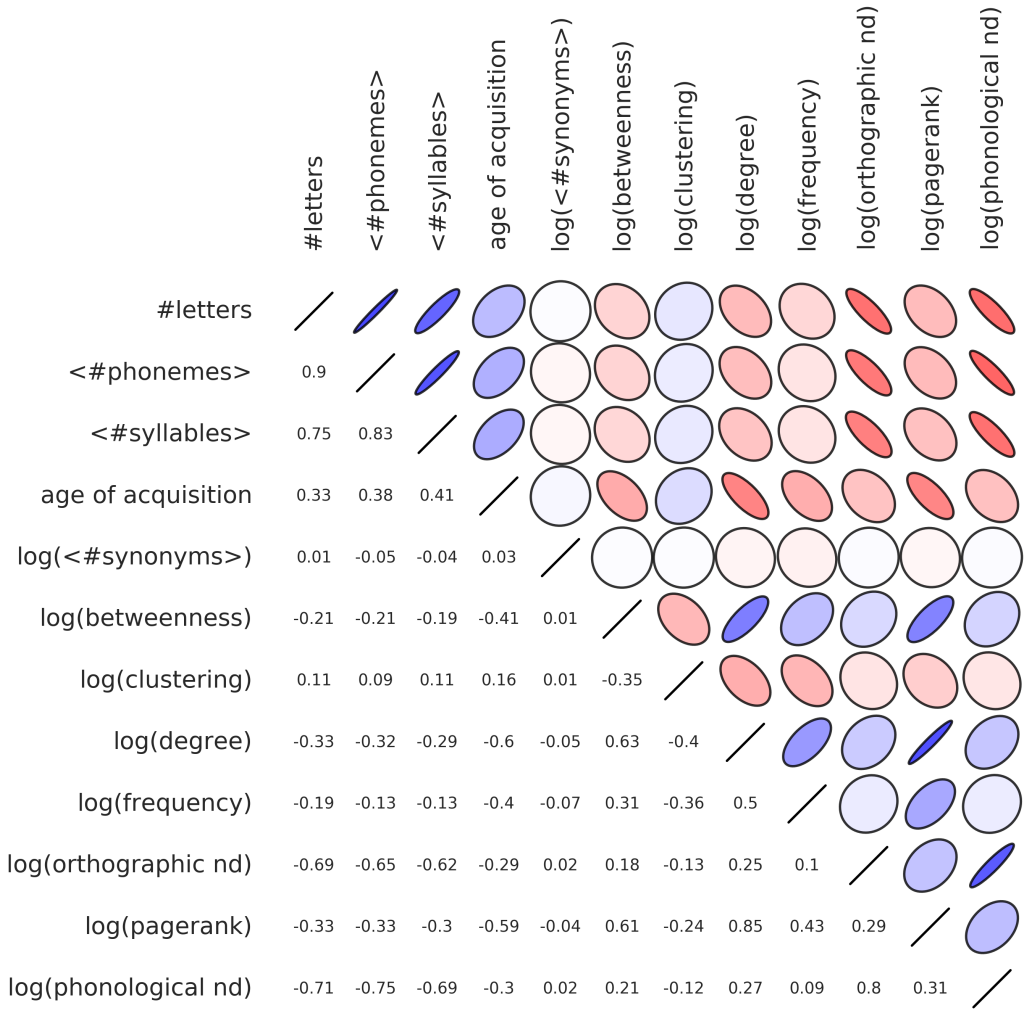


Figure 2.1: Spearman correlations in the initial set of features

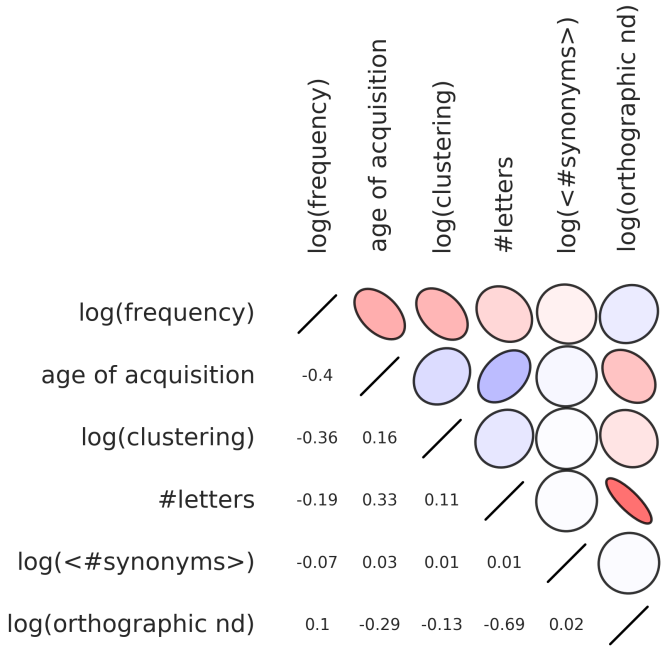


Figure 2.2: Spearman correlations in the filtered set of features

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 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.)

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.

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).

---

encoded in all features.

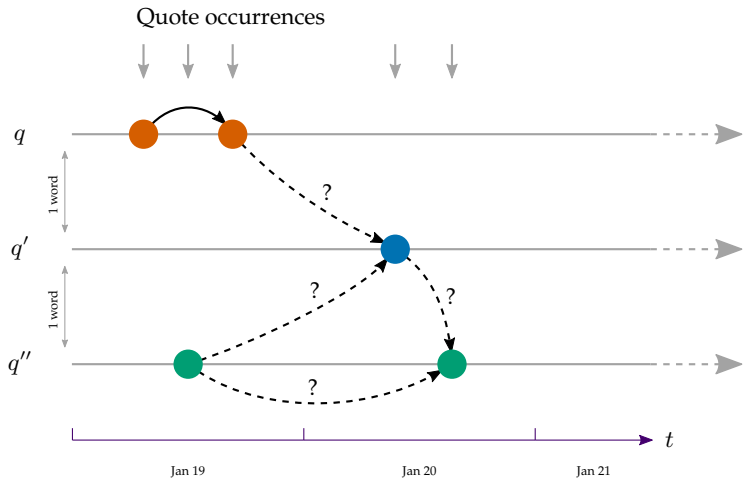


Figure 2.3: **Possible paths from occurrence to occurrence.** q, q' and q'' are three quotation variants belonging to the same cluster. q and q'' differ by two words, but q' differs from both q and q'' by one word. The second occurrence of q can safely be considered a faithful copy of the first, but 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.

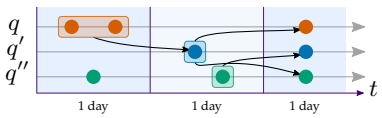
A substitution model, then, is the given of a value for each of those parameters; it considers valid all the substitutions (and only those) where the source and destination follow the rules set out by the parameters. If a destination has substitutions from multiple sources we count a single effective 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.

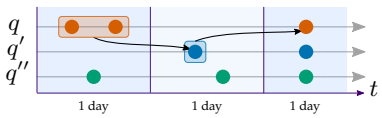
The results reported and discussed in the following sections are valid for all 16 models, and the graphics we present were produced by the model first introduced above. Finally, note that this inference procedure is one of the reasons we restricted our analysis to single-substitutions: looking for more 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

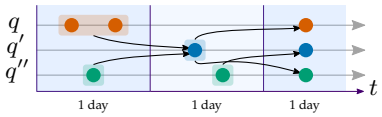
[13] We checked that this restriction does not bias the results discussed below by extending our protocol to two-substitution transformations; the results were unchanged.



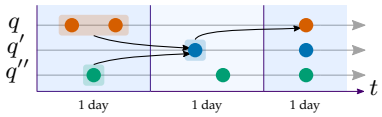
(a) Source must be majority in preceding bin, destination can be anything



(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.

out numbers, words unknown to WordNet, and deletions in substrings (which can appear as substitutions of non-deleted words); this eventually yields 6177 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:

[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

- "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, proof-reading — and may not be representative of the cognitive processes at work during alteration. The two situations are different enough to be studied separately, and we focus here on the latter.

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.

104 clusters have substitutions traveling only 1 hop, 110 traveling 2 hops, 137 traveling 3, 72 traveling 4, and 13 traveling 5.

[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.

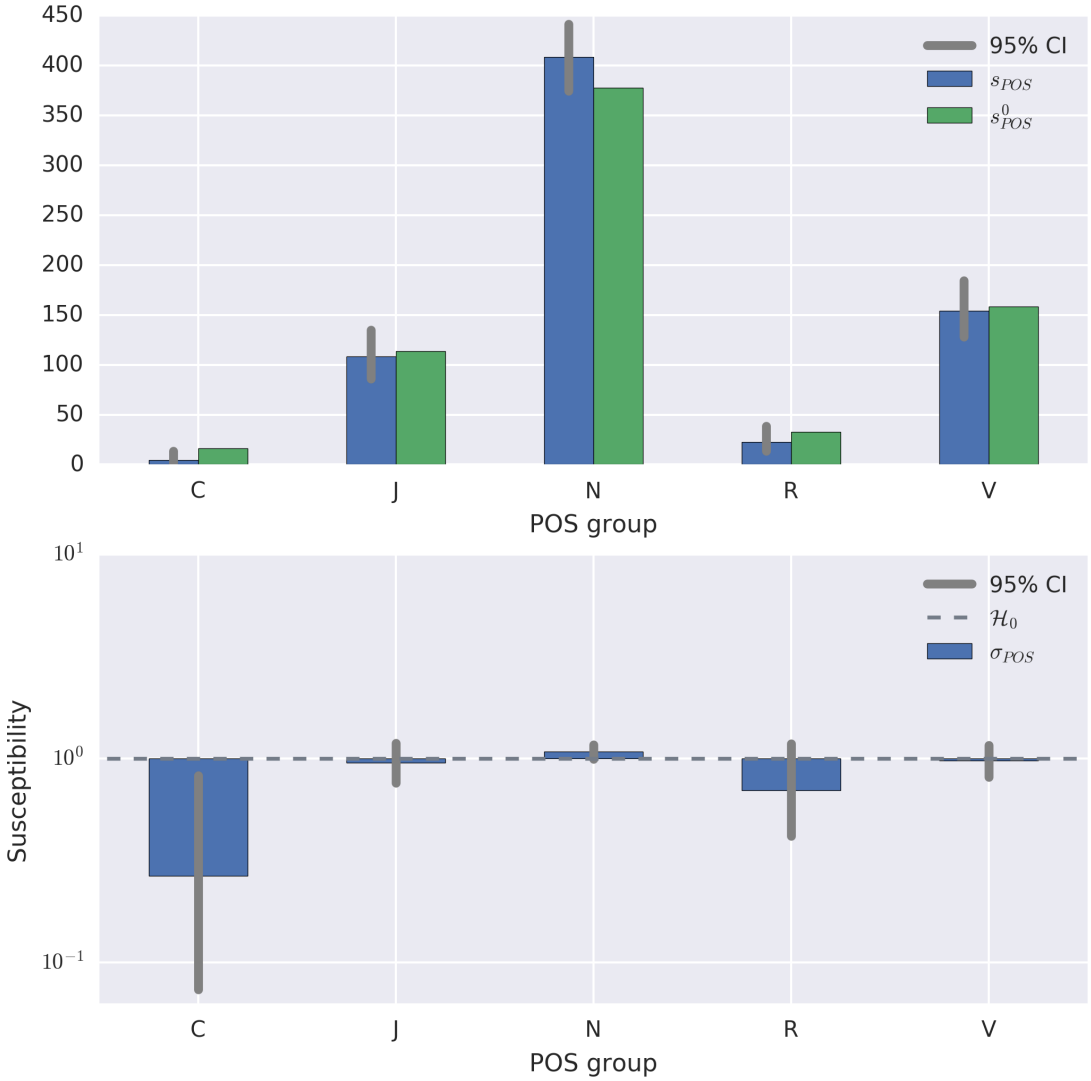


Figure 2.5: **POS-related results:** categories are simplified from the TreeTagger tag set: *C* means *Closed class-like* (see main text for details), *J* means adjective, *N* noun, *R* adverb, and *V* means verb. The top panel shows the actual s_{POS} and s^0_{POS} counts. The bottom panel shows the substitution susceptibility σ_{POS}, which is the ratio between the two previous counts. Confidence intervals are computed with the Goodman (1965) method for multinomial proportions.

Fig. 2.5 gathers the results for POS groups. A Goodman-based multinomial goodness-of-fit test (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.

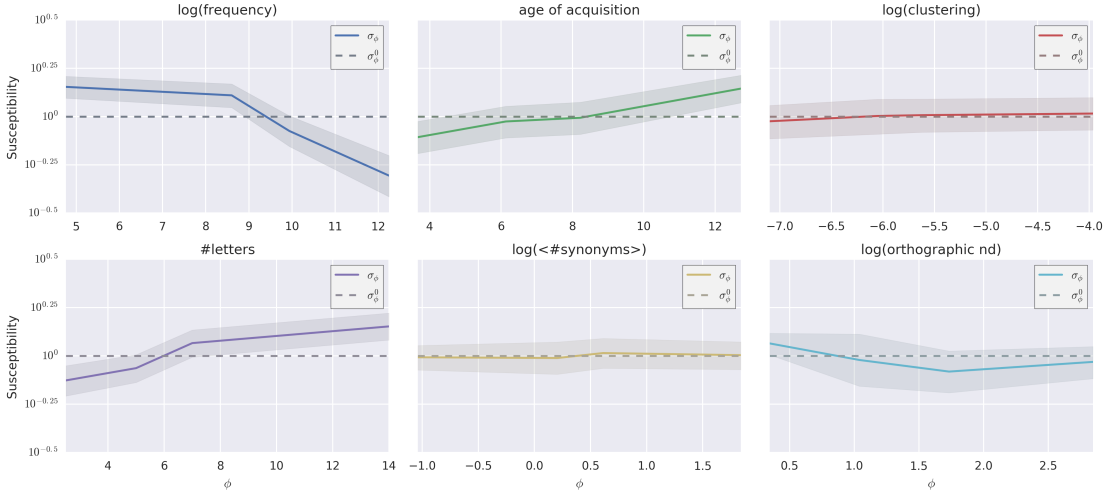


Figure 2.6: **Substitution susceptibility for feature values:** susceptibility to substitution versus feature value of a candidate word for substitution (binned by quartiles), with 95% asymptotic confidence intervals (Goodman-based multinomial).

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

[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*.

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 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 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:

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

$\mathcal{H}_{00}$ 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^{00}_ϕ does depend on f:[20]

$$\nu^{00}_\phi(f) = \left\langle \langle \phi(w') \rangle_{w' \in syn(w)} \right\rangle_{\{w | \phi(w) = f\}}$$

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^0_ϕ), and (c) an unbiased semantically plausible arrival (vs. ν^{00}_ϕ).

[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.

[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:

$$\nu^{00}_\phi(f) = \left\langle \left\langle \left\langle \phi(w') \right\rangle_{w' \in syn(m)} \right\rangle_{m \in meanings(w)} \right\rangle_{\{w | \phi(w) = f\}}$$

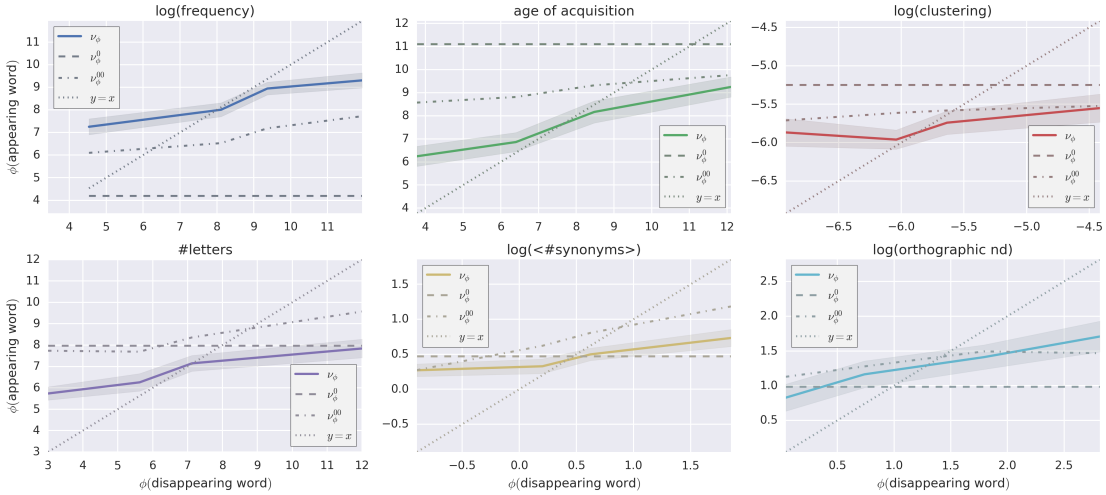


Figure 2.7: **Feature variation upon substitution:** ν_ϕ, average feature value of the appearing word as a function of the feature value of the disappearing word in a substitution (binned by quartiles), with 95% asymptotic confidence intervals based on Student's t-distribution. The overall position of 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 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.

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 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:

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

where $\boldsymbol{\phi}$ 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),

[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 point 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.

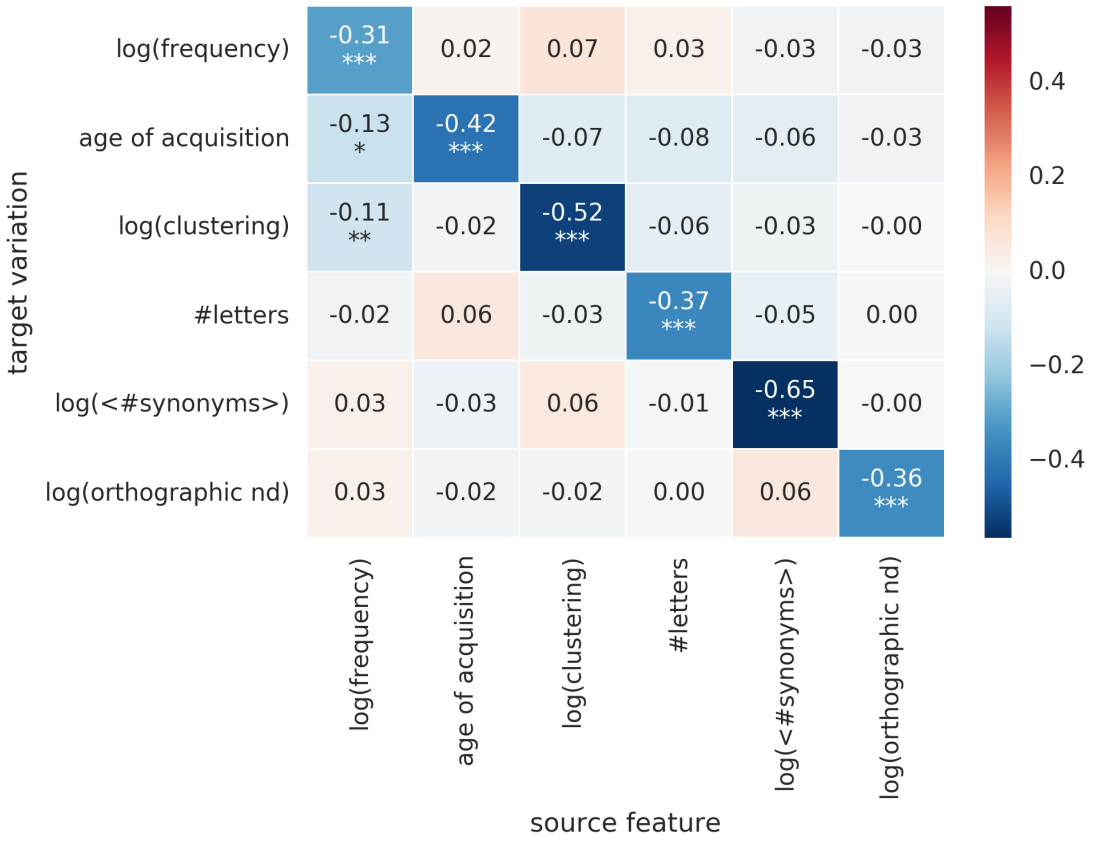


Figure 2.8: **Feature variations regression coefficients**: source feature values (columns) and feature variations (rows) were normalized to $[0; 1]$ to ensure the coefficients are comparable. Significance levels for individual t-tests against the hypothesis of a null coefficient are denoted by stars 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.

"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

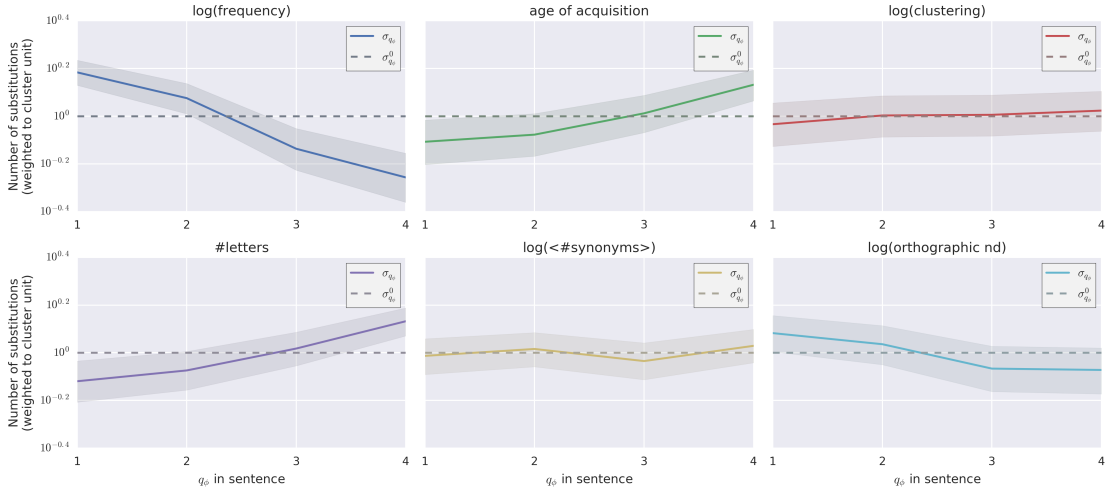


Figure 2.9: **Substitution susceptibility for in-quote feature quartiles:** susceptibility to substitution versus quartile of the feature distribution in the originating quote, with 95% asymptotic confidence intervals (Goodman-based multinomial).

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 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 $s^0_{q_i}$, 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

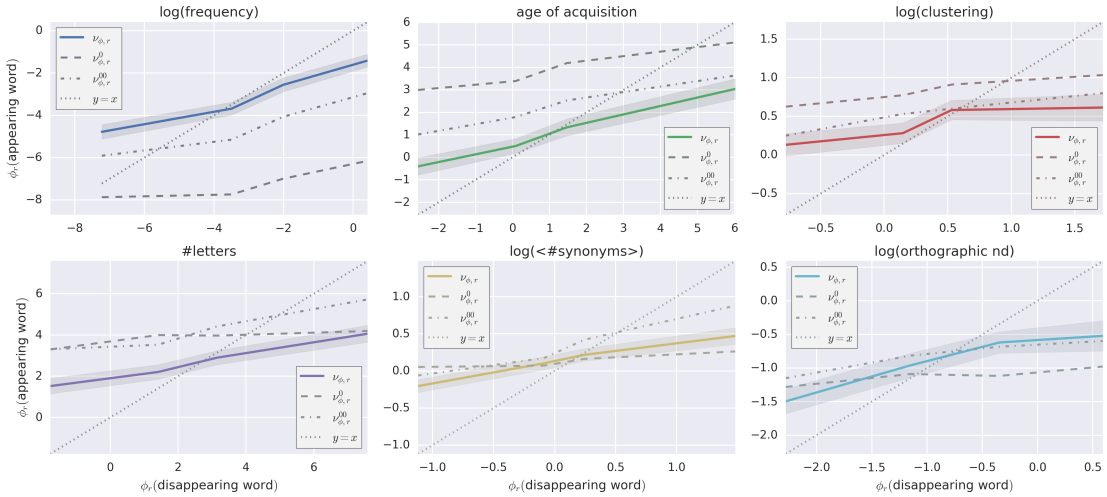


Figure 2.10: **Sentence-relative feature variation:** ν_{ϕ_r}, average sentence-relative feature value of the appearing word as a function of the sentence-relative value of the disappearing word (binned by quartiles), with 95% asymptotic confidence intervals based on Student's t-distribution. $\nu^0_{\phi_r}$ and $\nu^{00}_{\phi_r}$ 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 $\nu^0_{\phi_r}$ and $\nu^{00}_{\phi_r}$ 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.

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^0_ϕ and ν^{00}_ϕ each transpose to ν_{ϕ_r}, $\nu^0_{\phi_r}$ and $\nu^{00}_{\phi_r}$ (note that $\nu^0_{\phi_r}$ now depends on w since it is sentence-relative, whereas ν^0_ϕ 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 $\nu^0_{\phi_r}$ and $\nu^{00}_{\phi_r}$ on the other side (for Number of synonyms it is a triple intersection with $\nu^0_{\phi_r}$; for Clustering coefficient and Orthographic neighborhood density, a triple intersection with $\nu^{00}_{\phi_r}$). 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

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

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

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 (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

3.1 Introduction

The previous chapter demonstrated that it is possible to observe cognitive biases in the way quotations are copied from blog to blog. By grounding those biases in known effects in the recall of word lists, we also showed that the enquiry of cultural evolution for linguistic content can be related to lower-level cognitive mechanisms that help understand the way content is transformed in ecological situations. However in the online corpus we considered only extremely simple transformations, namely individual word replacements, so as to make it possible to infer missing links between quotations, thus make the analysis possible. While we observed a reliable bias in the way words are replaced, consistent with known psycholinguistic biases, our view of the overall transformations is extremely narrow: not only is it restricted to word replacements, it is limited to the replacements that were the only change in an utterance (other than sentence cropping). The analysis also remained at the low-level of lexical properties such as word frequency and age of acquisition, neither of which give much insight into the semantic changes that utterances can suffer. Finally, constraints of the data did not let us identify chains of transformations, and we could not observe the evolution of quotations beyond the individual transformation step.

We now wish to remedy most of these points by studying the evolution of short utterances in a controlled experimental setting. A controlled setting means having a less ecological situation, but also allows for the collection of all the available data for analysis. Once again, our approach is exploratory, and we aim to reach a more complete understanding of the transformation process that is at work in the propagation of online quotations, but also more widely in the evolution of linguistic content as it is transmitted in society. More precisely, our goal is to construct a descriptive model of the process that can bring insight into why utterances change the way they do, and how such observations can be connected to current knowledge in linguistics, on one side, and to the broader cultural evolution frameworks, on the other. If we succeed in creating such a descriptive model, explaining the details of the process with lower-level cognitive mechanisms should then be easier.

The ideal setup to tackle this question would be a standard transmission chain where we observe the accumulated transformations of subjects on a set of utterances that we choose. However, given our exploratory approach and the fact that we do not know in advance what the model will look like, it is important that we can run several such experiments in short cycles so as to adjust the task parameters and the sampling of utterances. It is also important that the collected data be of similar size to the

number of substitutions we extracted in the previous chapter, so that we will be able to compare and validate any overlapping results. We thus chose to run a set of transmission chain experiments on an online platform developed for the purpose: as we shall see, after an initial development phase this approach lets us collect large amounts of data in short periods of time, while maintaining a level of control similar to that of laboratory experiments.

We begin by discussing the works relevant to this endeavour, and in particular the bind in which current transmission chain experiments on linguistic content find themselves. We then present the procedure followed to develop the online experimental platform, and the measures implemented to achieve a high level of quality in the data. Next, we expose our analysis of the data sets collected, expose the descriptive model of transformations we construct from them, and highlight the main behaviours the model lets us observe. Finally, we discuss the relevance of these results in the broader context of the study of cultural evolution.

3.2 Related work

Inspired by the selectionist models of culture developed by Boyd and Richerson (1985) and Cavalli-Sforza and Feldman (1981), a sizeable part of the empirical work on cultural change has focused on identifying content and context biases in the way cultural items are transmitted. This line of work relies heavily on the transmission chain paradigm initially introduced by Bartlett (1995). For linguistic content in particular, studies using that paradigm now provide a catalogue of contrasts in the way utterances or short stories are transmitted. These effects range from the stereotypical personification of objects (Bangerter 2000), the favouring of negative story aspects (Bebbington et al. 2017) or the increased hierarchical encoding of events (Mesoudi and Whiten 2004), to biases in favour of social (Mesoudi, Whiten, and Dunbar 2006) or counter-intuitive aspects of stories (Norenzayan et al. 2006; Barrett and Nyhof 2001). Other effects such as the role of emotions in the selection of items to reproduce (Heath, Bell, and Sternberg 2001; Eriksson and Coultas 2014), or conformity and prestige biases (Acerbi and Tehrani 2017) have been studied by focusing on the individual transmission step on which the evolution of content hinges.

Often, such effects are identified by selecting two or more minimally different types of content and contrasting the way they evolve in transmission chains (for instance measuring the rate at which they are degraded). When a type of content is significantly better transmitted than other types, it signals that a bias is acting on that contrast dimension. The technique is useful in the context of selectionist models of culture, as it identifies examples of biases which could create selection pressures for specific cultural types and thus drive cultural evolution. It is also relevant to the Cultural Attraction framework, which focuses on the aspects of culture for which reconstructive processes are more important than selection. For instance, the approach introduced more recently by Claidière, Scott-Phillips, and Sperber (2014) proposes to use evolutionary causal matrices to model such attraction-based processes in cultural evolution, and could gain insight from the trends observed in transmission chains. In the terminology of Morin (2016), selectionist models focus on how culture survives in spite of wear-and-tear, and cultural attraction focuses on how culture survives in spite of possible flops, where a given item fails to elicit sufficient interest to be recreated at all. In theory, both these processes can be observed in transmission chains. However, in its current implementation focused on contrasting outcomes, the technique gives little insight into the underlying mechanisms at work, their commonalities and differences, what they depend on, and how exactly they can be explained in terms of cognitive and situated processing.

Indeed, understanding the mechanisms behind transformations in chains, or even only quantitat-

ively describing the details of said transformations, remains very much a challenge. This is especially true in the linguistic domain, where the complexity of language hinders most attempts to understand what is going on in a transformation. Up to now three main strategies have been developed to delve into to detail of transformations. The first is to use tightly constrained linguistic content, for instance sentences of a very specific type, or for which only pre-defined changes can happen. In that case the transformations can be directly modelled to identify regularities. The study of the recall of word lists (see Zaromb et al. 2006 for a review), and that of word replacements in short sentences (Potter and Lombardi 1990; Lombardi and Potter 1992), can be seen as employing that strategy: word lists and individual replacements in sentences are much simpler than transformations of complete sentences, and are thus more amenable to statistical analysis. Our analysis of word substitutions in the previous chapter can be categorised here too. A similar strategy is found in non-linguistic studies, such as iterated learning on sequences of colour items for which standard regularity metrics exist (Cornish, Smith, and Kirby 2013), or transmission chains of constrained visual patterns such as those used by Claidière et al. (2014). Both cases feature discrete and combinatorial pieces of content, for which it is possible to use natural notions of distance, equality, or regularity in transformation. This first strategy can be termed the "simple setting" strategy.

At the other end of the spectrum we find the "do-it-by-hand" strategy. This approach uses more ecological content but relies on exhaustively hand-coding it, and is used in most of the transmission chain studies mentioned above. The study of risk perception propagation developed by Moussaïd, Brighton, and Gaissmaier (2015), for instance, used a free-form interaction setting where subjects were taped while freely discussing a topic. The recorded conversations were later hand-coded for the presence of certain information items introduced at the beginning of the chains. The linguistic analysis of transformations of quotes in news stories provided by Lauf, Valette, and Khouas (2013) is also the product of exhaustively hand-coding differences between sentences.

Finally, the third strategy relies on pre-labelled data sets, often from online platforms, on which machine learning techniques can extract features that correlate to the transmission of pieces of content. This is the "already-coded" strategy. Danescu-Niculescu-Mizil et al. (2012), for instance, study the memorability of movie quotes by exploiting user ratings provided on the Internet Movie Database website. As we saw in the previous chapter, analysing the regularities that arise in unlabelled digital traces falls back into the first strategy, as having to infer missing information led to drastically simplifying the transformations considered.

Strategies two and three are additionally closely tied to data collection methods. Free-form interaction and more generally ecological content is costly to hand-code, and thus necessarily limited in size; it is also best used in controlled settings where the choice of content can be optimised. Conversely, using machine learning to extract features that relate to content transmission requires large amounts of pre-labelled data, which often means that an existing public data set must be used. Such studies thus seldom control the conditions under which the data is generated, which restricts the interactions they can explore to those encoded in existing data sets: any behaviour or piece of content that is not present in public data sets is off limits.

Overall, studies targeted at understanding the details of transformations of linguistic content seem forced to pick two of the following three properties, and relinquish the third: realistic content, computational analysis, and control over the generation of the data. Picking realistic content and computational analysis leads to the "already-coded" strategy. Picking realistic content and control over data-generation requires hand-coding a substantial part of the data collected, that is strategy two. Finally, computational analysis and data-generation control leads to the "simple setting" strategy. This bind thus appears as a major challenge to the better understanding of changes in linguistic content, and more broadly to the study of language-related cultural evolution. In particular, it hinders

attempts to model the low-level processes which could provide a more parsimonious account of the contrasts observed in linguistic transmission chains, and allow for a deeper integration of the study of cultural evolution with linguistics.

To overcome this obstacle we turn to two related fields of research. The first, which we term the Web and Smartphone experimental approach, is creating a middle ground between controlled laboratory experiments and the analysis of online corpora. This approach takes advantage of the ubiquity of internet browsers and mobile computing to develop large-scale controlled experiments out of the laboratory. Miller (2012) discusses the possibilities opened by developing experiments as smartphone applications in particular, and notes that this method changes the logistics and context-awareness of experiments: large amounts of subjects can be recruited online without having to manage meeting schedules, and experiments can probe participants without interrupting their everyday life, both advantages that been exploited in the study of mind-wandering and happiness (Killingsworth and Gilbert 2010; Mackerron and Mourato 2013; Bastian et al. 2017). A closely related method is the development of experiments as web applications, which similarly changes the set of experimental constraints. In linguistics, the possibility for large-scale data collection has been successfully used in the study of vocabulary size (Keuleers et al. 2015; Brysbaert et al. 2016); creating studies that involve many subjects at the same time is also made much simpler by the online logistics, an advantage that has been used for instance in the study of group conversations (Niculae and Danescu-Niculescu-Mizil 2016). More generally, these approaches relax the opposition between small-scale controlled experiments in the laboratory on one side, and analyses of large-scale but passively collected online data on the other side. Once the initial development cost is covered, they make it possible to collect relatively large data sets in short cycles, and combine simplified logistics with a level of control similar to that of laboratory experiments.

The second field we rely on creates an opening for the detailed modelling of utterance transformations: biological sequence alignment, the sub-field of bioinformatics which seeks to uncover commonalities in sequences of DNA, RNA, or amino acids in proteins from different species, has developed over the last 50 years a range of general algorithms to relate sequences of items. One such algorithm in particular, introduced by Needleman and Wunsch (1970), extends the principles of the Levenshtein distance and is particularly well suited to the analysis of linguistic transformations when combined with standard natural language processing methods. Inspired by Lauf, Valette, and Khouas (2013) who use similar tools to prepare their data for manual analysis, we use and extend the Needleman-Wunsch algorithm to reliably extract regularities in the way utterances are transformed through transmission chains.

3.3 Methods

3.3.1 Experiment design principles

Advantages and challenges of transmission chains

An obvious way to address the questions raised in the previous chapter is to use transmission chains in the laboratory to study the evolution of online quotations in a controlled setting: each subject reads, retains, and rewrites sentences that are then passed on to the next subject in a chain of reformulations. Such a setup can reproduce an idealised version of the read-remember-rewrite process which, we hypothesised, participates in the evolution of quotations in blogspace and media outlets. It also provides the information that our previous data set lacked in order to analyse the complete

transformations of quotations, as well as the long-term effect of those changes: the links between parent and child sentences are naturally encoded in the data, such that the transformations undergone by each sentence can be studied in full detail. There is no need to restrict ourselves to simpler changes as was necessary for the inference procedure used with digital traces from blogspace. By creating an artificial setting, the experiment design also lets us control the reading and writing conditions as well as the context in which sentences appear, which further removes one of the inevitable uncertainties of the previous protocol (albeit at the cost of less ecological conditions).

However, the laboratory transmission chain paradigm is not a good fit for our exploratory approach: we aim to collect data that will allow us to study both the complete set of transformations undergone by short utterances such as online quotations, and the interactions and cumulative effect of such changes; yet we do not know in advance the types of changes that subjects will make, or the extent to which such changes vary according to the type of linguistic content. Transmission chain studies typically start with an a priori hypothesis focused on a well-identified type of content, which is then empirically tested by contrasting the evolutionary outcome of two classes of sentences. Instead, our goal here is to provide first steps to characterise the process by which such evolution of linguistic content arises, and observe how it accumulates in the long term. The setup must thus allow us to collect enough data to extract regularities in successive transformations operated by different subjects on different sentences, and provide a resolving power similar to that of substitutions in online quotations so that we can compare results with the previous chapter. Since our main target is the set of detailed transformations and their interactions, a phenomenon of higher dimensionality than the contrast of accumulated outcomes, it is also crucial to fine-tune the difficulty of the read-write task and the complexity of the source sentences, in order to trigger a set of transformations varied enough that it could approach some of the changes encountered in real life situations. Our progress therefore involves a non-trivial trial-and-error component: indeed, a task made too easy or too difficult, and more so a set of source sentences that are too complex or too straightforward, will lead to either mass deletions or perfect conservation (or the former followed by the latter), none of which can help characterise the more intricate changes that linguistic content undergoes in the ecological setting we aim to simulate.

Web and smartphone experiments

Complementary to laboratory studies and to approaches using online digital traces, a new empirical approach based on Web browsers and mobile computing is striking a different balance in the trade-offs of experimental work; it seems very promising in addressing the problems outlined above. Indeed, browsers (both on desktop and mobile) and smartphones have evolved into powerful, ubiquitous application environments for which one can develop any kind of experiment involving text, graphics, and human interactions. At the cost of increased engineering requirements and a different approach to subject recruitment, Web and smartphone experiments give the designer full control over what data is collected and the way interactions are framed (similar to laboratory experiments), and make it possible to quickly collect data sets at scales comparable to what filtered and cleaned digital traces provide.

This approach makes a number of unusual trade-offs, the benefits of which can be summarised as follows:

- *Control*: similar to laboratory experiments, and unlike digital trace analysis, it is possible to use complex designs where all the interactions of the subjects are framed and observed by the experimenter. This includes for instance the presentation of the experiment (e.g. as a game or

a self-improvement aid, aside from being a scientific study) and, more importantly, the ways in which the system mediates the interactions between the subjects.
- *Scale*: if and when needed, the technical platform can scale the number of subjects to the tens of thousands at low marginal cost. Interactions between subjects can also scale to involve synchronous or asynchronous contact between hundreds of people, without having to manage per-subject scheduling.
- *Speed of data collection*: once the initial development is completed (see costs below), the data collection cycle is short. One day can be enough to collect 1000–10,000 usable data points, a size comparable to the final substitutions set extracted and analysed in the previous chapter. This is especially relevant for exploratory work which is made much easier with shorter trial-and-error cycles.
- *Flexible recruitment*: while also a challenge (see costs below), subject recruitment is more flexible than in the laboratory: services like Prolific Academic[1] let the experimenter recruit at reasonable costs in pools of tens of thousands of subjects with fine-grained demographic filters. Wider audiences can be achieved by offering non-financial rewards, framing the experiment as a self-improvement application, or turning it into a game.

The corresponding costs are the following:

- *Technical challenge*: developing Web and smartphone experiments involves a substantial amount of engineering, and makes use of technologies that most researchers, even technical, are not familiar with. While a couple of all-in-one kits exist,[2] creating an experiment that meets one's research questions requires learning average skills in most of the technologies at play: a native or cross-platform smartphone development environment, Web application development, backend server programming, and some server administration skills. Most importantly, the paradigms and problems encountered are new to researchers: control flow is asynchronous due to network communication and the user interface, and technicalities such as user management or email validation can grow into difficult engineering challenges.
- *Spam-control*: subjects are not constrained or encouraged by the face-to-face interaction of a laboratory experiment, neither are they (in most experiments) in the course of an interaction with friends that provides natural incentives for what they write, as can be the case with digital traces. Participants must have an incentive to perform the experiment's tasks well. If the spam introduced by one subject can be isolated in the design of the experiment, one possibility is to filter it once the data is collected and make payment depend on its prevalence. But if the spam introduced by one subject naturally propagates to data seen by other subjects in the experiment, as is the case for transmission chains, effective anti-spam pressures and motivations need to be factored into the design.
- *Recruitment cost*: while recruiting up to a few hundred subjects is cheaper than the equivalent for a laboratory experiment (not counting the development cost),[3] and is easy to manage for fast prototyping and pilot tests, recruitment cost rises linearly with the number of subjects and the time they spend on the experiment, unless a different strategy is used. Turning an experiment into a playful application or an application useful to the subjects (effectively making them users) involves yet another set of skills, can prove challenging, and must be factored into the development cost.

[1] https://www.prolific.ac/.
[2] See e.g. http://funf.org/ and http://www.epicollect.net/.
[3] Global competition on online platforms like Prolific Academic drives subject payments down.

General setup for Web-based transmission chains

The balance achieved by Web-based experiments is well adapted to the requirements we outlined above. Since no existing system would fit our needs, we chose to develop a tailored Web-based platform that could run transmission chains as Web experiments. Once ready, the platform would allow us to gather sufficient amounts of quality data in short cycles. We further decided to implement the simplest possible version of the transmission chain paradigm that is still viable, and leave the exploration of more complex setups for future research: the task we used asks subjects to read and memorise a short utterance, wait a few seconds, then rewrite what they have read as accurately as possible. We ran three main experiments using this evolving platform, and many smaller pilots in between to test lessons learned in the larger runs and adjust task parameters and source complexity. The overall quality of the data we gathered thus gradually increased. In what follows we present the general setup of the experiments, the data quality evaluation along with the changes implemented to improve it, and finally the adjustment of task complexity. Let us start with the architecture common to all experiments.

A transmission chain is defined by a type of content transmitted, a transmission task, and a layout defining which subject production is used as the source input for the next subject. In our implementation, subjects are presented with an utterance to memorise with no surrounding context; no distraction task is used between the reading and writing phases, and the material incentive for the task is purely monetary (although as we describe below we fine-tuned the interface to strongly encourage subjects to be conscientious). The experiment is available to subjects as a website, and passing it involves the following steps:

- Welcome and sign up (Figs. 3.1a and 3.1b),
- Answering a preliminary questionnaire (Fig. 3.2),[4]
- Training for the main task, where subjects are asked to repeatedly memorise and rewrite short utterances as accurately as possible. As the instructions illustrated in Fig. 3.3 indicate, an utterance is presented to the subject and after a short pause they are asked to rewrite it as remembered. The process loops until the subject has completed all the utterances assigned to them (calibrated so that completing the experiment lasts at most one hour). The real trials started after 3 to 5 training trials, depending on the overall experiment length.

This simple setup lets us quickly gather data sets of several thousand utterance transformations, ensuring our results were be comparable to those from the set of 6177 substitutions extracted in the previous chapter. Two parameters are then left to vary: the reading time for the source utterances, computed as the number of words in an utterance multiplied by a reading factor that is to be adjusted, and the set of initial source utterances.

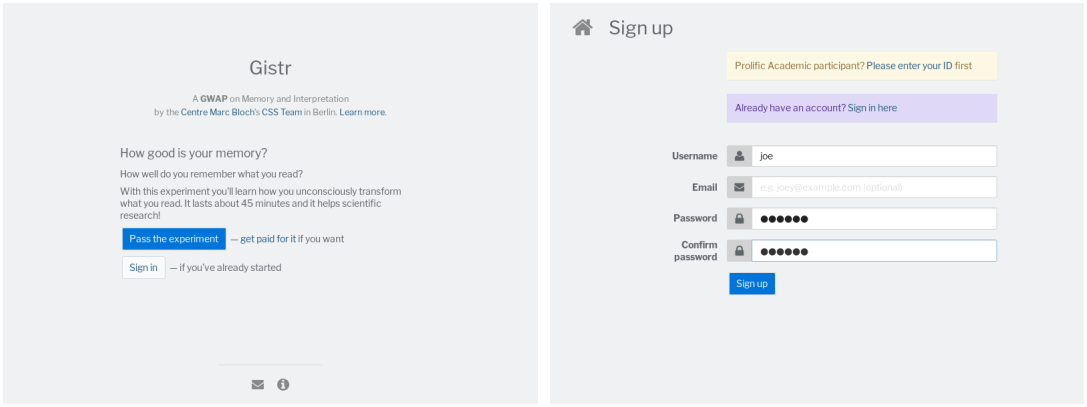
Each utterance from the initial set is used to create several parallel chains in order to allow for comparisons across chains with the same initial utterance. The final data thus consists in a set of reformulation trees, where each tree branch is a transmission chain started from the tree root, and continuing until it reaches a target depth defined for the experiment.[5] The number of branches in a tree is also adjusted for each run of the experiment. Except for those who drop out before finishing the experiment, all subjects are exposed exactly once to each tree in random order, such that all the reformulations in a given tree are made by distinct subjects, and nearly all subjects (excluding dro-

[4] An early version of the experiment also included a word span test at this stage. However, similarly to the age of subjects that we collect in the questionnaire, this data turned out to not be relevant in the analyses. The magnitude of transformations depends far more on the conscientiousness of subjects, and this non-trivial test was later removed during one of the frontend rewrites.

[5] We therefore use the terms “chain” and “branch” interchangeably in what follows.

pouts) are present in each tree. Satisfying this constraint means that we must always have at least as many subjects as there are reformulations in a tree. As noted, a few subjects from Prolific Academic will usually not complete the whole set of utterances assigned to them; we thus recruit additional subjects to fill the trees that were left incomplete. This leads the other, already complete trees to receive more reformulations than needed, making some of their branches run a little deeper than the target depth. All branches are cropped to the target depth for analysis.

Finally, note that when exposed to a tree, subjects are always randomly assigned to the tip of one of the branches that have not yet reached the target depth: subjects are thus randomly distributed across branches, but their depth-ordering loosely corresponds to the time of arrival on the tree. In particular, if a subject starts the experiment after most other subjects have completed it, he or she will be mostly exposed to utterances deep in the branches. Due to the chained nature of the data, there is no economical way of countering this ordering bias.[6] Fig. 3.4 shows a representation of the shape of the final trees.



(a) Welcome screen (b) Signup form

Figure 3.1: Initial steps for a subject entering the experiment.

Technically, the platform is a complete Web application based on current technologies, with accompanying backend server to collect and distribute utterances.[7] The experiment is available at https://gistr.io and subject recruitment was done using Prolific Academic, a service analogous to

[6] The following three approaches could be combined to counter ordering bias. (1) Have each subject do a single trial, that is, use as many subjects as there are reformulations in the full experiment; this is extremely expensive as there is a fixed minimal price for each subject, corresponding to the time needed to explore the interface, answer the initial questionnaire, and train for the main task. (2) Have each subject wait an adjustable amount of time between each trial, to open the possibility for ordering subjects differently than their time of arrival; this is also expensive, as it means paying subjects for waiting most of the time they spend on the experiment. (3) Optimise the order of tree presentations of each subject so as to spread subjects across depths; while this approach could achieve some level of spread when combined with (2), it is contingent on the starting times of subjects and their synchronisation, which we do not control (subjects find the experiment through Prolific Academic notifications and are free to start whenever they want).

[7] The frontend first used the Ember.js framework (Ember.js contributors 2017), and was later rewritten and extended using the Elm programming language (Czaplicki and Elm contributors 2017). Indeed, the assurance of no runtime exceptions that Elm provides was a strong argument in favour of switching, as was made clear by the trying "customer support" experience of a bug hitting 40 to 50 subjects at once during Experiment 1. The backend is a Python application written on top of the Django REST framework (Christie and Django REST framework contributors 2017). Most of the critical logic in the software is verified using automated tests, and the full source code is available under a Free Software licence at https://github.com/interpretation-experiment/gistr-app (frontend), and https://github.com/interpretation-experiment/spreadr (backend).

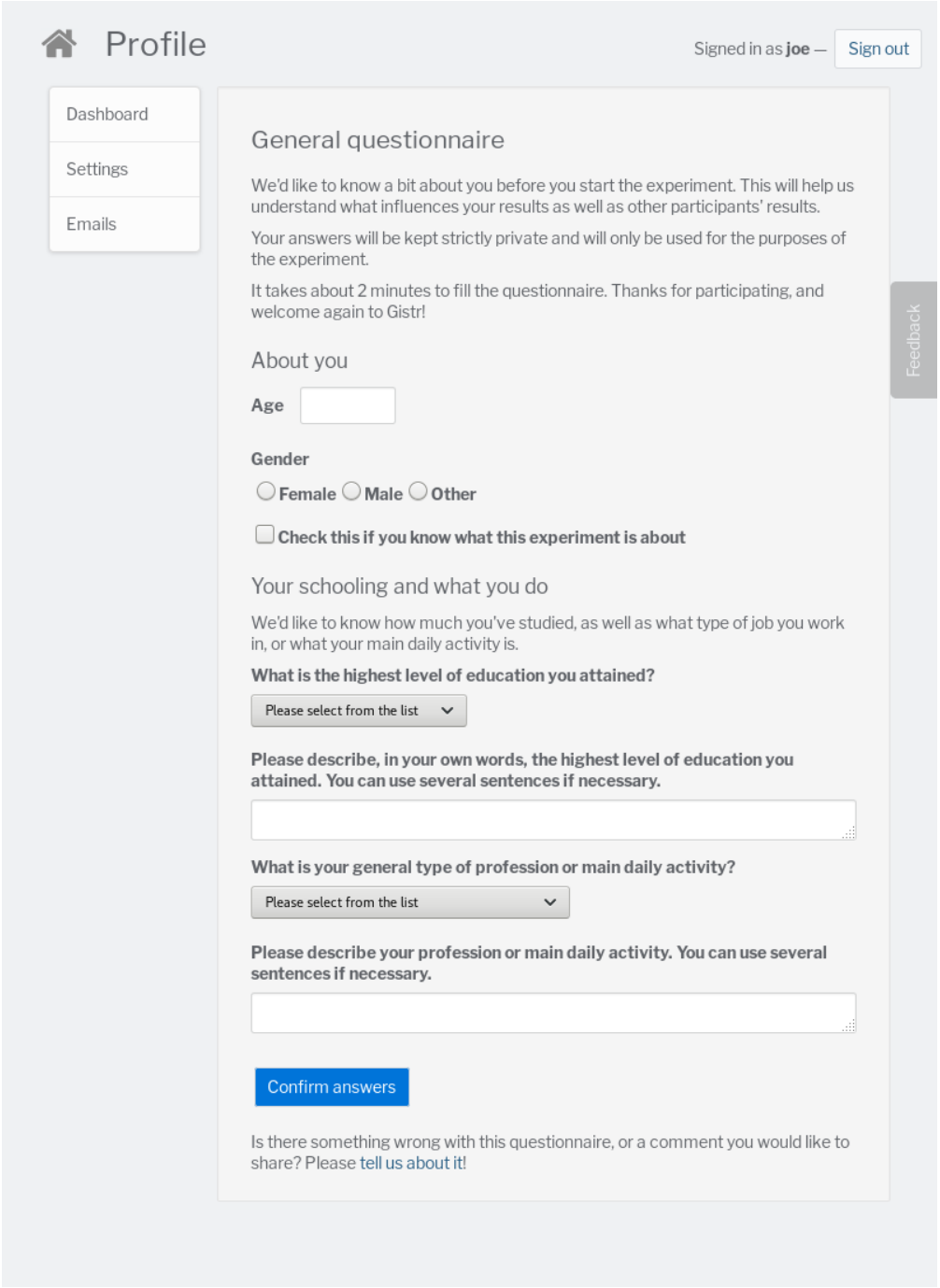


Figure 3.2: Initial questionnaire. Subjects can additionally submit feedback on the questionnaire or any other aspect of the experiment on most screens of the website.

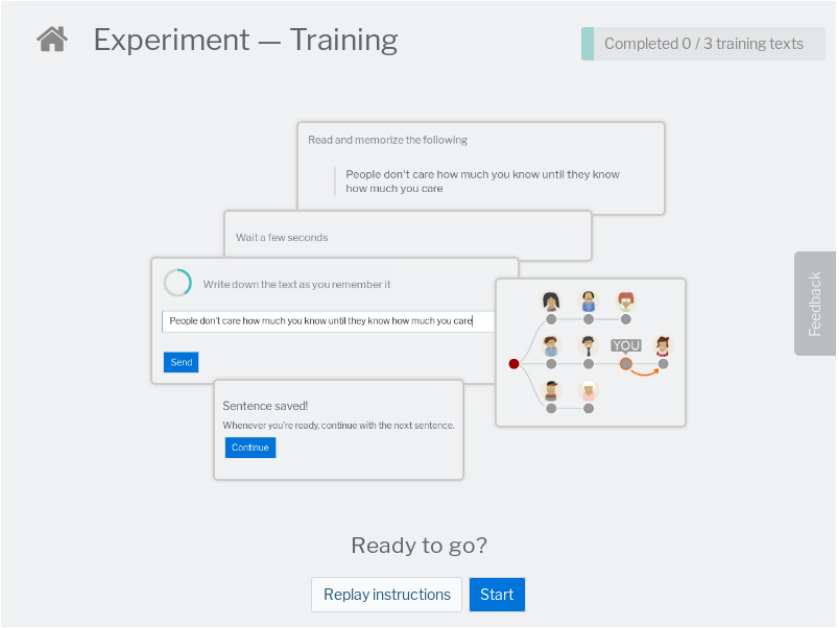


Figure 3.3: Instructions for the main task

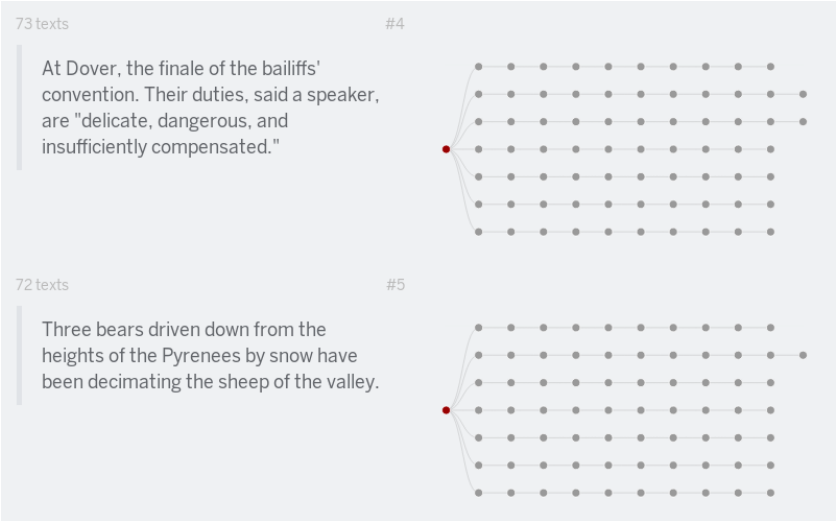


Figure 3.4: Two example reformulation trees generated by the setup, targeted at 7 branches of depth 10: the text on the left is the initial utterance for all branches of a tree, represented by a red dot in the right-hand graph; each grey dot in the graph represents an utterance produced by a subject on the basis of the preceding dot. Subjects create at most one reformulation in each tree, and most create exactly one per tree. The fact that some subjects drop out before completing all their trees leads us to recruit new subjects to fill in the missing reformulations, which is why some branches are uneven.

Amazon Mechanical Turk and geared towards academic research.[8]

Using the Prolific Academic service allowed us to select among a pool of over $26{,}000$ subjects, for which we used the following criteria:

- First language English speaker
- At least 18 years old
- Current country of residence and place of most time spent before turning 18 must both be in the UK
- Normal or corrected-to-normal vision
- No diagnosed literary difficulties
- Completed secondary school
- Not having participated in any of the preceding experiments

Only the first two constraints were enforced for the first experiment, and the full set was used for all subsequent runs. The full filter provided over 2300 eligible subjects, from which the service automatically sampled the number of subjects we requested.

Experiment 1 was the first non-trivial launch of the platform, with an initial 48 subjects, 54 root utterances, and trees targeted for 6 branches of depth 8. Subjects took an average 64 minutes to complete the experiment, and were rewarded with £6.5. A software bug that appeared and had to be fixed halfway through the experiment led the Prolific Academic service to recruit more subjects than was originally asked for, and the final number of participants was 53,[9] gathering a total 2695 utterance reformulations (above the planned $54 \times 6 \times 8 = 2592$ reformulations). Manual inspection showed that large portions of the data were of poor quality, both linguistically and because of technical inefficiencies leading to badly shaped trees; the sections below provide further details on these questions. Pilots following Experiment 1 were therefore aimed at improving data quality and solving tree shaping issues. Experiment 2 was launched with 49 subjects, 50 root utterances, and trees targeted for 7 branches of depth 7, gathering a total 2450 utterance transformations. Subjects took an average 43 minutes to complete the experiment, and were rewarded with £6. Quality issues in this data set were solved, but the choice of source utterances proved too easy to trigger varied transformations. After pilots exploring different fits of task parameters with source complexity, Experiment 3 took advantage of a more complex set of source utterances. It was launched with two batches of 70 subjects each receiving 25 root utterances, and trees targeted for 7 branches of depth 10, gathering a total 3546 utterances transformations. Subjects took an average 37 minutes to complete the 25 transformations, and were rewarded with £4.25 on average.

We now detail the evaluation of data quality and the measures that were taken to improve it. The section after that will focus on the fit of task parameters and source complexity, before moving on to the analysis and results.

3.3.2 Data quality

The choice of a Web-based setup sets the requirements of the interface much higher than for a laboratory experiment. There is no opportunity for a face-to-face walk-through of the experiment or for

[8] The public url of the experiment was not advertised anywhere else, and checking the subjects' Prolific Academic ID confirmed that only people from that platform participated in each experiment.

[9] The bug appeared only once a large proportion of trees had reached their target depth, and then affected all the subjects nearing completion of the experiment. The time taken to respond to complaints and realise that the experiment had to be paused led some subjects to exceed the maximum allowed time on Prolific Academic, and the service then sent the experiment out to new subjects. After fixing the bug, most subjects who had started the experiment accepted to finish it, leading the final subject count to be higher than originally requested.

questions, and subtle changes in the way the interface reacts to actions can lead subjects to interpret a signal where none was intended, or conversely to not notice an important message. The time of the subjects is not booked, and not having to travel to the laboratory or to talk to someone renders the interaction free of any commitment and generally more consumable: subjects can leave whenever they want, without having to feel bad about it (the only cost being the loss of their reward). The lack of human interaction with the experimenter also removes a natural incentive for subjects to take their time and perform according to what the experimenter in front of them explained. Combined together, these factors mean that if the interface is strenuous or ambiguous in any way, subjects will often pick the interpretations that make the process faster and either complete the experiment with minimal engagement or drop out. Redacting detailed textual instructions often makes matters worse. Instead, the interface must lead the subjects through the necessary explanations while remaining enjoyable, and must be unambiguous while still hinting towards the expected behaviour at the right moment, either through subtle interface reactions or through explicit contextual aids.

Manual spam-coding

Failure to properly encourage and wherever possible enforce the experiment's expectations led to data riddled with spurious transformations. Manual inspection of the data collected through Experiment 1, for which a substantial effort on instructions and for the overall interface had already been made, showed that large portions of the data were not usable as such. We therefore spam-coded all the utterances from this and subsequent experiments by hand. An utterance with any of the following properties was coded as spam:

- An ellipsis ("…") or other special characters (e.g. ">", "<") are present
- The utterance is partly or completely addressed to the experimenter (e.g. "Sorry, I can't remember")
- Over half the words are misspelled
- The utterance has no relationship to its parent utterance (i.e. it is an entirely new utterance)
- The utterance doesn't stand as an autonomous sentence, either because it is truncated or because so many words are garbled it becomes nonsense

Note that the last two criteria are not sharp, and several borderline cases had to be decided for the last one in particular. In Experiment 1 for instance, a subtle misunderstanding allowed by the interface led subjects to submit some sentences truncated at exactly 10 words, without regard to their meaning (see the details below); such utterances were unambiguously incomplete, and were thus coded as spam. In subsequent experiments however, utterances that could be made complete with the addition or the deletion of a single, sometimes unimportant word, were questioned by the same criterion. For instance the simple sentence "Mr Jones was robbed during" can be completed by adding the word "dinner" at the end, or by removing the word "during". Such sentences do not seem tied to a misunderstanding of the task, and are arguably attributable to temporary distraction whose effects are relevant to our analysis. The benefit of the doubt was given to such utterances, and they were not coded as spam.

Spam in transmission chains has the additional property of invalidating all the utterances that are made after it, such that the total number of utterances to discard is more than the spam introduced by subjects. Coded this way, Experiment 1 showed an accumulated spam rate of 22.4%. Combined with an initial technical oversight that led a small portion of utterances to be misplaced in the chains,[10]

[10] Ensuring that no two subjects are creating reformulations for the same chain tip at the same time, while not blocking other subjects from moving on with the experiment, is a non-trivial technical hurdle. Not solving it leads the chains to have "forks", that is, utterances with several children (possibly extending to sub-branches) instead of a single one. One of the children must

a total of 25.9% of the utterances generated by Experiment 1 were discarded. Apart from reducing the size of the usable data, spam also leads to uneven chains across trees, a heterogeneity that complicates the analysis. Accepting this level of spam was therefore not an option.

The main tool we used to reduce the level of spam is the user interface. As explained above, minor changes in the way the interface reacts to the subjects' actions combined with relevant context-dependent information can have a comparatively large impact on the spam rate.

User interface improvements

The situation is similar to that of surveys, where much effort is put into mitigating the risk of users engaging the minimum possible effort to complete the survey (Krosnick 2000). Successfully tuning the user interface is therefore a crucial factor in the quality of the data collected: what the interface might lead subjects to see as acceptable can easily be spam for the experimenter, and both perspectives must be aligned as much as possible. Interface design problems appeared repeatedly throughout the development of the platform and the pilots. The most important points can be summed up as follows:

- *Preventing digital copy-paste*: an obvious workaround to the task that most subjects will try in the first few trials.
- *Constraining the input*: a well-known behaviour in transmission chains of linguistic content is the rapid reduction in size of the content that is transmitted (Maxwell 1936; Bangerter 2000; Mesoudi and Whiten 2004). In order to encourage subjects to rely on what they remember, and prevent them from quickly reaching empty sentences, an early version of the experiment would disable the "send" button if the subject's input was shorter than 10 words (Experiments 2 and 3 later relaxed this constraint to 5 words). However, some subjects interpreted the button becoming active after 10 words as a signal that their input was ready to be sent as is, even if it was only a partial sentence. This ambiguity, corrected in later versions, is responsible for a large part of the spam found in Experiment 1.
- *Improving input quality*: Experiment 1 and subsequent pilots showed the need for strong incentives to write well-edited meaningful text. Indeed when pressed for time, some subjects will tend to write misspelled, poorly punctuated, or even meaningless utterances, which invalidate all the sentences that follow in the branch. Countering this tendency involved several changes to Experiments 2 and 3: emphasis was added to the fact that what is produced by one subject is later sent to other subjects, encouraging a more conscientious behaviour; a bonus was associated with high-fidelity trials, and the top 5 subjects with lowest transformation rates (as defined below in the analysis) received increased payment; most importantly, input from the subjects was also checked for repeated or inadequate punctuation, and for correct spelling against a combined British and American English dictionary. The interface asked subjects to correct any input that failed those tests, and presented them with a short explanation that emphasised the faulty behaviour and recalled the chain structure of the experiment. Inspecting the platform logs showed that this last measure led subjects to often correct their utterances, a fact that was also confirmed by the increased average writing time.
- *Relaxing the time pressure*: the interface of Experiment 1 made several mistakes that worsened the inherent pressure on subjects to complete the study as fast as possible (indeed, payment

then be chosen to form the main chain, and the others discarded. Solutions to the problem are difficult to test in practice, as they involve simulating dozens of subjects concurrently sending utterances to the platform. The approach adopted in Experiment 1 relied on client-side randomisation, but proved insufficient: 3.5% of the utterances posted by subjects were forks deep in the chains. Experiments 2 and 3 relied on a mix of client-side randomisation and server-side locking to solve the problem.

on Prolific Academic is per experiment, not per time spent – which, conversely, would encourage subjects to be very slow). First, subjects could terminate the reading time of an utterance at will. While this provided a measure of the effective reading time used by subjects, it also opened the possibility of speeding through the trials. Indeed over a third of the transformations of Experiment 1 were done by using less than half the allotted reading time. This pressure was increased by the presence of a "remaining time" clock in the reading and waiting phases, similar to the green clock shown on Fig. 3.3 for the writing phase. By removing superfluous clocks, keeping the reading time fixed, and proposing a break after each utterance, Experiments 2 and 3 relaxed the time pressure on the subjects and improved the final data quality.

- *Feedback channel*: survey design handbooks regularly insist on the importance of providing a channel for subjects to comment on the questions they were asked, and encourage the use of debriefing sessions to deepen that understanding (Leeuw, Hox, and Dillman 2008). Such feedback channels have also become a norm in online services, and we therefore chose to give the possibility for subjects to comment on most screens of the experiment (excluding the read-write screens) through a side-ribbon which, when clicked, would overlay a comment box (see Fig. 3.5). It seems, however, that a more interactive option would be more effective, as only a handful of subjects entered comments over the course of Experiments 2 and 3.
- *Instructions*: finally, a continuous effort was invested into fine-tuning the exact phrasing of instructions, and making the interface for instructions palatable using a now common pattern: for the instructions pictured in Fig. 3.3 for instance, different elements or images are successively foregrounded and highlighted, and a tooltip with short explanations appears next to the active element.[11] Here too, Experiment 1 and subsequent pilots allowed users to skip these instructions, leading a portion of the subjects to effectively never read them. Experiments 2 and 3 made navigating the complete list of instructions mandatory in order to start the trials.

These changes reduced the spam rate drastically. On the same criteria as Experiment 1, Experiment 2 showed an accumulated spam rate of .8%, which combined with misplaced utterances led to a total 1.4% of utterances discarded. Experiment 3 showed an accumulated spam rate of 1.0%, and with misplaced utterances had to discard a total 1.1% of the data.

3.3.3 Task difficulty and source complexity fit

In the simple task we used, difficulty is controlled by the reading time allotted to subjects and by the selection of source utterances. In order to unify reading times across utterances we decided that reading time would be proportional to the number of words in the utterance presented: for an utterance u, its number of words is noted $|u|_w$ and its reading time is defined as $|u|_w \cdot r$. We call r the *reading factor*. Average reading speeds for university students are usually between 200 and 300 words per minute, that is between 3.33 and 5 words per second (see Rayner, Slattery, and Bélanger 2010, where fast readers average at 330 wpm and slow readers average at 207 wpm). A reading factor of $r = 1$ therefore gives fast readers the time to read utterances more than 5 times, and slow readers about 3 to 4 times. A reading factor of $r = .3$ gives fast readers one or two readings, and slow readers at least one. $r = .2$ gives some readers one reading and others less than one, and $r = .1$ lets readers simply glance at the utterances without being able to read them completely.

Pilots and manual exploration indicated that the difficulty of the task is not linear with r. Whatever the value of r, longer utterances (more than 25 words) are often more transformed, relative to their length, than shorter utterances; longer utterances also give more space for the subjects to reformulate,

[11] The pattern was popularised by software libraries such as Intro.js (http://introjs.com/).

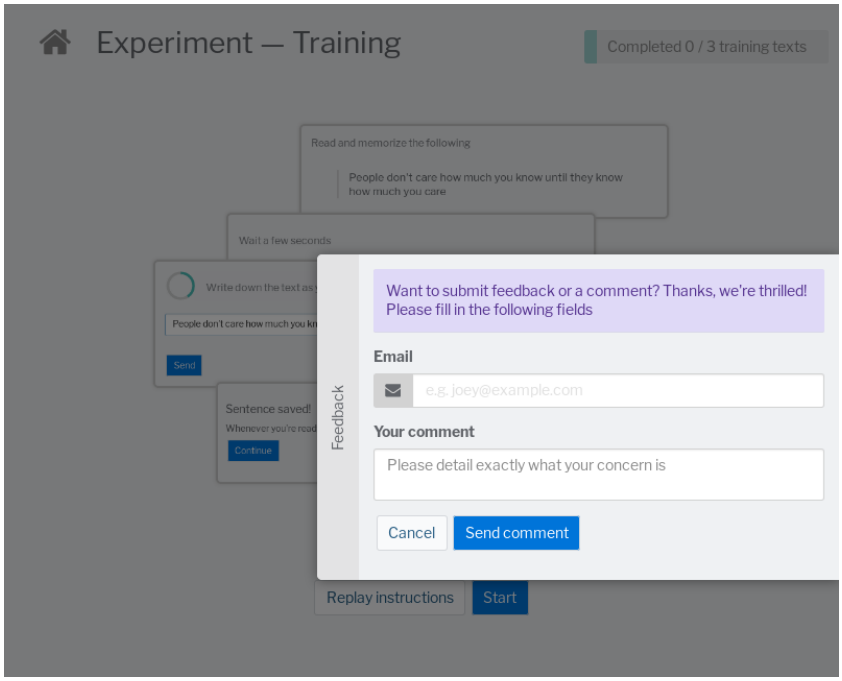


Figure 3.5: Overlay feedback box opened in the instructions screen from Fig. 3.3. The box is available in most screens of Experiments 2 and 3.

leading to more changes in style and permutations in the words. Changing r has less effect on shorter utterances than on longer utterances, and on utterances in oral versus written style. For short utterances in an oral style, pilots indicated that there is an abrupt transition between a low transformation regime when subjects can read the sentence at least once, and an extremely noisy regime when the subjects do not have the time to read the utterances entirely at their normal speed. Conversely, the transition is smoother for longer utterances or utterances with a more formal written style. Choosing an adequate set of source utterances is therefore an integral factor in adjusting the difficulty of the task.

Changing the source utterances also affects the sampling bias in ways that are difficult to measure given the multidimensionality of text. Contrasting minimally different utterances in different domains has resulted in domain-specific outcomes on for instance stereotypes, information hierarchy, and counter-intuitiveness (Kashima 2000a; Mesoudi, Whiten, and Dunbar 2006; Barrett and Nyhof 2001; Mesoudi and Whiten 2008), and authors have suggested that these outcomes are related to domain-specific biases in transformations (although to our knowledge these effects have not yet been studied jointly). In spite of this, we hypothesise that the low-level cognitive mechanisms underlying utterance transformation, that is the mechanisms that give rise to such accumulated outcomes, do not fundamentally change because of the type or the style of an utterance. If using news quotes instead of movie quotes or stories is likely to affect parameters of the observed transformations, it is less likely to affect the structure of the underlying cognitive mechanism, and therefore the general structure of transformations. Making this hypothesis lets us use utterance selection as an exploratory tool: by altering both the sampling of the transformations and the task difficulty, the exploration of different styles and types can help (1) improve data quality and (2) make general structure more visible, thus easier to measure and characterise. If this exploration yields insights about the struc-

ture of transformations and their effects in the long term, and if such insights are consistent with the previous chapter, then it will make sense to ask to what extent the uncovered structure is applicable to or varies with other types of utterances. Throughout pilots and experiments, our goal was therefore to find a set of utterances which would trigger varied transformations whose structure we could analyse, while at the same time helping the subjects to produce quality data by not creating too much pressure with reading time.

The set of sources used in Experiment 1 covered a broad spectrum of utterance types sampled from the following categories:

- Quotes from the MemeTracker data set used in the previous chapter,
- Famous compelling quotes from Wikisource[12] such as "Never doubt that a small group of thoughtful committed citizens can change the world, it is the only thing that ever has"
- Quotes extracted from the movie *12 Angry Men* such as "If you ask me I'd slap those tough kids down before they start any trouble, it saves a lot of time and money",
- Excerpts from news stories on controversial subjects (such as "How will the cultural and religious aspects of so many migrants impact E.U. society?") or risk-related subjects such as stories about the risks of Triclosan (used by Moussaïd, Brighton, and Gaissmaier 2015 in their study of the amplification of risk perception),
- The tale "War of the Ghosts" used by Bartlett (1995) in his original studies,[13] as well as excerpts from other tales,
- A small number of hand-crafted sentences such as surprising statements (e.g. "Don't forget to leave the door open when you leave the office") or stereotype-incongruent statements (e.g. "The young boy was suddenly hit by the little girl").

Each of these categories, we thought, could encourage the triggering of transformations. The spam level of Experiment 1, and especially the amount of misspelled words, made the exploration of the detailed transformations impossible and shifted the focus towards improving data quality through the interface. Nonetheless, it became clear that using such a heterogeneous set of utterances could surprise subjects, and was not the best approach to elicit regularities in transformations. Experiments 2 and 3 relied on a more thorough exploration of possible source data sets. Pilots explored utterances extracted from previous studies (Bangerter 2000 on personification and increased stereotypes; Heath, Bell, and Sternberg 2001 on the role of disgust; Maxwell 1936 on incoherent stories; Mesoudi, Whiten, and Dunbar 2006 for the role of social information).

Two larger and more homogeneous sets of utterances were reconstituted and finally used in Experiments 2 and 3. First, a set of movie quotes provided by Danescu-Niculescu-Mizil et al. (2012). This data set contains about 2200 pairs of quotes extracted from 1000 movie scripts; each pair is made of a quote that was marked as memorable by users of the Internet Movie Database, coupled with the closest quote in the same movie script that is spoken by the same character, has the same number of words, but is not marked as memorable on the Internet Movie Database. The 2200 pairs of quotes were filtered to keep only those which passed the spelling and punctuation quality tests from the previous section, and for which the number of words was strictly matched when excluding punctuation (this left 505 pairs). Second, a set of short stories from Fénéon and Sante (2007) was used. These stories are productions from Félix Fénéon originally anonymously published in the French newspaper *Le Matin* in 1906. They describe facts from everyday life such as accidents, suicides, or trials, in a terse and sometimes humorous style. A sample of 60 stories was extracted from the English version, for which French names and places were replaced with names and places more familiar to British subjects. Pilots explored these sets of utterances with reading factors of .1, .2, .3, .75 and

[12] https://en.wikisource.org/.
[13] Available online at http://penta.ufrgs.br/edu/telelab/2/war-of-t.htm.

1. Finally, tests were also made using these utterances with content words replaced with pseudo-words, in order to restrict effects to the grammatical dimension only.[14] The pseudo-word tests were inconclusive, as the task became too confusing and subjects often replaced unknown words with real words.

Experiment 2 used 25 of the 27 pairs of movie quotes that had exactly 15 or 16 words, providing a homogeneous set of 50 utterances in oral style, with a reading factor of .75. Experiment 3 used 43 of the 60 short stories by Fénéon (average number of words 21.2) coupled with 4 utterances extracted from Mesoudi, Whiten, and Dunbar (2006) (average number of words 60.3) and 3 utterances extracted from the story used by Maxwell (1936) (average number of words 40.7), with a reading factor of 1.

ADD: clean counts.

ADD: mispelling proportion in exp 1

ADD: a few example sentences for each experiment

3.4 Results

Recall that the goal we set ourselves is to provide a better understanding of the process at work than what low-level feature analyses such as that of the previous chapter. In doing so we also hope to bring some light to the processes underlying high-level contrasts of utterance categories that have been extensively studied in the literature. The analysis we present is thus geared towards creating an intermediary representation of the effect of transformations on utterances, one that is at a midpoint between the low-level of word features and the high-level of category contrasts, and can be usefully modelled to better understand the evolution of utterance chains. Since this work was exploratory in nature, our presentation will also loosely follow a step-by-step development of the analysis with intermediate results. Our analysis consists in five broad steps. First, a presentation of the general trends observed in the collected data, which provide a coarse but relevant view of the behaviour of utterance reformulations in these experiments. Second, the actual procedure developed to break down transformations into smaller blocks and grasp their detail. Third, we develop a descriptive model of transformations based on the detailed view that the previous step provided. We then refine this view by quantifying the main behaviours that the model lets us identify in the transmission chains. Finally, we characterise the lexical features of the words identified by the transformation model, and show how the accumulation of transformations gradually evolves the average features of utterances. We begin with the general trends observed in the data.

3.4.1 General trends

We begin the analysis of our three data sets by examining the evolution of aggregate measures as a function of depth in the trees. Here and in what follows, the analyses are made on the data cleaned of spam, and chains truncated at their target depth: the data from Experiment 1 is truncated at depth 8, Experiment 2 at depth 7, and Experiment 3 at depth 10, and all plots are aligned to the same depth axis to facilitate comparisons. The goal of this section is to give an overview of the shape of the data and highlight a few important trends to keep in mind in the rest of the analysis. The model

[14] Pseudo-words were generate using the Wuggy library (Keuleers and Brysbaert 2010).

we develop further down will then give a more precise view of the mechanisms underlying these trends.

Utterance length

A well-known effect in transmission chains with linguistic content is the quick reduction of utterance length as chains progress. These experiments are no exception: Fig. 3.6a shows a scatter plot of the number of words of an utterance $|u|_w$ versus depth in a tree.[15] The insets show the data restricted to trees for which root utterances have 30 words or less (thus most utterances in those trees also have 30 words or less); this boundary keeps all the Fénéon root utterances in Experiment 3, and we use it to separate longer from shorter utterances for the purposes of this figure. The plots confirm that word length quickly decreases as subjects read and rewrite utterances, and indicate that the reduction depends on the size of what is being transformed: very long utterances (above 100 words) are reduced to less than 100 words in 2 reformulations or less, whereas root utterances with up to 28 content words can maintain their size until the end of the branches of Experiment 3. Note that the differences in the speed of size reduction across the experiments are tied to surface features of the root utterances. Word count and average word frequency in particular, which we will later show are strongly related to transformation rate, have different distributions in the set of root utterances of each experiment: word counts are disjoint between Experiments 2 and 3, and root utterances from Experiment 2 are in an oral style, with a higher proportion of stopwords than in Experiments 1 and 3 (stopwords are always high-frequency words, and make up 67% of the root utterances of Experiment 2, versus 58% in Experiment 1 and 48% in Experiment 3). The steeper slopes in Experiments 1 and 3 compared to Experiment 2 are thus tied to the higher word counts and lower proportions of stopwords in their root utterances.

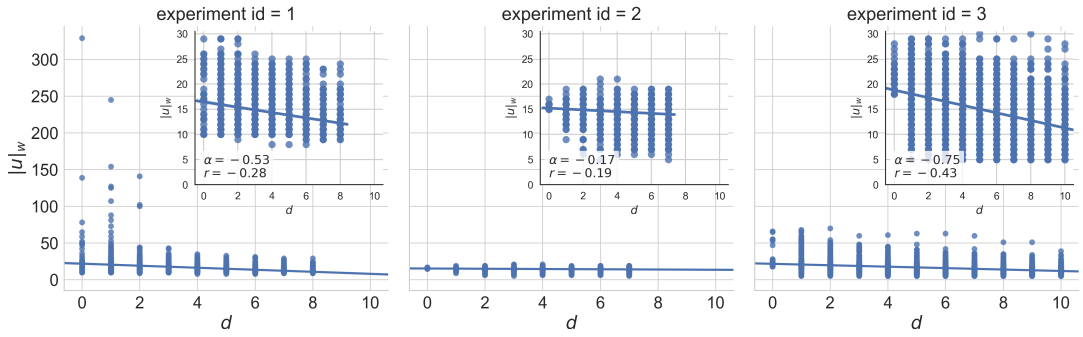
Next, we eliminate stopwords from the utterances and focus on the reduction in number of content words (notice however that stopword recognition is less reliable in Experiment 1 than in Experiments 2 and 3, because of spelling mistakes). For a given utterance u, the list of its content words is noted $c(u)$, and the number of content words is therefore $|c(u)|_w$, where $|\cdot|_w$ is extended to provide the length of a list of words (aside from counting words in a string). Fig. 3.6b plots the content word counts for the same utterances as those in Fig. 3.6a (in particular, insets show the same utterances in both figures), and shows a similar reduction in counts across all experiments. Here too, the differences in regression values across experiments and between the two figures are tied to the differences in distributions of word count and proportion of stopwords in the roots. In other words, the size reduction is sampled differently by each experiment: these figures show how the effect acts on each set of root utterances, but do not indicate that the mechanism is any different across experiments nor that it depends on the actual meaning of the utterances (versus primarily on surface features such as word count and average word frequency).

Utterance to utterance distance

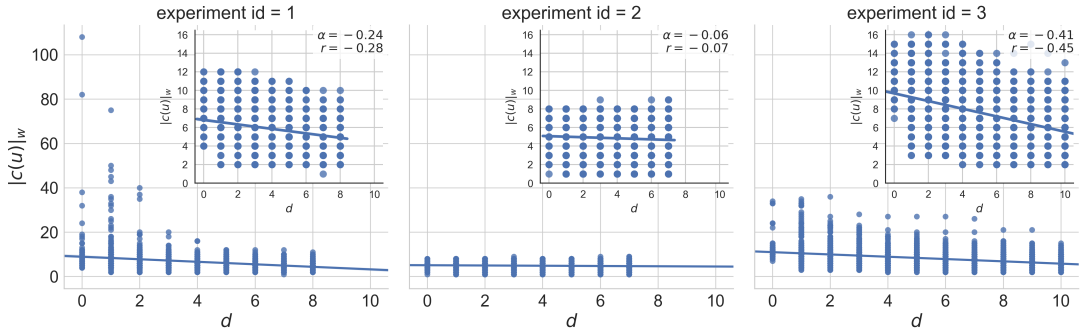
As a first approximation to the magnitude of transformations we introduce a measure of the distance $\lambda(u, u')$ between two utterances u and u', defined as the Levenshtein distance[16] between the lemmas of the content words of u and u':

[15] All NLP computations in this chapter are performed using the spaCy library for Python, version 1.9.0, available at http://spacy.io/.

[16] The Levenshtein distance (also known as the edit distance) is defined between two lists of items, and counts the minimal number of insertions, deletions, and replacements that are needed to transform the first list of items into the second. It has all the properties of a metric (non-negativity, identity of indiscernibles, symmetry, and subadditivity).



(a) Numbers of words, $|u|_w$



(b) Numbers of content words, $|c(u)|_w$

Figure 3.6: Reduction in utterance word count and content word count as a function of depth (d) in the three experiments. Each blue dot represents an utterance. For a given experiment, the same utterances appear in panels (a) and (b). The insets show the utterances for which the tree root has 30 or less words ($|u_{\text{root}}|_w \leq 30$), with the numerical values of the linear regression slope and correlation coefficient. All regression slopes are non-zero with $p < .001$.

$$\lambda(u, u') = \text{lev}\left(\text{lemmatize}(c(u)), \text{lemmatize}(c(u'))\right)$$

For example, consider the three following utterances taken from Experiment 1 (in a tree whose root is from the MemeTracker data set):

u_a: "This crisis did not develop overnight and it will not be solved overnight" (3.1)

u_b: "the crisis did not developed overnight, and it will be not solved overnight" (3.2)

u_c: "The crisis didn't happen today won't be solved by midnight." (3.3)

After removing the punctuation and converting all words to lowercase, the lemmas of the content words of these utterances are as follows:

lemmas($c(u_a)$): "crisis", "develop", "overnight", "solve", "overnight"

lemmas($c(u_b)$): "crisis", "develop", "overnight", "solve", "overnight"

lemmas($c(u_c)$): "crisis", "happen", "today", "solve", "midnight"

Such that $\lambda(u_a, u_b) = 0$ and $\lambda(u_a, u_c) = \lambda(u_b, u_c) = 3$. λ thus measures differences in content lemmas and obviates minor transformations such as changes of stopwords or word inflexions. In order to have a uniform quantity across utterances of different lengths, we define the *transformation rate* ρ as the normalised distance between two utterances:

$$\rho(u, u') = \frac{\lambda(u, u')}{\max\left(|c(u)|_w, |c(u')|_w\right)}$$

ρ thus measures the magnitude of the transformation between the contents of u and u', relative to the size of the contents of those utterances. It takes its values between 0 and 1: $\rho(u, u') = 0$ if and only if u and u' have exactly the same content words in the same order, and $\rho(u, u') = 1$ means that the content words of u and u' have so little in common that rewriting from scratch is quicker than changing one into the other with word insertions, deletions, or replacements. Here, $\rho(u_a, u_b) = 0$ and $\rho(u_a, u_c) = \rho(u_b, u_c) = .6$. A major caveat of this measure is that it does not know about synonyms or expressions with similar meaning, such that two sentences separated by a transformation rate of 1 can have the same meaning at a higher level. For instance with the following sentences,[17]

u_d: "Will you investigate the gravest crimes of the Bush administration, including torture and warrantless wiretapping?" (3.4)

u_e: "Will you research the worst problems of the 2004 mandate, like its surveillance?" (3.5)

u_f: "Don't forget to leave the door open when you leave the office" (3.6)

we have $\rho(u_d, u_e) = \rho(u_d, u_f) = 1$. The measure misrepresents the changes between these utterances, as u_d and u_e can easily be considered to have similar meanings at a high level, and their

[17] u_d and u_f are from Experiment 1 (u_d being originally from the MemeTracker data set), and u_e was created for this comparison.

difference is far less important than the difference between u_d and u_f. Nonetheless, the measure performs reasonably well on utterances inside the same tree: in that context all utterances come from the same source and have therefore some level of meaning in common, and there is no need to differentiate between the types of transformations that $u_d \to u_e$ and $u_d \to u_f$ represent.

Transmissibility and transformation rate

Together with transformation rate, we examine a measure derived from it: the *transmissibility* of utterances, defined as the proportion of utterances at a given depth whose content lemmas are perfectly transmitted to their child, computed over all the branches of all the trees of an experiment (this measure was introduced as 'average success' in Claidière et al. 2014). If we note $\mathbb{1}_{\lambda(u,u')=0}$ the success of a subject in transmitting an utterance's content (it equals 1 if the content lemmas of u and u' match perfectly, and 0 if there was any change in content lemmas), the transmissibility $\eta(d)$ of utterances at depth d can be expressed as:

$$\eta(d) = \left\langle \mathbb{1}_{\lambda(u,u')=0} \right\rangle_{u \text{ at depth } d,\ u' \text{ child of } u}$$

Transmissibility provides a coarser measure of the evolution of transmission success than transformation rate (a change in transmissibility implies a change of transformation rates), but lets us better differentiate between the two important alternatives: perfect transmission, and transformation. A classic effect of transmission chains for various types of content is that transmissibility increases with depth in the chains.

Fig. 3.7 shows the transmissibility and one minus the transformation rates ($1 - \rho$) for the three experiments, both overall and grouped by content length of the utterances. Fig. 3.7a shows an increase in transmissibility with respect to depth (from .40 to .67), when considering the whole data set from Experiment 1. However, the plots on the right show only a slight increase in transmissibility (or even no increase at all for $|c(u)|_w \notin [7, 10]$) for utterances of a given content length. The right-hand side also indicates that transmissibility depends on content length, as the transmissibility lines become gradually lower when content length increases (average .92 for 2 content words, .20 for 12 content words). Together, these trends indicate that the overall increase in transmissibility with respect to depth could be mostly due to the rarefaction of utterances with long content length: as depth increases, the proportion of shorter utterances increases; shorter utterances are better transmitted, and as consequence global transmissibility increases too.

Fig. 3.7b shows the same analysis for Experiment 2. Contrary to the previous case, transmissibility here is stable at .82-.88 with respect to depth, both for the whole data set and at fixed content length. It also depends less on content length than in Experiment 1, as utterances with 2 content words have an average transmissibility of .95, and utterances with 8 content words an average transmissibility of .69.

Experiment 3 (Fig. 3.7c) features an increase in transmissibility with respect to depth both globally (from .18 to .71) and at long fixed content length. This effect is stronger than in Experiment 1, and indicates that long utterances in the data set become slightly easier to transmit as they are transformed. As noted previously, utterances found at the end of a chain will often come from much longer utterances at the start, such that improved transmission success along a single branch is always mixed with the shortening of content. However, for long utterances (content lengths 8 and above), utterances found at the end of all chains are on average better transmitted than utterances *of the same content length* at the start of all chains, meaning that transmission along the chains has an

effect on transmissibility of long utterances beyond the shortening of the content. Finally, the different behaviours across experiments are here again tied to the differences in word count and stopword proportion distributions in the root utterances.

Variability

We close this overview of the general trends in all experiments with a final measure: the variability of utterances at a given tree depth. For a given tree t, the variability $\kappa(t, d)$ measures the average transformation rate between all pairs of utterances at depth d in t (henceforth the *slice* of t at depth d):

$$\kappa(t, d) = \langle \rho(u, u') \rangle_{\{u,u'\} \subset \{u \text{ at depth } d \text{ in } t\}}$$

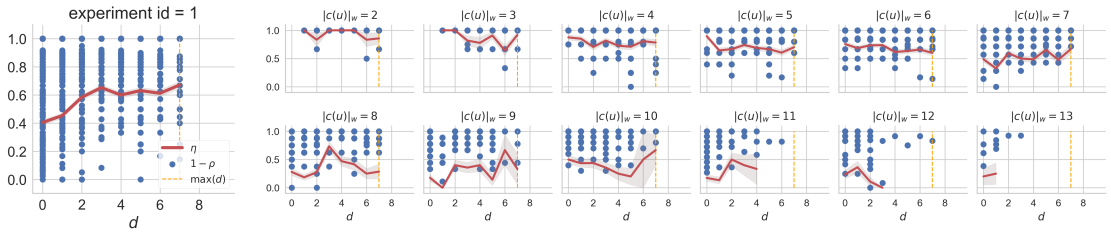
This measure gives a sense of how fast branches diverge between each other. For each experiment, Fig. 3.8 plots the variability of all slices of all trees, and the average variability averaged across trees. All three increase significantly, meaning that utterances from different branches in a tree become more and more different as the chains progress. The increase is sublinear and plateaus for Experiments 1 and 3, suggesting that branches diverge most at the beginning and less at the end. This is consistent with the increases in transmissibility. The different divergence rates correspond to the transformation rates observed in Fig. 3.7 (Experiment 2 has lower transformation rates, and diverges slower), and are therefore again tied to the differences in root utterances.

These three measures provide a coarse view of the speed at which utterance length is reduced, whether or not transformations make utterances easier to remember, and how fast the specificity of branches develops. However, they provide little insight into the detail of these trends and the transformation mechanisms that underlie them. We address this question by constructing a model of transformations in three broad steps: break down the transformations into a more detailed encoding of operations, visualise these operations to create a descriptive model of transformations, and finally quantify the main behaviours that the model allows us to observe. In what follows we focus on the data set from Experiment 3, which provides the best quality of data and sampling of transformations. The procedure we present is applicable to the other two experiments, but we will not discuss those applications here.

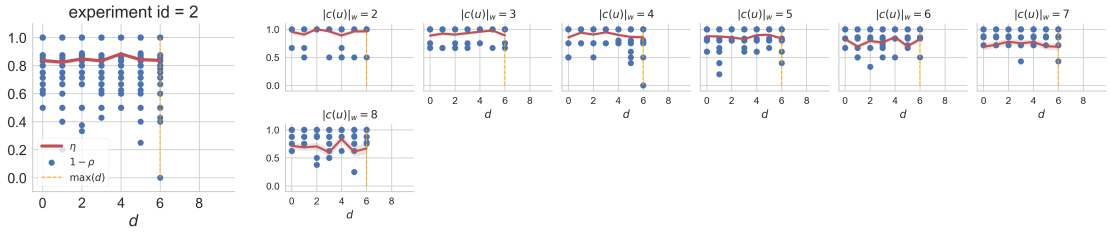
3.4.2 Transformation breakdown

Sequence alignments

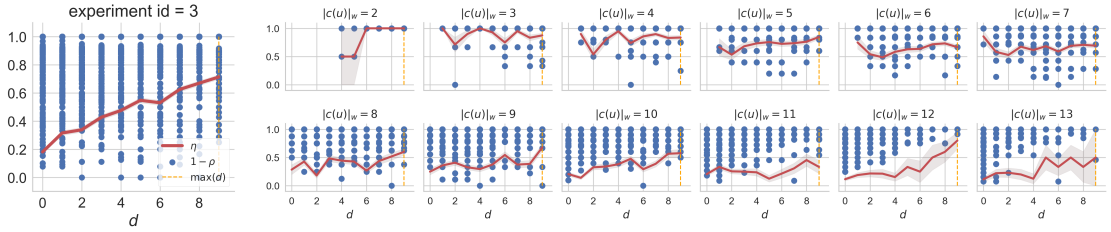
Our first step to construct a model of transformations is to take advantage of existing generalisations of the Levenshtein distance underlying the transformation rate measure ρ. Recall that the Levenshtein distance between two sequences of items s and s' computes the minimal number of insertions, deletions, and replacements necessary to turn s into s' (and vice-versa). This problem can equally be formulated as that of aligning the items of s and s': each item of s can be paired either with an item from s' (signifying a conservation if both items match, or a replacement if the two items are different), or with nothing (signifying a deletion in the transformation of s into s'). Symmetrically, items from s' can also be paired with nothing (aside from being paired with items from s), signifying an insertion in the transformation of s into s'. In this formulation, insertions and deletions are unified into the same operation: a "gap" (or "indel" for insertion-deletion), found



(a) Experiment 1



(b) Experiment 2



(c) Experiment 3

Figure 3.7: Transmissibility and conservation rate for each experiment. Each individual graph shows both transmissibility (red line) and one minus the transformation rate (blue dots) for a subset of all utterances. Light red areas are the 95% confidence intervals for transmissibility, based on Student's t-distribution and considering each transformation as an independent event. A blue dot at $y = 1$ is an instance of perfect transmission ($\rho = 0$), and pulls transmissibility upwards; a blue dot anywhere below is a transformation instance ($\rho > 0$), and pulls transmissibility downwards. The large plot on the left shows both measures for all the utterances of an experiment. The small plots on the right show both measures for utterances that have a given number of content words (up to 13, after which the data is nonexistent or very sparse in all experiments). The orange dashed line marks the maximum depth in the experiment, so as to differentiate content lengths reaching the limit from content lengths disappearing before the limit.

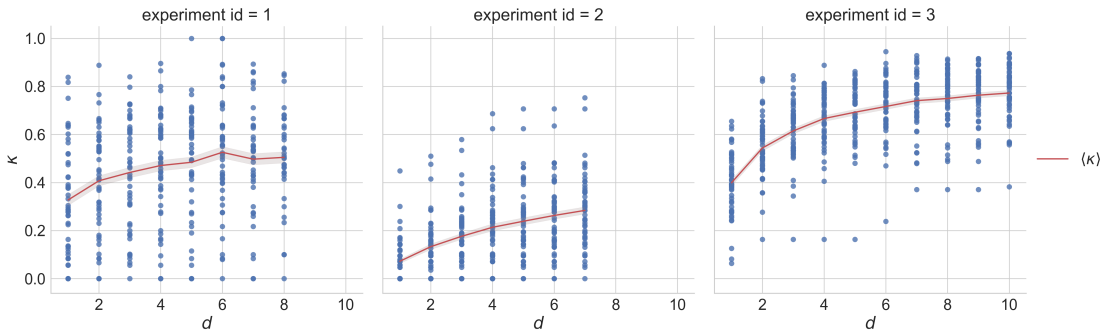


Figure 3.8: Slice variabilities in the three experiments. Each plot shows the variabilities of each slice of each tree (blue dots), as well as the average variability across slices of all trees at a given depth (red line with 95% confidence interval based on Student's t-distribution, considering each tree slice as an independent measure).

either in s or in s'. The problem thus formulated has become extremely important over the last 50 years in the subfield of bioinformatics known as sequence alignment.

Sequence alignment is in the business of looking for similarities between sequences of DNA, RNA, or amino acids in proteins that could indicate evolutionary or structural relationships between two or more species. Research on this problem has led to the development of several generalisations of the algorithm underlying the Levenshtein distance; these are geared towards assigning different weights or costs to the individual operations transforming one sequence into the other, finding optimal alignments of subparts of the two sequences (a task known as local alignment, in contrast to global alignment), or aligning more than two sequences simultaneously (multiple sequence alignment, in contrast to pairwise alignment).

The structure of the problem is strikingly similar to our present task: we aim to decompose the transformation of a parent utterance into a child utterance into a combination of small basic operations. In sequence alignment terms, this task is a pairwise global alignment of lists of words, for which the Needleman-Wunsch algorithm (Needleman and Wunsch 1970, henceforth NW) provides a flexible generalisation of the Levenshtein distance. For two sequences of items of any type, s and s', the NW algorithm assigns different scores to each basic operation (gap, mismatch a.k.a. replacement, and match, which is considered a scored operation like the first two), and returns the list of alignments between s and s' with maximal total score. There can be several such alignments, and each of them can be directly interpreted as a minimally scoring list of operations to transform s into s' (and vice-versa).

More precisely, let us note $s = (s_1, ..., s_n)$ and $s' = (s'_1, ..., s'_{n'})$ the items in both sequences, with n and n' the lengths of the sequences. The NW algorithm returns pairs of sequences a and a' of lengths $m \geq \max(n, n')$, made of the items from s and s' (respectively), in the same order, but interspersed with a "gap item". We noting the gap item g, and the alignment sequences $a = (a_1, ..., a_m)$ and $a' = (a'_1, ..., a'_m)$. Each tuple (a_i, a'_i) then represents the pairing of an item from s with an item from s' (either match or mismatch), or with a gap if $a'_i = g$ (and vice versa if $a_i = g$). Considered as a transformation from s to s', gap items in a' represent deletions, and gap items in a represent insertions. Each pair can thus be seen as an operation taking an item from s to construct s', and a and a' are such that the sum of scores of the operations they represent is maximal.

Take for instance the DNA sequences $s =$ AGAACT and $s' =$ GACG. An example alignment between

the two sequences can be represented as follows (with the gap item represented as "-", and matches shown with vertical bars):

$$\begin{array}{r} a = \texttt{AGAACT-} \\ \texttt{\quad\ \ | \ ||} \\ a' = \texttt{-G-AC-G} \end{array}$$

The power of the NW algorithm is that gap, mismatch and match scores can be defined at compute time, knowing what items are being compared (or evaluated for a gap), and what operations would have been made up to that point if this operation were to be part of an optimal alignment. This flexibility has been used in biological sequence alignment to account for the fact that, in a DNA sequence for instance, the deletion of a base in the middle of an otherwise intact portion of DNA is less probable than the continuation of a gap that has already started. In other words, in biological sequence alignment opening a new gap is more costly than extending an existing gap, and the compute-time scores of gaps can reflect that. The same goes for mismatches: not all bases are equally likely to replace another base, and mismatch scores can factor that into the evaluation of alignments. As is hopefully clear by now, the situation is strikingly similar for sequences of words. In the next sections we detail our application and extension of the NW algorithm to decomposing utterance transformations.

Application to utterance alignment

The NW algorithm can be straightforwardly applied to sequences of any kind, provided we define scores for opening and extending gaps and a function to evaluate the comparison of two items (henceforth the match scoring function). We thus apply it to sequences of words without punctuation, with a match scoring function that takes into account the semantic distance between the two words compared. For a given pair of utterances u and u', we start by tokenising them and removing all punctuation. We then apply the NW algorithm[18] on the resulting sequences of tokens, with a match scoring function computed as an affine transformation of the similarity between two words w and w':

$$\text{similarity}(w, w') = \begin{cases} S_C\left(\boldsymbol{w}, \boldsymbol{w}'\right) & \text{if we have word vectors for both } w \text{ and } w' \\ \delta_{\text{lemma}(w),\text{lemma}(w')} & \text{otherwise} \end{cases}$$

where S_C is the cosine similarity function (one minus cosine distance) and $\boldsymbol{w}$ is a 300-dimensional vector representation of w encoding the word's semantics,[19] such that the $S_C(\boldsymbol{w}, \boldsymbol{w}')$ provides a measure of semantic similarity between w and w'. Finally, $\delta_{i,j}$ is Kronecker's delta which equals 1 if and only if $i = j$, and 0 otherwise. This function thus provides a "best effort" similarity measure which depends on whether we have semantic information about the words being compared or not.

[18] We used Biopython's implementation of the NW algorithm (Cock et al. 2009).

[19] Vector representations (also known as "word embeddings") encode words as vectors in a high-dimensional vector space. The high-dimensionality allows the vectors to bear part of the semantic information of the words as they appeared in a training corpus. Large pre-trained vocabulary sets are available in many NLP libraries, and the standard spaCy English language model includes "vectors for one million vocabulary entries, using the 300-dimensional vectors trained on the Common Crawl corpus using the GloVe algorithm" (https://alpha.spacy.io/docs/usage/word-vectors-similarities). The GloVe algorithm was introduced by Pennington, Socher, and Manning (2014), and is one of several possible methods to train such word vectors (another well-known family of methods being word2vec).

Adding an affine transformation to similarity lets us adjust its importance with respect to gap scores, for which we only differentiate opening and extension scores. This definition thus uses an initial 4 scalar parameters (two gap scores, two affine parameters) that define the way each operation is scored against the others. Since the final score of an alignment is computed as the sum of the scores of its individual operations, a linear scaling of all the parameters by the same amount does not change the choice of best-scoring alignments, such that we can further reduce the number of parameters by one. We choose to set the slope of the affine transformation of similarity to 1, and are then left with 3 alignment parameters:

- $\theta_{mismatch}$, the base score for the match scoring function, such that

$$\text{score}(w, w') = \text{similarity}(w, w') + \theta_{mismatch}$$

- θ_{open}, the score for opening a gap; θ_{open} is negative since it is a cost,
- θ_{extend}, the score for extending a gap; θ_{extend} is also negative.

Given the right set of parameters, the alignment produced by the NW algorithm to transform one utterance into another is a good approximation of the internal operations of said transformation. Take for instance the following two utterances from Experiment 3:

> "Finding her son, Alvin, 69, hanged, Mrs Hunt, of Brighton, was so depressed she could not cut him down." (3.7)
>
> "Finding her son Arthur 69 hanged Mrs Brown from Brighton was so upset she could not cut him down" (3.8)

With the set of parameters that we obtain through training as explained below, the algorithm aligns these two utterances as follows (noting any gaps with "-", and emphasising replacements):

```
Finding her son Alvin  69 hanged Mrs Hunt of -     -    Brighton was so depressed
Finding her son Arthur 69 hanged Mrs -    -  Brown from Brighton was so upset

she could not cut him down
she could not cut him down
```

Detecting exchanges

Applying the NW algorithm in this manner works well for simple transformations such as the example above. However, more complicated transformations include operations that the algorithm does not know about. Hand inspection of the data showed that exchanging sub-parts of an utterance, in particular, is a relatively common operation for which our current tool has no representation. Consider the following two utterances from Experiment 3 for instance:

> u_a: "At Dover, the finale of the bailiffs convention, their duty said a speaker are delicate, dangerous and detailed" (3.9)
>
> u_b: "At Dover, at a Bailiffs convention. a speaker said that their duty was to patience, and determination" (3.10)

The current alignment algorithm, with parameters trained according to a procedure outlined below, produces the following:

```
At Dover the finale of the -  - bailiffs convention - -       -    -    their duty
At Dover -   -      -  -   at a Bailiffs convention a speaker said that their duty

said a speaker are delicate dangerous -   -  -        and detailed -
-    - -       -   -        -         was to patience and -         determination
```

This alignment misses the fact that the deleted part "said a speaker" is found as "a speaker said" earlier in the reformulated utterance. The general idea to detect such exchanges is that blocks of insertions and blocks of deletions can be matched against one another with the same alignment algorithm, and the resulting deep (recursive) alignment can be scored and compared to the initial shallow alignment. If the final deep score $\chi_{deep}(u_a, u_b)$ is higher than the initial shallow score $\chi_{shallow}(u_a, u_b)$, then we adopt the deep alignment with exchange as the best solution. Suppose that for the alignment of the deletion block u_- "said a speaker are delicate dangerous" with the insertion block u_+ "a speaker said that", we are able to compute an optimal deep alignment with associated score $\chi_{deep}(u_-, u_+)$; then the deep score for the top level $\chi_{deep}(u_a, u_b)$ is as follows:

$$
\begin{aligned}
\chi_{deep}(u_a, u_b) = \chi_{shallow}(u_a, u_b) \quad & \text{initial shallow score} \\
+ \theta_{exchange} \quad & \text{score the addition of an exchange operation} \\
- \text{score(deletion of } u_-) \quad & \text{recover the cost of the deletion block} \\
- \text{score(insertion of } u_+) \quad & \text{recover the cost of the insertion block} \\
+ \chi_{deep}(u_-, u_+) \quad & \text{add the deep alignment score of the exchange}
\end{aligned} \tag{3.11}
$$

where $\theta_{exchange}$ is a new negative parameter that defines the cost of creating an exchange, to be added to the existing three shallow alignment parameters. The deep alignment extension we implemented follows exactly that recursive principle, but accommodates for the possibility of multiple exchanges at each level of the recursion. Algorithm 3.1 provides an overview of the way this tree of alignments can be constructed. Note that for long utterances, the size of the deep alignment tree can grow very fast:

- For a given deep alignment, there is a list of mappings between deletion and insertion blocks,
- Each mapping is a set of (deletion block, insertion block) pairs,
- Under each such pair, there is a list of deep alignments; and from there on recursively.

Also, our implementation of the exploration of that tree is mostly brute force, and does not try to be smart in predicting which branches are dead-ends. In spite of this, we did not need to optimise the computation any further (aside from obvious gains in caching repeated computations), as most of the time of a deep alignment is spent computing shallow alignments, and most alignments of utterances are very shallow anyway. Finally, note that this approach provides no guarantee of finding the globally optimal deep alignment. Indeed, it starts from optimal shallow alignments, and explores the tree of possibles from there on. But the initial shallow alignments it extends may not be the best starting point, such that the exploration may return locally optimal deep alignments.

Nonetheless, given a good set of parameters (see the next section where we derive those), this deep alignment algorithm produces surprisingly satisfying results given the simplicity of its underlying principles. In the case of the two utterances exemplified at the beginning of this section, the algorithm produces the following deep alignment tree. First, the top-level alignment:

```
At Dover the finale of the -  - bailiffs convention |-Exchange-1------| their duty
At Dover -   -      -  -   at a Bailiffs convention a speaker said that their duty
```

Algorithm 3.1 An implementation of the deep alignment extension for detecting exchanges in NW alignments. Note that this implementation is inefficient but presentationally clearer than the implementation we made.

function SHALLOWALIGN(u, u')
  (Implemented by Biopython)
  return List of optimal shallow alignments of u and u'
end function

function MAPPINGS($a_{shallow}$)
  $\mathcal{D} \leftarrow \{d | d$ deletion block in $a_{shallow}\}$
  $\mathcal{I} \leftarrow \{i | i$ deletion block in $a_{shallow}\}$
  return $\mathcal{D}^{P(\mathcal{I})}$ ▷ $P(\Omega)$ is the power set of Ω
end function

function SCOREMAPPING($a_{shallow}$, D_M)
  $s \leftarrow 0$
  for $((u_{e,-}, u_{e,+}), D_e)$ in D_M **do**
    $s \leftarrow s + \theta_{exchange}$
    $s \leftarrow s -$ SCORE(deletion of $u_{e,-}$) $-$ SCORE(insertion of $u_{e,+}$)
    $s \leftarrow s + \max\{\chi_{deep}(a_{deep}) | a_{deep} \in D_e\}$
  end for
  return s
end function

function DEEPALIGN(u, u')
  $D \leftarrow []$ ▷ D is the list of deep alignments trees we have explored
  for $a_{shallow}$ shallow alignment in SHALLOWALIGN(u, u') **do**
    if $a_{shallow}$ has no gaps or has only gaps **then**
      $D \leftarrow (a_{shallow}, [], \chi_{shallow}(a_{shallow}))$
    else
      for M mapping in MAPPINGS($a_{shallow}$) **do**
        $D_M \leftarrow []$
        for $(u_{e,-}, u_{e,+})$ exchange in M **do**
          $D_M \leftarrow D_M + ((u_{e,-}, u_{e,+}),$ DEEPALIGN$(u_{e,-}, u_{e,+}))$
        end for
        $D \leftarrow D + (a_{shallow}, D_M, \chi_{shallow}(a_{shallow}) +$ SCOREMAPPING$(a_{shallow}, D_M))$
      end for
    end if
  end for
  return Recursively pruned version of D with only maximally scoring deep alignments
end function

```
said a speaker are delicate dangerous -   -  -        and detailed -
|-Exchange-1-----------------------| was to patience and -        determination
```

For which $\chi_{shallow} = -2.93$ and $\chi_{deep} = -2.89$. Then the alignment of Exchange 1:

```
said a speaker are delicate dangerous |E2-----|
|E2| a speaker -   -        -         said that
```

For which $\chi_{shallow} = -1.01$ and $\chi_{deep} = -0.99$. And finally the alignment of Exchange 2, from inside Exchange 1:

```
said -
said that
```

For which $\chi_{shallow} = \chi_{deep} = -0.18$.

Notice how in this deep alignment the phrase "are delicate and dangerous" was initially included in Exchange 1, only later to be recognised as a deletion in the alignment of Exchange 1. The same happened for "that", initially included in Exchange 1 and finally recognised as an insertion in the alignment of Exchange 2. Most cases of deep alignments look like this one, where a single path exists in the tree of recursive alignments. For longer utterances however, there can be several exchanges at each level, and the tree of alignments becomes much larger.

Training alignment parameters

Finally, we need to determine a set of alignment parameters that produce useful results with this procedure. Recall that the parameters are:

- $\theta_{mismatch}$, the base score for the match scoring function,
- θ_{open} and θ_{extend}, the scores for opening and extending a gap,
- $\theta_{exchange}$, the score for creating an exchange.

In order to make the problem of finding usable parameters tractable, we decided to restrict parameter training to the shallow alignment parameters only (henceforth noted $\boldsymbol{\theta} = (\theta_{mismatch}, \theta_{open}, \theta_{extend})$), and fine-tune $\theta_{exchange}$ by hand after the first three were defined (this also corresponds to the fact that deep alignments are made on the basis of optimal shallow alignments). Our general approach for this task is therefore to hand-code shallow alignments for a random set of utterance transformations in Experiment 3, then train the shallow alignment parameters to that standard before adjusting the exchange parameter by hand. Since there are only three dimensions to explore, the training step is easiest to accomplish by brute force.

We thus start by evaluating the size of the training set that is necessary to obtain a set of parameters that extrapolates well to untrained data. Indeed, a training set too small in size might provide too weak a constraint on the set of parameters, such that brute forcing would find many parameter sets that do not extrapolate well. On the other hand, manually coding alignments is time-consuming and we do not wish to code more than necessary. We used the following procedure to decide this trade-off:

1. Uniformly sample a random parameter set $\boldsymbol{\theta}^0 \in [-1, 0]^3$ and use it to generate artificial alignments for all the non-trivial transformations in Experiment 3 (i.e. for which the transformation rate ρ was positive, which amounts to 2159 transformations); note these alignments $\mathcal{A}^0$.
2. Sample a training set of size n from the artificial alignments; note the training set $\mathcal{A}^0_t$, and the remaining evaluation set $\mathcal{A}^0_e = \mathcal{A}^0 \setminus \mathcal{A}^0_t$.

3. Brute-force the sets of parameters $\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_m$ that best reproduce the training set $\mathcal{A}_t^0$, by exploring the sampling space $[-1, 0]^3$ with a discretisation step of .1; parameters that perfectly reproduced the training set were always found, such that no finer-grained exploration was needed.
4. Evaluate the worst fit $\hat{f}_n$ between the evaluation alignments $\mathcal{A}_e^0$ and the alignments produced by each of the $\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_m$ on the same transformations.

For a given set $\mathcal{T}$ of transformations, the alignments generated by parameters $\boldsymbol{\theta}$ can be written:

$$A(\mathcal{T}, \boldsymbol{\theta}) = \{\text{aln}(u, u', \boldsymbol{\theta}) | (u, u') \in \mathcal{T}\}$$

Where $\text{aln}(u, u', \boldsymbol{\theta})$ is the set of alignments between u and u' produced by $\boldsymbol{\theta}$. $A(\mathcal{T}, \boldsymbol{\theta})$ is thus a set of sets of individual shallow alignments (indeed each pair of utterances generates its own set of shallow alignments). The fit between two such sets of sets of alignments $A(\mathcal{T}, \boldsymbol{\theta}_1)$ and $A(\mathcal{T}, \boldsymbol{\theta}_2)$ is then computed as:

$$f(\mathcal{T}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2} \sum_{(u,u') \in \mathcal{T}} \max_{((a_1,a_1'),(a_2,a_2')) \in \text{aln}(u,u',\boldsymbol{\theta_1}) \times \text{aln}(u,u',\boldsymbol{\theta_2})} \{\text{lev}(a_1, a_2) + \text{lev}(a_1', a_2')\}$$

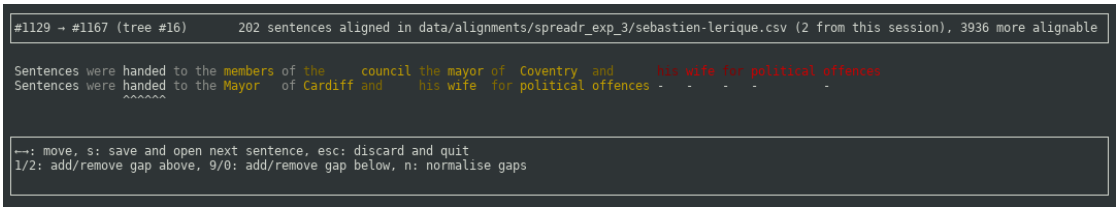
TODO: unify mathematics notations

The value of the fit thus loosely corresponds to the total number of words whose alignments would need to be changed in order to go from one set of alignments to the other. Divided by the number of transformations $|\mathcal{T}|$, it tells us the average number of word alignment errors per transformation. The worst fit $\hat{f}_n$ then gives us an upper bound estimation of the error that can be produced by training on a set of size n. One caveat in this evaluation approach is that there is no guarantee that the hand-coded alignments on which we will train could be produced by this parametrisation of alignments. We have no workaround for this caveat, other than hand-evaluation of the parameters after the training step.

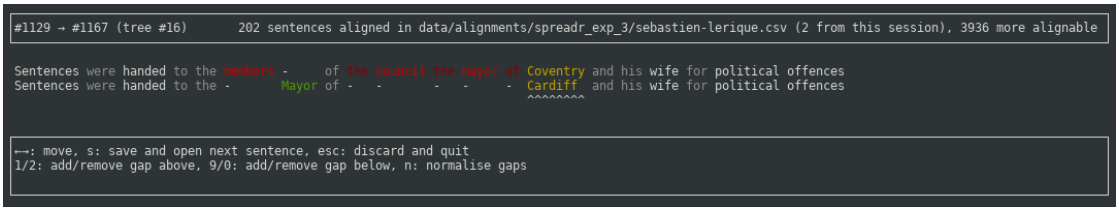
After having sampled a parameter set for step 1, we used $n = 20, 50, 100, 200$ and ran steps 2-4 ten times for each value of n. The worst values of the ten runs were $\hat{f}_{20} = 3652.5$ (1.71 errors per transformation), $\hat{f}_{50} = 1377.5$ (.65 errors per transformation), $\hat{f}_{100} = 847$ (.41 errors per transformation), and $\hat{f}_{200} = 636.5$ (.32 errors per transformation). For $n = 100$, we further resampled $\boldsymbol{\theta}^0$ ten times (step 1) and ran steps 2-4 ten times for each of those 10 parameter sets, yielding an overall $\hat{f}_{100} = 1437.5$, that is .70 errors per transformation. We conclude from this evaluation that a training set between 100 and 200 alignments is enough to reduce the final error below one word per transformation.

We thus hand-coded 200 alignments of non-trivial transformations, using a simple console interface illustrated in Fig. 3.9. The manual alignments were used as a parameter training set, on which brute forcing the best $\boldsymbol{\theta} \in [-1, 0]^3$ with discretisation step up to .025 achieved a training fit of 240 (i.e., 1.2 errors per transformation), confirming that our hand-coded alignments are most likely not possible to reproduce perfectly with this parametrisation. The final parameters obtained with this approach are $\theta_{mismatch} = -.89$, $\theta_{open} = -.29$, and $\theta_{extend} = -.12$. $\theta_{exchange}$ was then set to $-.5$ after manual trial and error.

Finally, we manually evaluated the overall quality of these parameters by hand-coding the number of errors in a random set of 100 non-trivial alignments generated by the parameters. Errors were counted as the number of words whose alignment would have to be changed in order to obtain a



(a) Before manual alignment



(b) After manual alignment

Figure 3.9: Console interface for manual transformation alignment. The user moves their cursor (underline below "handed" and "Cardiff") along the word sequences to insert or remove gaps and align the two utterances by hand.

perfect alignment. Of those 100 alignments, 79 were perfect, 12 had 1 error, 4 had 2 errors, and the remaining 5 had between 3 and 6 errors. Counting 1 error as acceptable, this method yields a successful alignment in 91% percent of the cases. To make sure this is also the case for deep alignments, we hand-coded errors in 100 random alignments for which the algorithm had explored exchanges (though not all of them are deep alignments, as it may be that the shallow alignment was the best). An error was counted for each exchange that was missing, mistaken, or should not have been present at all. Of the 100 alignments, 81 had no errors, 17 had 1, and 2 had 2 errors. The parameters obtained here were thus used for all further analyses. They are also the ones used in the example alignments discussed in the previous sections.

3.4.3 Transformation model

The alignment procedure we just presented provides a list of deep alignment trees for each transformation in the data set. At this point we have showed that such deep alignments reliably encode the details of a transformation broken down into smaller operations, thus completing the first step of our modelling approach. We now proceed to the next step: first, create an accurate visualisation of the details of transformations captured by the alignments, then derive a transformation model based on that visualisation. The step after that will then refine the model and quantify the behaviours we can measure with it.

Consensus filiations

The first step to visualise the process encoded by alignments is to reduce the possibly multiple deep alignments encoding a transformation into a single version, which we call the consensus alignment. Indeed for a given transformation $u \to u'$, a word in u could for instance be assigned to different words in u' for different choices of deep alignments. Other cases are possible, as w could be deleted

in one deep alignment and not in others, and so on and so forth. A method to resolve conflicts across multiple deep alignments is therefore required.

We adopt the following procedure to construct the consensus alignment. For a given transformation $u \to u'$, start by flattening each tree of deep alignments into a list: each branch in each tree becomes a different deep alignment, and we are left with a list of uniquely defined deep alignments for the the transformation. Now for a given word w in u, determine if it is conserved either exactly or through replacement in at least half of all the deep alignments; if so, select the child word in u' which appears in most deep alignments (i.e. the majority child); if not, the consider w deleted. Any word in u' that has no assigned parent word is then considered an insertion.

A few details are worth mentioning here. First, since a word that is stable in exactly half the deep alignments and deleted in the other half is still considered stable, the procedure sometimes maintains one or two more stable words than the deep alignments it synthesises. Such conflicts are inherent to any consensus method, and the alternative in this case would be to consider the word unstable, adding deletions and insertions instead of stabilities to the consensus alignment; we choose to favour stabilities so as not to inflate operations artificially. Second, an analogous conflict can arise when a stable word has two equally probable candidate children, or conversely a given child word has two equally probable parent words. In those cases we decide in favour of the word closest to the end of the utterance # TODO: Favour the first, so as to stop suspicion of bias in position effect., and the procedure is consistent in both directions: the consensus alignment for $u \to u'$ is the same as for $u' \to u$. Finally, a small number of cases create new single-word exchanges, as two words are assigned not the same child but different children at exchanged positions; manual inspection of these cases showed that such exchanges are consistent with the transformation. In practice, only 53 of the 3461 transformations in Experiment 3 have more than one deep alignment, 46 of which have 2 (the other seven having 3 to 6 deep alignments), such that any change here has virtually no impact on the results.

With a single consensus alignment thus constructed for each transformation, it is possible to follow the ancestry and descent of individual words through parent and child transformations in a branch. Consider for instance a toy branch $u \to u' \to u''$. A word w' in the middle utterance u' is now uniquely identified either as an insertion, or as the conservation or replacement of a parent word $w \in u$ (with or without movement due to an exchange). Continuing down the branch, w' can also be linked to its child w'' (if it was not deleted), thus creating a lineage for this specific word along the branch.

Branch and utterance axes

Fig. 3.10 represents the lineages produced by this procedure on the branches of an example tree taken from Experiment 3, whose root is the following utterance:[20]

> « At Dover, the finale of the bailiffs' convention. Their duties, said a speaker, are "delicate, dangerous, and insufficiently compensated." » (3.12)

At first glance the plots reflect the rapid shortening of utterances, and the fact that transformations are less important on shorter utterances deeper in the branches. The figure also indicates that word replacements, studied in the previous chapter with data from blogspace, are quite speckled: they

[20] This is tree #4, which is also shown in Fig. 3.4. The transition from depth 1 to depth 2 in branch #49 of the figure also corresponds to the transformation of utterance 3.9 to utterance 3.10 discussed when introducing deep alignments.

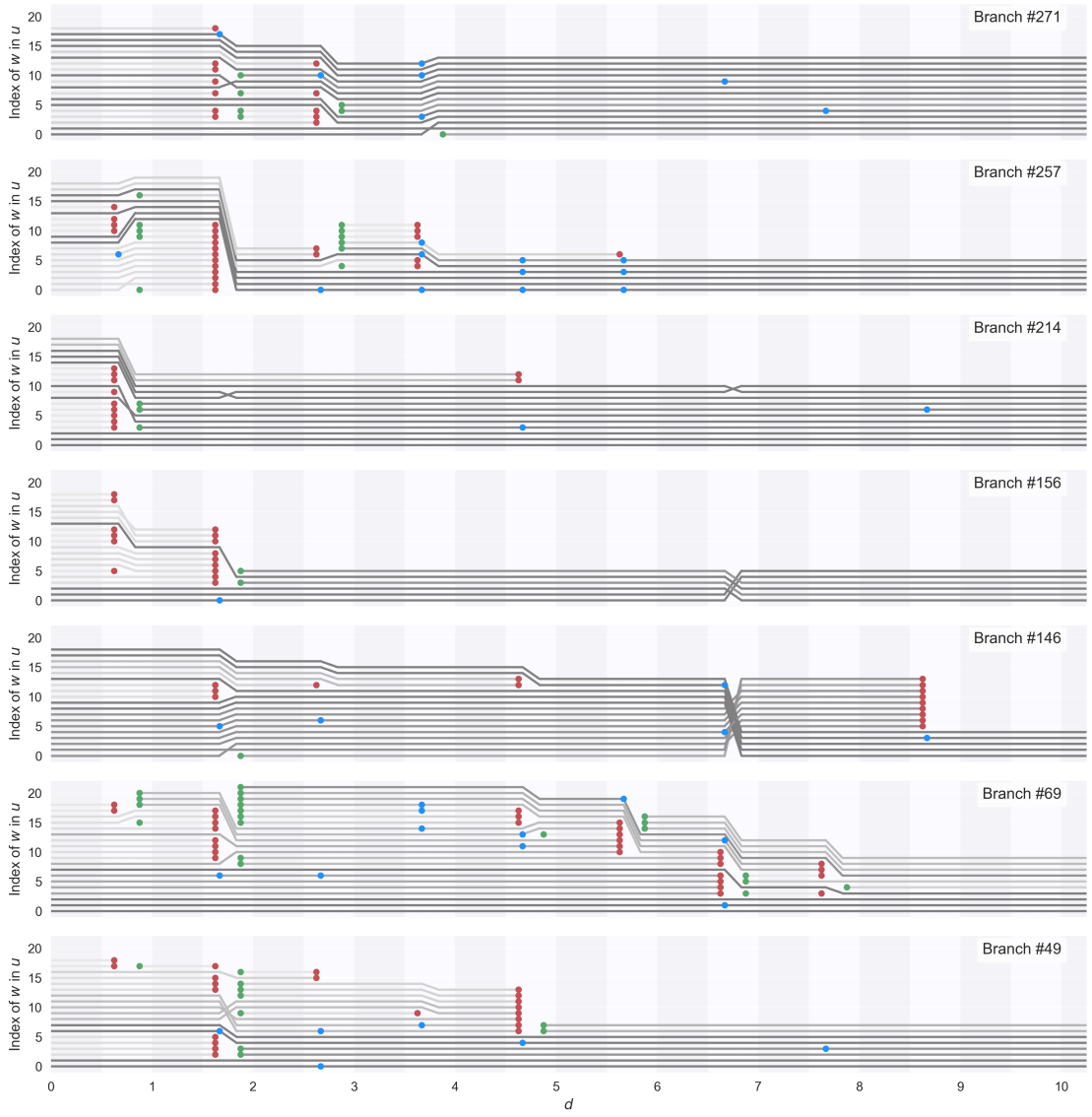


Figure 3.10: Example lineages for all the branches of tree #4 from Experiment 3. Each subplot corresponds to a different branch. The horizontal axis is the depth in the branch, and the vertical axis is the index of each word in its utterance. A grey line represents a word lineage along the branch, and the darkness of the line corresponds to the length of the path between the word's first appearance (or the branch start) and its disappearance (or the branch end); darker lines thus represent words that last longer across transformations (since branches eventually stop, however, our view of the process is truncated and the darkness is less reliable for words that appear towards the end of a branch). At each depth, the darker background band indicates what the subject sees, and the lighter band indicates the transformation that the subject made. Inside lighter bands: red dots are word deletions, green dots are word insertions, blue dots are word replacements, and exchanges can be seen when bundles of lines cross each other. Dots inside each light band are spread out on the horizontal axis so as make them easier to distinguish visually, but the horizontal position of a dot inside its band has no further meaning.

are less frequent than deletions and insertions, and affect smaller portions of the utterances when they appear. As replacements were the only process that could be extracted from the blogspace data set, suspecting this caveat was one of the motivations for our current experimental approach.

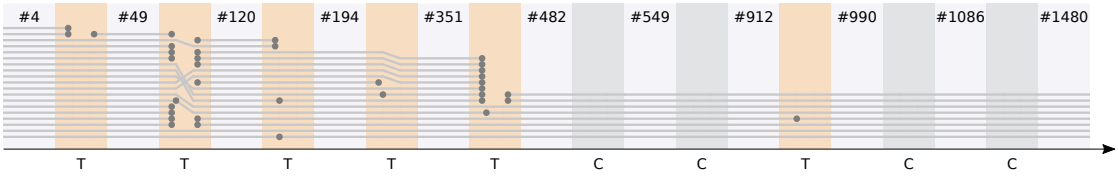
The plots also show noticeable regularities in the way transformations vary. To discuss these we distinguish between the two axes of Fig. 3.10 as two scales of the representation, each of which corresponds to a different set of events. First the horizontal axis: an event at this level is the bulk transformation or conservation of an utterance by a subject, without going into the detail of the way an utterance changes. This yields a series of conservation and transformation events, one at each depth in the branch. Call this the branch dimension, pictured in Fig. 3.11a. A salient feature on this dimension is the apparent burstiness of transformations. Since successive subjects perform transformations independently, confirming this trend would indicate a behaviour reminiscent of punctuated equilibria: a transformation occurring after a period of stability would result in a new utterance that is more likely to be transformed again. We quantify this trend further down.

Second, the vertical axis which delves into the detail of a transformation represented as a set of word insertions, deletions, conservations and replacements with or without exchange. Call this the utterance dimension. An important feature of this representation of transformations is its uniqueness. Indeed, at the mathematical level a consensus alignment encodes a transformation as a pair of word sequences with gaps (and possible sub-alignments of exchanged parts), an encoding that is not unique. Insertions and deletions that happen together can be reordered (putting insertions before their neighbouring deletions instead of the other way around, or alternating an insertion with a deletion); The exchange of two parts around a stable chunk can also be re-encoded by inverting the roles of stable and exchanged chunks, all without changing the transformation represented by the encoding. By compressing the gaps in this encoding, the utterance dimension merges these equivalent versions together and produces a unique diagram representing the transformation. We picture this correspondence between the transformation diagram in the utterance dimension and the compressed form of consensus alignments in Fig. 3.11b.

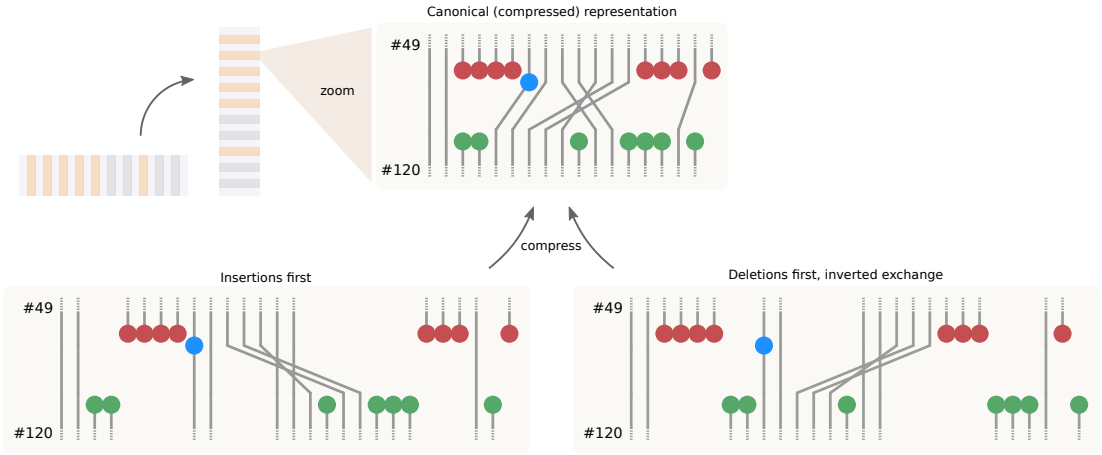
This subtlety in the encoding of transformations is not a coincidence. # TODO: this would be interesting to detail. It relates to the fact that, in spite of the one-dimensional nature of text written on a line, utterance transformation is a multi-level process that can operate on whole groups of words at a time (for instance when insertions and deletions happen together) and does not necessarily reduce to a sequential series of events. Manually inspecting the branches' transformations on the utterance dimension indicated several trends to that effect, several of which are visible in Fig. 3.10:

- Deletions, exchanges, and insertions seem bursty, that is they appear in large chunks (a behaviour that replacements do not seem to have). The bursts also seem longer if the utterance they appear in is longer.
- Insertions seem to rarely occur without deletions. When they appear with deletions, the two tend to be close to each other and of similar magnitude.
- All operations are less frequent at the very beginning of utterances.

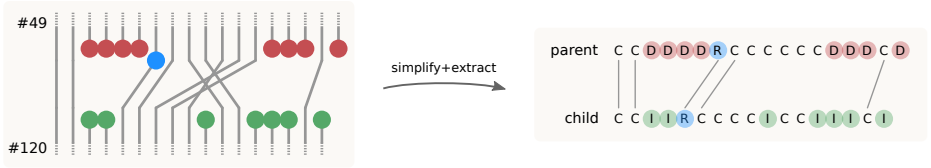
As we just noted, burstiness at the word level is no surprise: words are not processed independently and transforming parts of an utterance is likely to depend on syntactic and semantic boundaries. However, the behaviour of the bursts impose constraints on the kind of model that can account for the transformation process. In particular, to account for the possibility of insertion and deletion bursts that match in size when close to each other, a generative model would need to involve memory and attention span mechanisms that allow bursts to relate to their neighbouring operations (both preceding and following). Similarly, accounting for exchanges with a generative model also requires at least a memory component that is capable of recalling the postponed part of an



(a) Branch dimension. This level looks at whether or not an utterance is transformed, without going into the detail of changes (hence the greyed out dots and lines). Similar to Fig. 3.10, light grey bands are what subjects see, and the bands between those represent what the subjects do with what they read. An orange band indicates that an utterance was transformed, that is a T event, and a dark grey band indicates that an utterance was perfectly conserved, that is a C event. The corresponding ordered series of events is shown underneath the axis' arrow.



(b) Utterance dimension. This level looks at the detail of a transformation, and represents it with a diagram that compresses the pair of sequences produced by aligning parent and child utterances. This diagram uniquely represents the transformation, and merges any variations in encoding that can exist in pairs of sequences with gaps. The top level of the figure shows how the canonical representation comes from the lineage plots. The bottom level shows two equivalent encodings of the same transformation (as would be produced by the alignment tool), which compress to the same canonical representation.



(c) Parent and child arrays of operations. The canonical representation is further simplified by discarding the change in position encoded by word exchanges, and only keeping the information on whether a word is conserved or replaced. The procedure results in two arrays of word operations: a parent array made of conservations (C), replacements (R) and deletions (D), and a child array made of conservations, replacements and insertions (I). Conservations and replacements in the parent array, if not involved in exchanges, are linked to their corresponding operation in the child array, such that we can compute the distance between a block of insertions in the child and a block of deletions in the parent (except when exchanges separated the blocks).

Figure 3.11: Analysis dimensions. Transformations are analysed along two dimensions. The branch dimension only looks at whether utterances are transformed or not, thus sees a series of T (transformed) and C (conserved) events. The utterance level looks at the detail of the transformations, and after simplification represents them with two arrays of operations, one for the parent utterance (made of C, R and D operations) and one for the child utterance (made of C, R and I operations). This example is built on branch #49 from Fig. 3.10. # TODO: Make this work in grayscale

exchange. Such mechanisms are beyond the score of this chapter, and we therefore focus on developing a descriptive—rather than generative—model. While it will not allow for a reconstruction of the transformation process, this approach will provide a synthetic understanding of the transformation behaviour without needing to rely on cognitive mechanisms. By abstracting out the basic building blocks of transformations, we will then be able to gradually increase the level of detail with which we understand the regularities of their interactions.

Descriptive transformation model

Our model relies on a simplification of the transformation diagrams in the utterance dimension of lineage plots, which we take to be the canonical representation of a transformation. In order to keep the model palatable, we first set aside part of the information provided by exchanges. Indeed, the natural way of analysing an exchange in a transformation diagram is to see it as a permutation of a sub-sequence of words in the utterance, with possible replacements, insertions and deletions added in-between. Analysing the regularities of such a process is matter for a model in itself, and we chose to leave this aspect of transformations for further research. We instead focus on insertions, deletions, and replacements, and keep from exchanges only the conserved or replaced status of a word. Note that while this excludes any shifts in position from our model, the approach still benefits from having detected exchanges earlier in the procedure: it guarantees that the remaining insertions and deletions correspond to actual appearances and disappearances, not undetected exchanges.

From a given transformation diagram we then extract two arrays of word-level operations,[21] one for the parent utterance and one for the child utterance. The parent array contains conservation, replacement and deletion operations, and the child array conservation, replacement and insertion operations. The transformation diagram further provides us with the correspondence of conservation and replacement operations between the two arrays (except for operations that were involved in an exchange, for which we lose position information), such that we can measure the distance between two blocks of insertions and deletions (except if the two blocks are separated by operations involved in an exchange).

Fig. 3.11c illustrates this simplification of transformations, which we use as our model for the process: it represents transformation as two arrays of word-level operations, one for the parent utterance made of word conservations, replacements, and deletions, and one for the child utterance made of word conservations, replacements, and insertions. Operations that happen on several contiguous words are called chunks. Conservations and replacements in one array can additionally, but not necessarily, be paired with another conservation or replacement in the other array. When insertion and deletion chunks are separated by paired conservations or replacements, it is then possible to define the distance between the two chunks of operations as the number of conservations or replacements separating the two. When unpaired conservations or replacements separate an insertion and a deletion chunk, this distance is undefined.

3.4.4 Model refinement

Having defined our model for transformations, we now delve into the detailed behaviour that it captures. We do so in three stages. First, we quantify the extent to which transformations are bursty,

[21] We use the phrase “array of operations”, and not “series of events”, to emphasise that these operations exist on the one-dimensional utterance axis, but do not necessarily come from a sequential generation process. The two terms refer to the same mathematical object, and simply change the interpretation of the index: for a series of events the index represents time, for an array of operations it does not.

both in the branch dimension and in the detailed transformation model (utterance dimension). In doing so we establish the prevalence of operation chunks in the transformation model. We then characterise the number of individual and chunk-level operations that occur in utterances, linking their magnitude and probability to the length of the parent utterance and the position at which they occur. Finally, we examine the dependencies between each operation type, and highlighting a close relationship between insertions and deletions.

Bursty behaviours

We begin by measuring the extent to which each dimension features bursty behaviour. Following Jo et al. (2012; who rely on Goh and Barabási 2008), we measure the burstiness of a series of events through the parameter B defined as

$$B = \frac{\sigma_{intervals} - \mu_{intervals}}{\sigma_{intervals} + \mu_{intervals}}$$

where $\sigma_{intervals}$ and $\mu_{intervals}$ are respectively the standard deviation and mean of the distribution of inter-event times in the series of events. The same computation applies to arrays of operations (the two have the same mathematical description). B has values between -1 and 1; $B = -1$ corresponds to a perfectly regular process ($\sigma_{intervals} = 0$, and $\mu_{intervals} > 0$ is the constant period of events), $B = 0$ indicates a burstiness equivalent to that of a Poisson process, where the occurrence of a new event does not depend on the presence of previous events (and $\sigma_{intervals} = \mu_{intervals}$), and $B = 1$ corresponds to an asymptotically perfectly bursty process (it is the limit $\mu_{intervals}/\sigma_{intervals} \to 0$). Intuitively, a process with average inter-event time shorter than its standard deviation will often have events close to each other with a few long intervals without events, and a process with an average inter-event time longer than its standard deviation will have events more evenly spaced relative to their mean spacing.

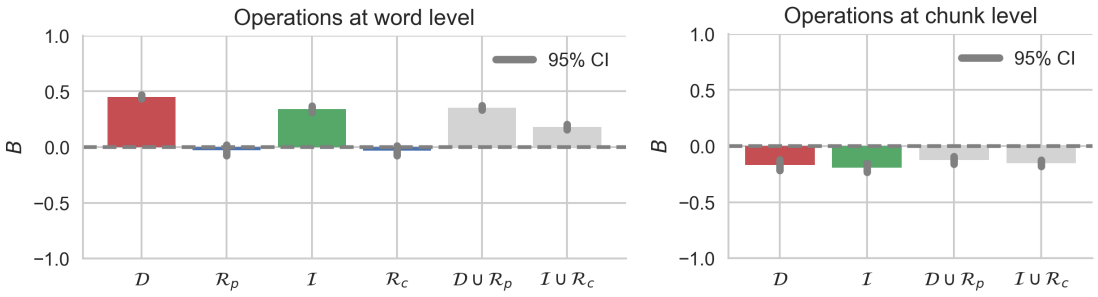
In the branch dimension, an event is the transformation of an utterance, and the absence of event is the conservation of an utterance. Note that our data in this dimension is truncated due to branches not being infinite. When the last subject in a branch does not transform the utterance they reproduce, we do not observe the actual duration of that stability: had the branch continued, the stability could have been interrupted immediately, or could have lasted for many more reproductions of the utterance. Including these truncated intervals in the distribution of inter-event times artificially inflates the burstiness (because it adds underestimated intervals to the distribution), but removing them biases our sample towards inter-event times for longer utterances (earlier in the branch), which could also inflate burstiness. We thus present measures for both distributions, with and without the truncated intervals.

Burstiness in the branch dimension with truncated intervals is $B_{branch,all} = 0.252 \pm 0.029$, and burstiness without the truncated intervals is $B_{branch,observed} = 0.304 \pm 0.031$ (both error estimates correspond to the 95% confidence interval based on Student's t-distribution, considering each tree as an independent burstiness measure). Both measures show that the transformation process in the branch dimension is significantly bursty. This is consistent with our intuition that when a transformation appears after a period of stability, it is likely to trigger other transformations following it until a new stable (often much shorter) utterance is reached.

The situation in the utterance dimension transformation model involves more event types. In the parent array, we note the series of deletion events $\mathcal{D}$ and the series of replacements $\mathcal{R}_p$. In the child array, we note the series of insertion events $\mathcal{I}$ and the series of replacements $\mathcal{R}_c$. A conserved word

is considered an absence of event. Note that because of inserted and deleted words, replacements may not appear with the same distributions in the parent and child arrays. As a consequence, $\mathcal{R}_p$ and $\mathcal{R}_c$ may not have the same distribution of inter-event times. The burstiness measures for each of these series are shown in Fig. 3.12a, along with the burstiness of the series made of all parent or child events without distinguishing their type. The plots show that deletions and insertions are both bursty, while replacements are undistinguishable from a non-bursty process such as a Poisson process. When all event types are joined together, the process is also bursty, albeit slightly less.

Given the strength of this behaviour for deletions and insertions, we further look at these series by collapsing each contiguous chunk of deleted or inserted words into a single event. This leads to a series of deletion and insertion chunks separated by word replacements and word conservations (non-events). For inter-event times, it corresponds to removing the null values in the previous distributions of inter-event times (which separated words in the same chunk); computing the burstiness of the chunk process is therefore straightforward. The values plotted in Fig. 3.12b show that none of the chunk processes are bursty; rather, they are slightly more regular than a Poisson process would be.



(a) Burstiness of word operations

(b) Burstiness of chunks

Figure 3.12: Burstiness of operations in the utterance dimension. The left pane shows the burstiness of each type of word-level operation in parent and child arrays, as well as the burstiness of the series made of all operations joined regardless of their type. The right pane shows the burstiness for deletions, insertions, and joined events, where contiguous blocks of operations are collapsed into single events. This corresponds to the burstiness of *chunks* of deletions, insertions, and joined events (i.e. only considering strictly positive inter-event times). Grey lines are the 95% confidence intervals based on Student's t-distribution, considering each tree as an independent burstiness measure. # TODO: FIXME: is it okay that here we count each tree as an independent measure, whereas in what follows we count each transformation as an independent measure?

Although this behaviour is consistent with our intuition of the way an utterance is reformulated, there is a question as to whether the alignment procedure does not favour burstiness. Indeed, the scores of operations are parametrised in such a way that insertion and deletion gaps are assigned different costs for initial opening and extension. However, while this parametrisation makes it possible for burstiness to be more easily identified, it does not make it a necessity: setting the gap opening cost to the same value as the gap extension cost would make the alignment tool neutral with respect to burstiness (setting it lower would be biased against burstiness, and the alignment algorithm would favour word mismatches over gaps to encode differences). In our case, the parameters we trained set the gap opening cost to a much higher value than the extension cost (.29 vs. .12 in absolute values), such that the alignment tool does find bursty insertions and deletions more easily. However, these

parameters are learned from hand-coded alignments and their output has been validated on test samples: any bursty insertions or deletions detected by the alignments is therefore the product of the data itself.

Position and utterance length

The general trends presented in Section 3.4.1 indicated that utterance length has a strong effect on the probability and magnitude of transformations. The transformation models now lets us explore in detail the way word and chunk operations depend on the size of an utterance, on one side, and on the position at which they occur, on the other.

We begin by looking at the probability of each operation as a function of utterance length. Fig. 3.13a plots the logistic regression of the presence or absence of deletions, insertions, and replacements as a function of the number of words in the parent utterance. The length of the parent utterance has a significant effect on all three operations, with deletions being the quickest to increase in probability, followed by replacements then insertions: the threshold for having deletions over half the time is 19 words, 22.6 for replacements, and 28.1 words for insertions; the slopes of the regressions are also ordered this way. In other words, a longer utterance will have a higher risk for all operations, and the increase is strongest for deletions, then for replacements, then for insertions.

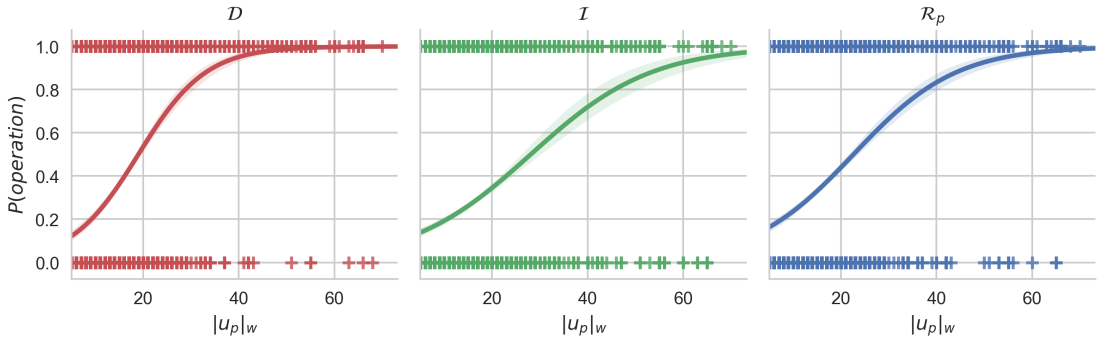
Fig. 3.13b further shows the number of operations as a function of parent utterance length, either counting one for each word affected or counting one for each contiguous chunk of words affected. The number of word and chunk operations increase close to linearly as a function of parent length. Deletions have by far the strongest link to parent length, both at the word and chunk levels, followed by insertions then replacements. Note that the replacement counts barely change between word and chunk level since this operation is not bursty: it affects mostly isolated words instead of chunks of words. In short, a longer utterance has a higher probability of suffering any type of operation, with on average over a quarter of the words deleted, the equivalent of a fourteenth of the original utterance in new words, and about a twentieth of the words replaced.

Manual exploration of the lineage plots also indicated that operations are not positioned evenly in the utterances. To quantify this behaviour we apply the susceptibility measure developed in the previous chapter to positions in an utterance. For words at position $x \in [0, 1]$ relative to their utterance's length ($x = 0$ for words at the beginning, $x = 1$ for words at the end), the susceptibility $\sigma_O(x)$ to an operation $O \in \{D, I, R\}$ is defined as the ratio of $s_O(x)$, the number of times words at relative position x are the target of operation O, to $s^0_O(x)$, the number of times those words would be the target of operation O if the choice of words were random:[22]

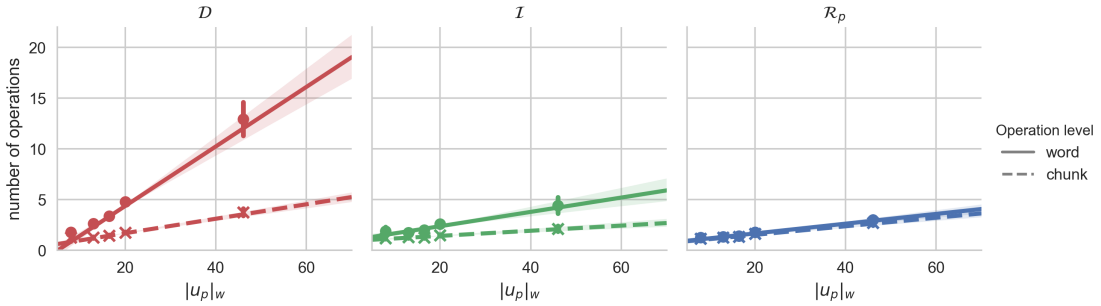
$$\sigma_O(x) = \frac{s_O(x)}{s^0_O(x)}$$

Fig. 3.14 plots σ_D, σ_I and σ_R (for replacements on the parent side) both overall and for binned parent lengths. The leftmost plots show that deletions and insertions are half as likely to appear at the very beginning of an utterance as they would at random, and more likely than random in the second half of an utterance. This is consistent with the well-known primacy effect in recall of word lists. In this case, subjects transform the beginning of an utterance on average much less than the rest. Replacements feature this primacy effect to a lesser extent, with the addition of a slight recency

[22] Since operations in a given transformation are not independent, we scale both $s_O(x)$ and $s^0_O(x)$ such that each transformation has a maximum contribution of 1 to the total counts. This procedure is similar to the susceptibility scaling approach we followed in the previous chapter.



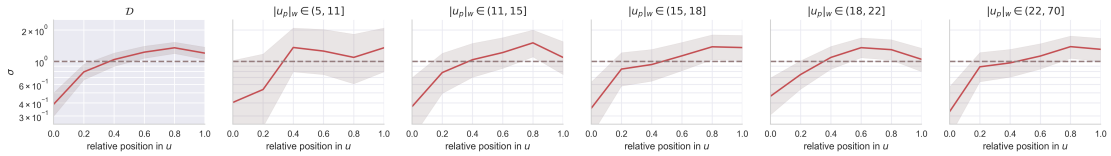
(a) Probability of word operations w.r.t. parent length, computed as the log-odds logistic regression of the presence or absence of a given operation in the transformation of u_p (parent) into u_c (child), versus the number of words in u_p. Colours correspond to the colour-coding used in Fig. 3.10. Light shades are 95% regression confidence intervals.



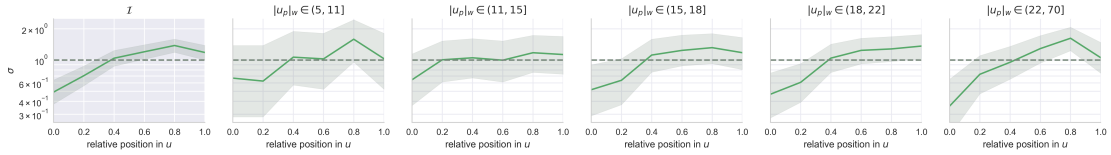
(b) Number of word and chunk operations w.r.t. parent length. Parent lengths are binned into 5 quantile-based bins. Word-level counts the number of individual words affected by an operation (deletion, insertion, replacement). Chunk-level counts the number of contiguous chunks of words affected by an operation. Light shades are 95% regression confidence intervals, and vertical bars are 95% confidence intervals for the value of a bin (Student t-based, here counting each operation as an independent measure # TODO: FXME: vertical bars should count each transformation as independent).

Figure 3.13: Probability and number of word-level or chunk-level operations. # TODO: add raw distributions?

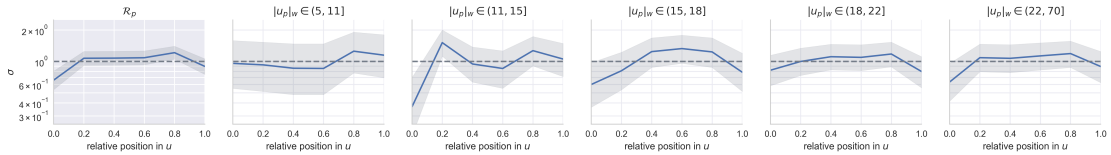
effect: words at the end of an utterance are slightly less replaced than non-extremity words. The plots at binned parent lengths show little to no variation in these patterns: each pattern is more or less marked depending on the parent sentence length (especially for replacements, which seem more uniform for short utterances), but the general behaviour is the same for different parent lengths.



(a) Deletions



(b) Insertions



(c) Replacements

Figure 3.14: Susceptibility for word operations as a function of relative position in the utterance. The leftmost plot of each sub-figure (blue background) shows σ computed over all transformations. The plots with the white backgrounds show σ computed over transformations with binned parent utterance lengths, indicated in the plot titles. Parent length bins are quantile-based, that is computed to have the same number of utterances in each bin (the bins are identical to Fig. 3.13b). Light shades are the 95% confidence intervals computed following the Goodman (1965) method for multinomial proportions, considering each transformation as an independent measure.

Finally, we examine the dependence of operation chunk size on its position in an utterance. The manual exploration of lineage plots did not hint to any effect at this level, but the question now appears legitimate: since subjects delete words on average more often towards the end of an utterance, it is possible that those deletions are also longer if they correspond to larger memory loss. Fig. 3.15 shows the dependence of chunk size on position in the utterance, for deletions, insertions and replacements, both overall and for binned parent length. Deletions exhibit a slight effect of position on chunk size, which is significant for parent lengths between 11 and 15 words.[23] That is, for those lengths, deletions towards the end of the utterance are significantly larger than deletions at the beginning (4.1 words versus 1.7 words on average), in addition to being more frequent (see the susceptibility plots above). The trend is present for deletions at all lengths, though most of the time not significative. Other operations do not seem to exhibit this behaviour (the variations for insertions are not significative).

[23] The plots also indicate that the overall chunk size increases with parent length, a slight effect which was confirmed for deletions and insertions with dedicated regressions, but which we do not discuss further given its mildness (slopes respectively .030 and .013, both significative with $p < .001$).

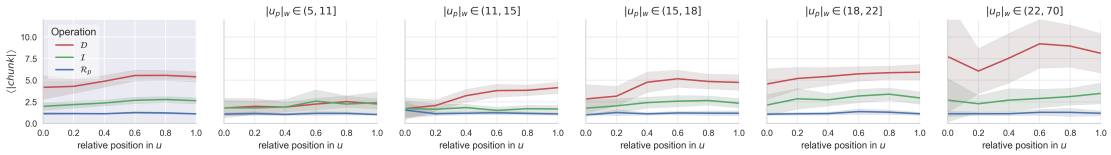


Figure 3.15: Chunk size w.r.t. parent length and position in utterance. The leftmost plot (blue background) shows the average chunk size w.r.t. parent length for all utterances. The plots on its right (white background) divide that data into binned parent lengths (bins identical to Figs. 3.13b and 3.14). In each plot, the height of a line for a given relative position x corresponds to the average size of the chunks in which words at position x are found; for instance, a deletion chunk that spans the second half of an utterance will be spread over $x \in [.5, 1]$. Average sizes are weighted such that each utterance contributes 1 unit. Light shades are the 95% confidence intervals (Student t-based, considering each transformation as an independent measure).

Overall, these measures show that deletions are more frequent than insertions, which are more frequent than replacements. Operations happen preferentially in the second half of utterances (except replacements which favour all positions except extremities), and their number of words and number of chunks are proportional to the parent length. Deletion chunks are also larger in the second half of utterances, compared to in the first half.

Dependencies between operations

Manual exploration of the lineage plots indicated that operations have non-trivial dependencies between each other. The contingency table combining the presence or absence of each operation gives an overview of these dependencies:

		Deletion			
		no		*yes*	
Replacement		*no*	*yes*	*no*	*yes*
Insertion	*no*	1381	415	286	308
	yes	66	94	399	512

Fig. 3.16 illustrates this data with a mosaic plot, rendering some of the trends more visible. One way to look at these figures is by considering deletions first. Without deletions, insertions are very unlikely (8.2%), and replacements are also unlikely (though less so: 26.0%): the most likely event without deletion is by far a transformation with no change at all (70.6%). With deletions, all four possibilities are of comparable probabilities: having both insertions and replacements is the most likely case (34.0%), followed by insertions without replacements (26.5%), then replacements without insertions (20.5%), then neither replacements nor insertions (19.0%). Overall, deletions can be seen as a gate for other transformations: without them the most likely outcome is no change at all, with them all situations have relatively similar probabilities. A second way to look at the contingencies is to consider that insertions trigger deletions: without insertions, deletions happen only 24.9% of the time, whereas with them deletions are extremely likely (85.1%). Replacements are also linked to insertions, either with or without deletions: the presence of one always increases the probability of the other.

The process is joint of course, and separating it into different stages would require more knowledge

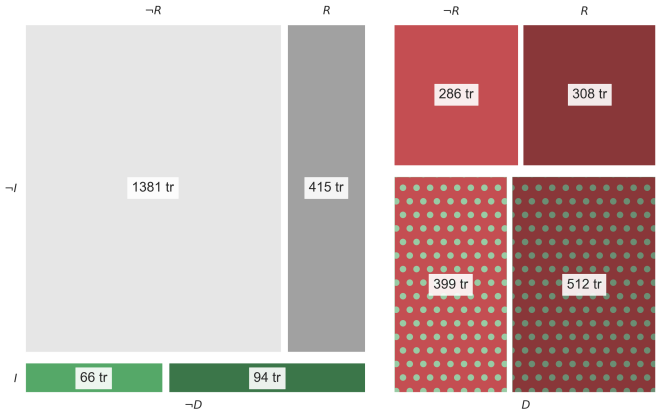
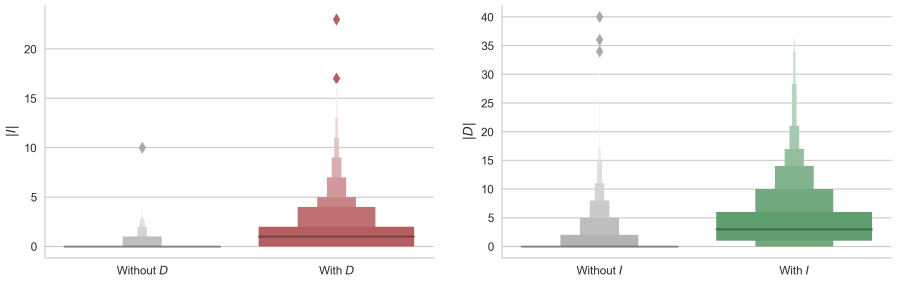


Figure 3.16: Mosaic plot of the contingency table between deletions, insertions, and replacements. Red rectangles indicate deletions are present; green rectangles or green dots indicate insertions are present; darker colours indicate replacements are present. Each rectangle also indicates the number of transformations it represents (corresponding to the rectangle area).



(a) Number of insertions conditioned on the presence of deletions.

(b) Number of deletions conditioned on the presence of insertions.

Figure 3.17: Letter-value plots (Hofmann, Kafadar, and Wickham 2011) of deletion and insertion counts conditioned on the presence of one another. In a given plot, the boundaries between boxes are placed at the $1/2^i$-th quantiles: the middle line is the median, and above and below it the biggest box stops at the first and third quartiles, the second biggest stops at the first and seventh 8-quantiles (octiles), and so on and so forth. Diamonds are outliers that do not fit into the smallest box.

of the cognitive mechanisms that underlie these transformations. In spite of this, the relationship between insertions and deletions seems to be well constrained, a fact we see not only in the probability of presence or absence, but also in the number of operations inside a given transformation. The link between insertions and deletions can be seen by plotting the distribution of the number of insertions conditioned on the presence of deletions, and vice-versa. Both plots are shown on Fig. 3.17: aside from being less probable, insertions without deletions are also much smaller in number compared to with deletions. A similar behaviour is observed in the opposite case: deletions that happen in the presence of an insertion are much greater in number than without insertions.

Deletions and insertions thus seem closely linked, as our intuition of the process suggests: deletions could be the first manifestation of the subject having forgotten something in the parent utterance, and their presence then opens the door to further reformulations, possibly to make up for the forgotten content.

This relates to the last observation produced by our manual exploration: insertions and deletions seem to occur in similar sizes when close to one another. To quantify this observation we estimate a correlation function between the sizes of insertion and deletion chunks separated by fixed numbers of conserved or replaced words. More precisely, for each insertion chunk in the data set we identify the nearest deletion chunk either before or after it, separated by words that are not involved in an exchange.[24] If an insertion chunk has such a nearest neighbour (it may not if there were no deletions, or if it occurred in the middle of exchanged words such as in Fig. 3.11c), we note r the separation between the two chunks. If insertion and deletion chunks face each other, $r = 0$; otherwise, $r < 0$ if the deletion comes before the insertion in the utterances, $r > 0$ if the deletion comes after, and $|r|$ equals the number of conserved or replaced words separating the two. For a given value of r, we compute a robust linear regression of insertion chunk size against deletion chunk size for all insertion-deletion chunks separated by r. We then take the slope of that regression as an indicator of the correspondence between the sizes of r-separated insertion and deletion chunks.[25]

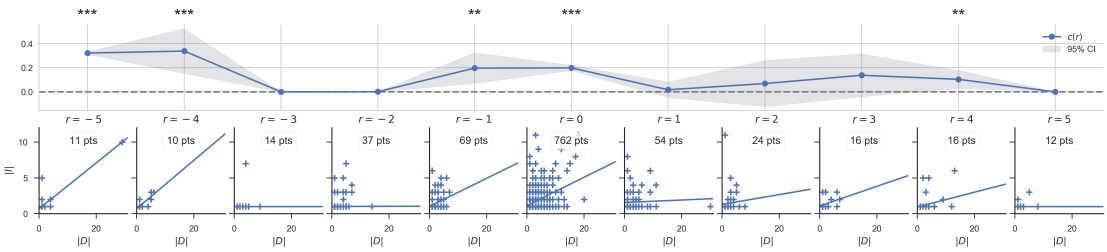


Figure 3.18: Size correlation between nearest-neighbour insertion and deletion chunks at different distances. The bottom subplots show the robust regressions for couples of insertion-deletion chunks separated by a given value of r. The text at the top of each subplot indicates the number of insertion-deletion couples that the subplot represents. The top plot shows the values of the regression slopes aligned to the bottom subplots, with 95% regression confidence intervals and star-coded significance levels (*** for $p < .001$, ** for $p < .01$, * for $p < .05$ and nothing otherwise).

[24] As alluded to when introducing the transformation model, when exchanges separate an insertion chunk from a deletion chunk there are several paths from one to the other, depending on when one traverses the exchange; different paths can have different final distances, none of which are more or less plausible than the others. The distance between an insertion chunk and a deletion chunk separated by an exchange is thus not clearly defined.

[25] The robust regression lets us minimise the impact of outliers in the distributions of insertion chunk sizes, which otherwise had a strong effect on more common correlation measures. The regression is computed using the Statsmodels statistics library for Python, which implements robust M-estimation using Huber's T norm (Huber 1981) with a default parameter of 1.345.

Fig. 3.18 shows the robust regressions and the estimated correlation function for $r \in \{-5, ..., 5\}$ (outside of which there was always less than 10 insertion-deletion couples). The plot shows three important points. First, the vast majority of nearest neighbours insertion and deletion chunks face each other ($r = 0$), and their sizes significantly correlate. Second, the correlation initially decreases to become non-signifcant as $|r|$ increases. The third and most interesting point is that the correlation function is skewed towards the left: it is significantly above zero for $r = -1$ but not for $r = 1$, then also for $r = -4, -5$ at higher values than for $r = 4$. Note however that the last three points represent only 10 to 12 insertion-deletion couples each and are thus more susceptible to outliers (especially to deletion outliers, i.e. the x axis, which the M-estimation technique we used does not counter). Overall, the correlation is positive for insertion and deletion chunks facing each other, and also often for deletion chunks preceding insertions by a few words.[26] This trend is consistent with the intuition we outlined above, according to which insertion chunks could come as tentative replacements for the content that was lost in the deletions that directly precede them.

The transformation model we introduced thus captures several important behaviours in the way subjects change utterances. Looking at transformations as made of word-level replacements, deletions and insertions, we see that both insertions and deletions are bursty, and that the presence and magnitude of an operation depends strongly on utterance size and the position at which it appears in the utterance. We further see that insertion and deletion chunks are closely related: insertions behave as if they were gated by the presence of a deletion, and their sizes tend to correlate to that of deletions appearing at the same time or shortly before them.

3.4.5 Lexical feature makeup

We finally descend to the lower level of lexical word features to characterise the words involved in insertions, deletions and replacements. We thus extend the feature analysis of the previous chapter to our current situation, and verify the consistency of the new results with previous word susceptibility and feature variation measures. Finally we extend the analysis to a point we could not reach in the previous data set, by looking at the accumulation of transformations along the branches and the evolution they cause in the lexical makeup of utterances.

Word features

For the sake of conciseness, we restrict features to the four lexical variables that showed relevant effects in the previous chapter: word frequency, age of acquisition, Free Association clustering and number of letters, thus leaving aside number of synonyms and orthographic neighbourhood density. Age of acquisition, clustering and (obviously) number of letters are identical to the previous chapter. Word frequency was previously computed from the complete set of online quotations; however, since the present data set is much smaller we relied instead on external word frequency ratings based on subtitles (Heuven et al. 2014), a source which has repeatedly beaten previous predictors of standard lexical decision times (see Heuven et al. 2014 for more details). These frequencies are provided on what the authors introduce as the Zipf scale, computed as $\log_{10}$(Frequency per billion words). The frequency values thus use a different source than those of the previous chapter, but their final computation only differs by an affine transformation.

[26] The detail of these plots is sensitive to cropping in the data, and especially to constraints on the maximum deletion size since it can remove x-axis outliers. The general trend we observe is always conserved however: deletions preceding insertions correlate more than deletions following insertions.

The situation is otherwise parallel to previously, and its procedure can be directly applied. We measure the susceptibility of words to being the target of an operation (either by deletion or replacement) in a similar manner to substitution susceptibility. We additionally measure the susceptibility to being the new word of an operation (either as replacing word or inserted word). For a given grouping of words g (e.g. grammatical category or feature value), we compute its susceptibility σ_g^- to being a target and its susceptibility σ_g^+ to newly appearing as the ratio of the number of times it is a target (s_g^-) or a new word (s_g^+) to the number of times it would be if the process were a random sampling from the available utterances (s_g^0):

$$\sigma_g^- = \frac{s_g^-}{s_g^0} \quad \text{and} \quad \sigma_g^+ = \frac{s_g^+}{s_g^0}$$

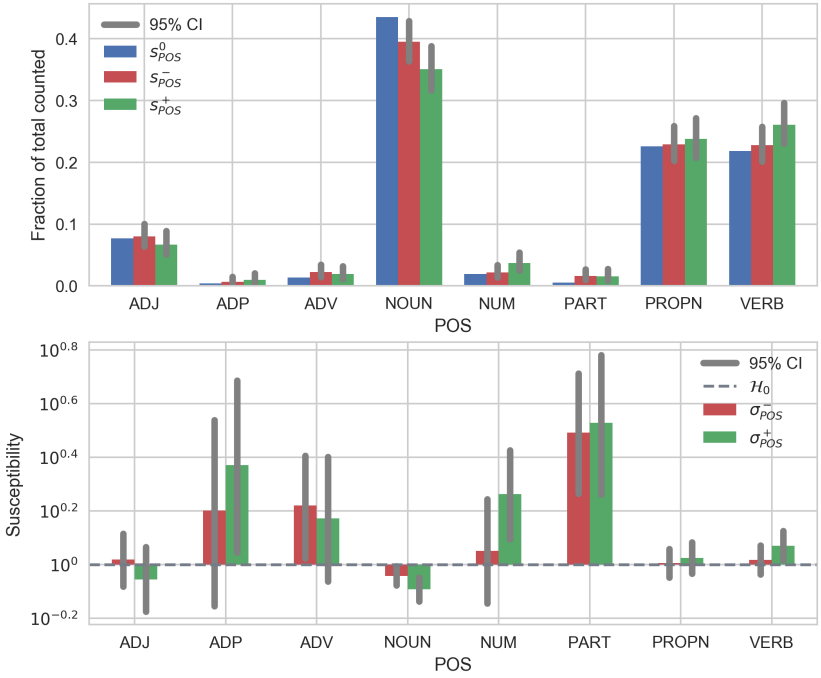
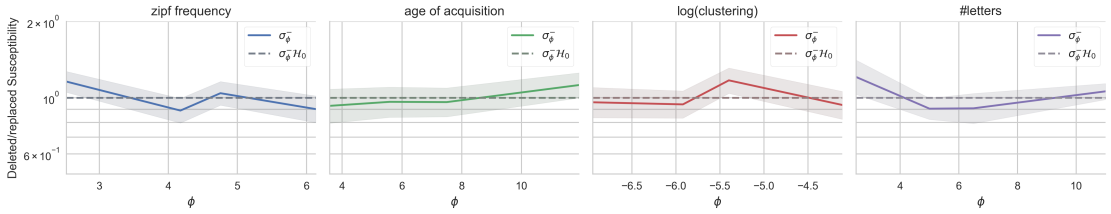


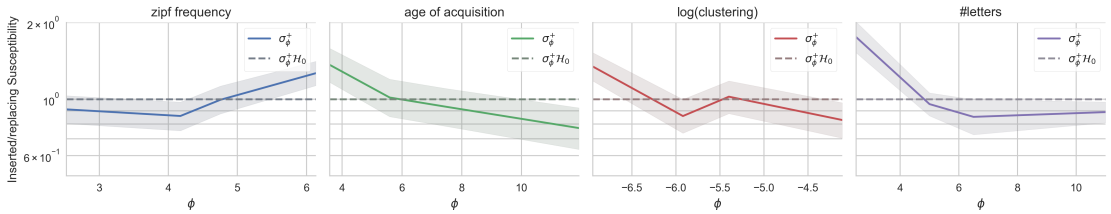
Figure 3.19: POS susceptibility to targeting (deletion and replacement on the parent side) and to appearance (insertions and replacement on the child side). The top panel shows the proportions of POS categories observed in utterances overall (s^0_{POS}), in targeted words (s^-_{POS}) and in appearing words (s^+_{POS}). The bottom panel shows susceptibilities, that is the ratio of s^-_{POS} and s^+_{POS} to s^0_{POS}. 95% asymptotic confidence intervals are shown in grey (Goodman-based multinomial proportions, considering each transformation as an independent measure). POS tags are from the Universal Dependencies tag set.

In order to render the results more comparable to the previous chapter, in this section we also filter out stopwords in all the utterances. Fig. 3.19 shows POS susceptibilities for being the target or the new word of an operation. The two measures are very close to each another, and similarly to the online case there is little to no effect of the main categories on susceptibility: adjectives are involved at random, nouns appear slightly less than at random, and verbs slightly more. Verbs, nouns and proper nouns are all irrelevant for targeting, but adverbs are targeted very slightly above random.

The other categories (adpositions, numerals and particles) total negligible amounts because they are affected by the stopword filter. Overall, the behaviour for targeting is consistent with what we observed in blogspace (the only difference being the trend for adverbs, which are also less present overall in this data set), and the behaviour for appearances indicates a slight bias in favour of verbs and against nouns (note that appearance susceptibilities were not analysed in the blogspace data set, so we have no point of comparison).



(a) Susceptibility to targeting



(b) Susceptibility to appearance

Figure 3.20: Feature susceptibilities of words to targeting (deletion and replacement on the parent side) and appearance (insertion and replacement on the child side), binned by quartiles, with 95% asymptotic confidence intervals (Goodman-based multinomial, considering each transformation as an independent measure).

Fig. 3.20a plots the susceptibilities for targeting and appearance for word frequency, age of acquisition, clustering and number of letters. The trends for the first three are consistent with previous results: low frequency, high age of acquisition words tend to be very slightly more targeted, and clustering is mostly not relevant to the process. Number of letters has a different behaviour than previously, as short words are slightly more targeted than random, beyond the effect for long words. At this stage it is unclear where this change of effect comes from, as it could be due to the choice of source utterances, or to the fact that subjects could be more inclined to replace some words because of a different task context (the feature variation analysis, further down, gives more insight into this change). Fig. 3.20b shows the corresponding feature susceptibilities for appearance, where the trends for frequency and age of acquisition are reversed: more frequent, lower age of acquisition words are more susceptible to appearance. Low clustering and short words appear also more than random.

Note that while we cannot produce variation measures for deletions and insertions (but we do for replacements further down), comparing the targeting and appearance susceptibility curves for each feature gives an idea of the effect of transformations: targeting susceptibilities indicate what kinds of words are preferentially picked for transformation, and appearance susceptibilities indicate what kinds of words appear instead (although not necessarily in one-to-one replacement). Thus transformations preferentially remove low frequency, high age-of-acquisition words, and preferentially add high frequency, low age-of-acquisition and low clustering words. The case of number of letters

is interesting, as transformations remove more short words than long words, but also insert more short words than long words; however, the difference is stronger for appearance than for targeting, such that the final effect should be in favour of shorter words (a fact we confirm below). All these trends are also consistent with the variation patterns observed previously.

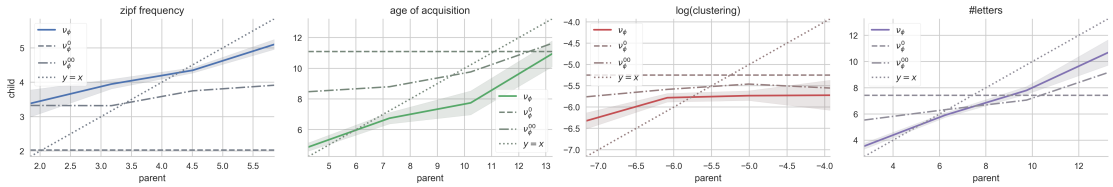


Figure 3.21: Feature variation upon replacement. ν_ϕ, average feature value of the appearing word as a function of the feature value of the targeted word (fixed bins), with 95% asymptotic confidence intervals based on Student's t-distribution. Refer to Fig. 2.7 (# TODO: ref to BCP variation figure) for the detailed interpretation of the curves.

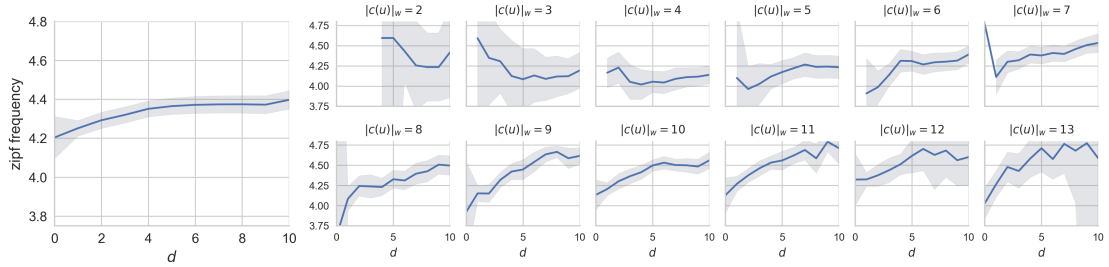
Our previous analysis of feature variation can also be directly applied to word replacements (though not to deletions or insertions), and Fig. 3.21 shows the results for the current data set. The plots for frequency, age of acquisition and clustering are strikingly similar to previous results. The $\mathcal{H}_{00}$ curve (ν_ϕ^{00}) is slightly changed for high age of acquisition, as it gets much closer to and eventually crosses the first null hypothesis, ν_ϕ^0. The variation curve (ν_ϕ) shows the same uniform negative bias w.r.t. both null hypotheses, but its attractor point has moved to 6. Clustering has changed very little, and the attractor point is at -5.75. Here too however, number of letters has a different behaviour than previously: ν_ϕ and ν_ϕ^{00} are substantially changed and do not show the previous uniform negative bias: both are much closer to word conservation ($y = x$) than previously, and ν_ϕ crosses both ν_ϕ^0 and ν_ϕ^{00}. In other words, word sizes are better conserved in this data set than in blogspace. Nonetheless, the process still features a slight negative bias with an attractor point close to 5 letters. Two factors could have influenced this change of effect: first, the alignment procedure favours replacements for closely related synonyms (evaluated by their vector similarity), which could explain the fact that ν_ϕ and ν_ϕ^{00} are much closer to each other and to $y = x$.[27] Second, the fact that ν_ϕ^{00} changes so much from the its values in the previous chapter indicates that the sampling of source utterances also has a role. Recall that in this case ν_ϕ^{00} is the average word length of the synonyms of replaced words: ν_ϕ^{00} being closer to $y = x$ then indicates that synonyms of words in the current utterances are closer in size to their originals than is the case in the blogspace utterances, a fact that could contribute to the overall better conservation of number of letters.

Branch evolution

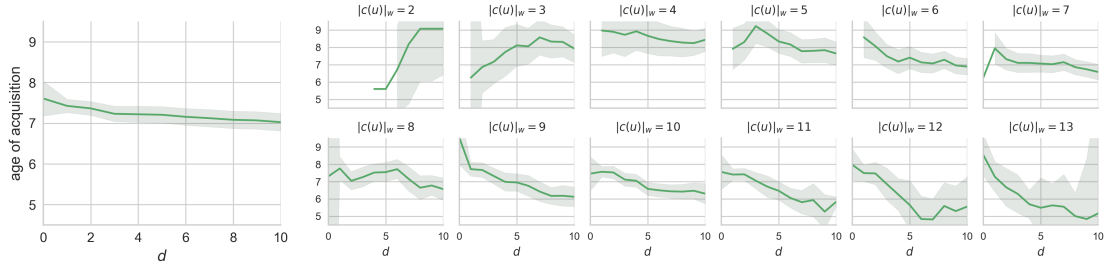
Beyond the confirmation of previous findings, we are now in a position to observe the evolution of lexical features along the branches, and relate any trends to the step-wise transformations. This question could not be answered in the blogspace data set for lack of detectable chains.

Fig. 3.22 plots the evolution of the average features of utterances as a function of branch depth, both for all utterances and divided into fixed content lengths. The evolution of each feature is consistent

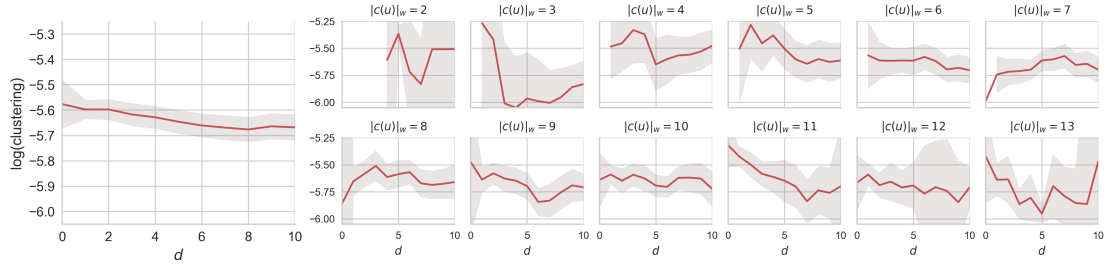
[27] The role of synonym replacements could also explain the change of susceptibility for short words observed in Fig. 3.20a. Indeed, separating susceptibility plots for deletions and replacements on the parent side shows that short words have a susceptibility to replacement higher than random (whose effect is seen in the aggregate figure), but not to deletion. The behaviour could therefore come from shorter words having more synonyms, or more frequent synonyms.



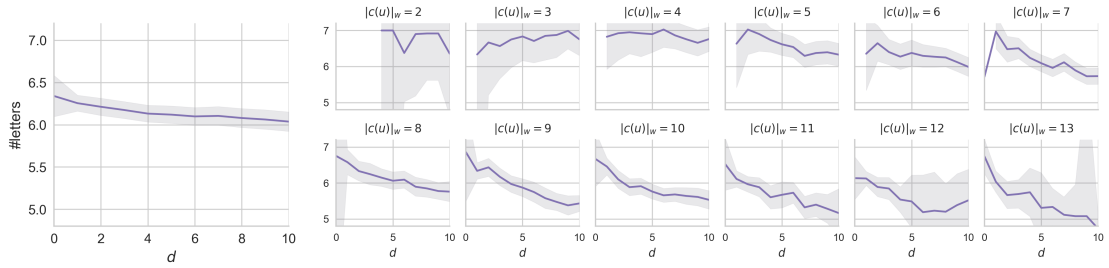
(a) Word frequency



(b) Age of acquisition



(c) Clustering



(d) Number of letters

Figure 3.22: Evolution of average utterance features as a function of depth in the branch, with 95% confidence intervals based on Student's t-distribution (considering each utterance as an independent measure).

with its susceptibility to targeting and appearance, and its variation upon replacement. Average word frequency significantly increases with depth, both globally and at fixed content length. This is consistent with low frequency words being more susceptible to targeting and high frequency words more susceptible to appearing (Fig. 3.20), as well as with frequency increasing upon replacement (Fig. 3.21). The reverse is true for age of acquisition, which decreases with depth (albeit significantly for certain content lengths only). Clustering and number of letters both decrease also, though the clustering trends at fixed content lengths are less uniform than for the other features. It is worth noting that for number of letters, in spite of the targeting bias in favour of short words, the much stronger appearance bias in favour of short words wins in the long run: average number of letters decreases along the branches, even at fixed content length. We also note that the asymptotic values (if we may, after 10 transformations) for each feature do not correspond exactly to the attraction values for replacements; this could be due to the fact that deletions and insertions are much more important in magnitude and are likely to change the exact attractor points, or to finer interactions (for instance in semantics) which the averaged variation and susceptibility plots we presented do not capture.

Nonetheless, it is noteworthy that trends consistent with the step-wise behaviours are visible at the level of utterance averages: in less than 10 iterations, transformations which mostly maintain the overall meaning of the utterances have a significant effect on these features, beyond the shortening of utterances (and consequent removal of words that could have an effect on the features). Through transformations, subjects thus gradually evolve the utterances to use more frequent, shorter words, learned earlier and with lower free association clustering coefficients. We reserve the discussion of this phenomenon for the next section.

3.5 Discussion

ADD: refs to substantiate; this should be easier once the final discussion chapter is clear

We set out to better understand the process at work in the short term evolution of linguistic content. In an approach complementary to the previous chapter, we decided to design a controlled experiment that would provide the complete data needed to develop a model of the process. We developed an online platform for the purpose, and after adjusting task difficulty and source complexity we were able to gather relatively large data sets of linguistic transmission chains with low levels of spam. Then, by combining standard NLP methods with our extension of a biological sequence alignment algorithm, we decomposed the utterance transformation process into small, analysable operations that subjects use in their reformulations.

A few important points are worth noting to qualify the results we just presented. First, several choices in the alignment procedure we followed are obviously sub-optimal, and were made in the interest of time. The Needleman-Wunsch algorithm we used does not allow the detection of chunk replacements. For instance, abbreviations or short paraphrases, which Lauf, Valette, and Khouas (2013) identify as non-negligible in their data set (e.g. “has no idea” → “doesn’t know”), are not captured by our approach. Extending the algorithm to allow this is technically possible, but involves a substantial amount of additional work. The algorithm is also blind to syntactic boundaries such as punctuation, which insertions and deletions are likely to respect at least part of the time. Manual inspection of the alignments showed a few cases where a deletion would affect two contiguous parts of an utterance separated by a comma, for which distinguishing the parts could help in improving the final alignment. Finally, deep alignments are only explored on the basis of optimal shallow alignments, which are not guaranteed to be the best starting point: it is possible that the search will find a

locally optimal alignment, when a better solution could have been found by starting from other (non-optimal) shallow alignments. There are many other possible improvements that we did not explore fully for lack of time, for instance matching insertion-deletion blocks from exchanges at different depths to overcome local optima, or starting with local instead of global shallow alignments. The manual evaluation of alignments indicated that the current approach was good enough however, and that these optimisations could be left for further research without jeopardising our model.

Concerning the design of the task, we note that the incentive provided for subjects to perform well also leaves room for improvement (a point related to the spam levels). Subjects had a monetary incentive with bonuses for outstanding performance, but they did not experience the bonus until after the experiment was completed. More generally, the setup puts participants in the position of a subject and not of a user: setting aside what the interface encourages, there is no intrinsic incentive for people to put extra effort into the task. As Gauld and Stephenson (1967) phrased it already long ago: "Errors could, it seemed, be avoided, if the subject was so inclined." However, analysing the transformation rates of individual subjects showed nothing to that effect: although some subjects average better than others, it seems that no individual subject is uniformly good or bad at the task. Initial explorations of the word span of subjects (in Experiment 1) also showed little to no correlation to subject performance. The best solution to this problem would be to create an intrinsic motivation for the participants by aligning their interests with performance. Claidière, Trouche, and Mercier (2017) implement this kind of incentive, by asking participants to defend and convince others of a choice they have previously made. In our case, such an incentive could guarantee subjects' involvement but would not suffice to improve the quality of the written productions, as people will very easily understand badly edited text in the course of a live interaction (this would make the computational analysis even harder).

Finally, our setup entirely obviates the question of the context in which utterances are processed and reproduced. This was a deliberate choice, as we decided to use the simplest transmission chain task possible and reserve the introduction of more complexity (such as context effects) for later research. As a consequence however, we have no control over the situation in which subjects read an utterance, nor did we add contextual paragraphs or preceding utterances to examine framing or priming effects on the interpretations and transformations of subjects. Preceding text is very likely to have effects on the transformations, as these are reliably observed in the study of intrusions in recall of word lists (Zaromb et al. 2006). Manual exploration of the data also showed (rare) cases of words from one utterance bleeding into later transformations: in one case, a subject reintroduced a word that they had read in an utterance three trials earlier. This phenomenon is difficult to control for beyond the randomisation we applied, and we note that the cases we observed were extremely rare.

In spite of these caveats, we showed that transformations can be usefully analysed as made of bursty deletions and insertions, speckled with word replacements (and exchanges, which we left aside in the analysis). Deletions are by far the most frequent operation, and they act as a gate for insertions; in turn, the size of insertions tends to correlate to the size of deletions that they closely follow. We further observed that all operations are less probable at the beginning of an utterance, as well as in shorter utterances, and that deletions tend to grow in chunk size as well as in chunk numbers towards the end of an utterance. Finally, we observed that transformations are also bursty at the level of the branch, suggesting that the process follows a punctuated equilibria pattern: when a subject makes a transformation on a previously stable utterance, the next subjects in the chain might add transformations until the changes are regularised into a newly stable utterance.

Overall, we suggest that adopting this descriptive model provides a clearer picture of the process at work in the evolution of linguistic content than has been previously achieved. The model creates an intermediary scale between the detailed level of lexical word features and the high level of contrasts

in aggregated evolution, thus rendering the process more intelligible. In particular, we believe that visualisations such as the lineage plots we presented are extremely helpful in identifying the underlying mechanisms that can be then connected to known effects in linguistics. In the context of Cultural Attraction Theory, this type of approach could prove useful to construct more parsimonious models of the evolution of representations. In particular, it brings detail to the linguistic instantiation of the wear-and-tear and flop problems introduced by Morin (2016): the first could correspond to the way utterances are gradually transformed by replacements, exchanges, and insertions making up for deletions; the second could correspond to downright mass deletions in a transformation.

Transposing the analysis from the previous chapter to the current data set also confirmed the trends in lexical features observed in blogspace: less frequent, longer words that are learned late and have higher clustering coefficient are on average replaced by higher-frequency, shorter words, learned earlier and with lower clustering coefficient. As we discussed in the previous chapter, these words are overall easier to produce in standard naming or word recall tasks. Examining the evolution of these features along the branches also showed that the process significantly transforms utterances to use easier words on average: transformations can thus be seen as creating a gradual drift of utterances at the low-level of lexical features due to a cognitive bias in favour of certain word types.

While one could consider this phenomenon as relevant to Cultural Attraction Theory, it remains extremely low-level and does not indicate any consistent drift or attraction in the semantics of the utterances (nor does it invalidate it). Manual inspection of the data also gave no sign of an attraction phenomenon at the semantic level. Two points might be noted nonetheless. First, the details of transformations seem to follow the patterns identified by Lauf, Valette, and Khouas (2013) in news stories: complements, adverbs, modals, and more generally any details not essential to the main meaning seem often deleted or replaced. Second, examining episodes of bursty changes also suggested that there is a chaotic aspect to chains: relatively minor changes in the middle of a transformation sometimes lead to a comparatively larger change in meaning down the branch, as an ambiguity is created and resolved differently than previously. In Experiment 3 for instance, a subject presented with the following utterance,

> "A dozen hawkers who had been announcing the news of a non existent bomber in Kings Cross have been arrested." (3.13)

introduced several changes including a typographical error that replaced the word "news" with "new"; he or she produced the following utterance:

> "A dozen hawkers announcing non existent **new** about a bomber at Kings Cross have been arrested" (3.14)

The "news" $\rightarrow$ "new" change, while superficially minor compared to the rest of the transformation, is quite important once we remove the context provided by the previous utterance. Indeed, it later went through a long regularisation such that the final utterance of the branch read:

> "A dozen hawkers, at New Kings Cross have been arrested." (3.15)

Examining the transformations in the branch suggested that the small typographical error rendered its surroundings ("non existent … about a bomber") confusing and irrelevant, such that "new" was

finally integrated as part of the "Kings Cross" proper noun instead. This behaviour is not frequent, as many times typographical errors are corrected by subsequent transformations, but it appears to be possible whenever an ambiguity is created or enhanced (not only through typographical errors).

TODO: unpack some of the discussion points and suggestions below

Other intriguing semantic effects were observed. In one case for instance, small changes that accumulated in different parts of an utterance ended up combining into a larger semantic change, because the relationship between parts of the utterance had eventually changed. More broadly, tackling the question of semantic attraction (or more simply semantic transformations) could require the definition of semantic levels at which the transformations should be examined. As the chaotic behaviour described above illustrates, changes at the semantic level can be created by any ambiguity that is picked up by subjects. This can range from a typographical error, to a change in punctuation, or a minor change of vocabulary which seems ambiguous to one subject but not to another. The question is thus shifted from the structure of utterances themselves to what subjects attend to, in an utterance, for a given task or in a given interaction situation. In other words, analysing the semantic changes of utterances goes hand in hand with a determination of what aspects of an utterance are relevant for a given task or interaction, that is, it requires an approach to utterance pragmatics.

Without delving into utterance pragmatics, a more palatable development of this approach would be to ask about the role of simple semantics and syntax in the transformations: beyond lexical features and word categorisations, one could attempt to quantify and thus characterise the change in meaning at each transformation and overall through the chains, possibly through deeper integration with existing NLP methods. Conversely, the extent to which word meanings (or word relationships with the rest of the utterance) participate in the transformation process could be explored. Better integrating these results with what is known of the way utterances are cognitively processed and produced could also be fruitful: known mechanisms in sentence parsing and processing could explain the patterns observed by our model, and integrating with current knowledge of sentence production could make it possible to develop generative models of the transformation process. Finally, many questions from the last chapter also remain pending. In particular, if context is an important factor in the transformations observed, we wonder about the role of feedback effects and path dependence in the evolution of content online: how much are transformations determined by the distribution of utterances that readers are exposed to, in what way, and how does this influence feed back into the distribution itself? An important role of feedback in the transformations would be grounds for a strong path dependence in the evolution of linguistic content, maybe even at lower cognitive levels.

Answering these more approachable questions would already provide a much more complete view of the dynamics at work in the short-term evolution of linguistic content. As noted above however, it is likely that further progress in this area will require considering the role of pragmatics in utterance transformations, for instance by exploring more ecological interaction situations than the simple read-and-rewrite task we used.

Chapter 4

Discussion

4.1 Introduction

In this chapter, we aim to take a broader view on what would be necessary to achieve a fuller understanding of the processes at work in cultural change at the linguistic level. So far we have adopted wholesale the paradigm put forward by Cultural Attraction Theory, by seeking to identify and elucidate situations where linguistic representations are transformed as they are transmitted, and assessing, on one side, the extent to which the empirical evolution of content agrees with what is expected under CAT, and on the other side, the extent to which CAT provides productive guiding questions in understanding what is at work in the situations studied. This has led us to identify a number of behaviours which are consistent with Cultural Attraction Theory: studying word substitutions in online quotations first, and more general transformations in controlled transmission chains of short utterances second, we showed that the low-level lexical features of words evolve in a systematic manner to make utterances easier to produce, and that the direction of the evolution is consistent with the attraction pattern that can be observed in the individual step of word replacements. However, these approaches did not bring us any closer to understanding the semantic changes that utterances undergo when they are transformed, be it online or in controlled transmission chains.

We now wish to discuss the reasons for this limitation. Our purpose is first to convince the reader that meaning[1] is a crucial aspect in the evolution of content which must eventually be analysed in order to fully understand the way representations circulate and change. Manual exploration of the changes in transmission chains in particular show that the surface measures that we used in quantitative analysis have no handle on the evolution of meaning. Indeed, meaning will appear as a deeply context- and interaction-dependent property, which cannot be understood by simply focusing on the utterances themselves. Second, we aim to show how this challenge can be traced to what is known in philosophy of mind as the "hard problem of content", and to how approaches to pragmatics deal with it. We will thus discuss two important approaches to studying the meaning of utterances in relation to the context and interaction in which they appear: Relevance Theory and the Enactive approach. The first is better developed and integrated with linguistics, but is complex to implement and must face some version of the hard problem of content in order to provide a complete account of meaning. The second starts from a simpler endogenous notion of meaning

[1] In what follows, we will always assume that meaning is meaning *to someone*. In other words, meaning to a listener is the listener's interpretation, and meaning to a speaker is the meaning intended to be communicated.

which avoids the problem of content, but has yet to prove its viability and usefulness for the study of language. Favouring one or the other, or possibly combining parts of the two, is further related to the overall construal of cultural evolution and to the importance of representations in a theory of cultural change, as critiques of the cultural attraction framework have shown. Finally we believe that the question of meaning in cultural change, and the debates it relates to, can be moved forward through informed empirical investigation. After having laid out the alternatives, our final goal is therefore to put forward the approaches we believe are most useful to turn this problem into an empirical question.

We begin by discussing detailed examples of the role of semantics in our transmission chain experiment, to show how the lack of an account of utterance meaning renders the empirical question of attractors in this case under-specified. Next, we present in more detail two possible approaches to pragmatics and meaning, and discuss their relationship to an overall view of cultural change. Finally, we present possibilities for refining and advancing the debate through empirical investigation.

4.2 Empirical epidemiology of linguistic representations

4.2.1 Relevant results

The path we took so far has consisted in entirely adopting the cultural attraction paradigm and developing experiments to evaluate one of its strong hypotheses, namely the existence of attractors in the evolution of representations. Indeed, cultural attractors are in many ways a cornerstone for the theory, as they reflect its explanation of the stability of culture in spite of strong micro-level transformations (they are the product of ecological and psychological factors interacting with each other), and they provide intelligibility into the complexity of cultural change as a whole. Linguistic utterances appeared as a good proxy to study representations that are part of everyday life and for which large corpora are readily available. Language is also one of the most versatile means by which representations circulate, making linguistic utterances an important study case for the theory.

Our initial high-level question was thus whether attraction could be observed in the evolution of linguistic utterances as they are interpreted and produced anew by successive people. The first case-study we developed relied on online quotations, a type of representation for which an implicit rule mandates perfect copy, yet which often changes as it propagates across blogs and news outlets. Our investigation of single-word replacements showed that, when transformed, words are reliably replaced by words easier to produce. Evaluated on standard lexical features, individual word replacements showed an attraction pattern specific to each feature and consistent with the hypothesis of an attractor at the lexical level, which could be due to cognitive biases in word production. Our second case-study explored utterance transformations in a more controlled situation, by setting up artificial transmission chains of short utterances on an online platform. Here, the analysis first focused on developing a descriptive model that would provide an overview of transformations decomposed into more basic operations. We showed that transformations can be reliably described as made of chunks of word insertions and deletions interspersed with individual word replacements (as well as chunk exchanges, which were set aside for the analysis). This level of description showed that the transformation process has several regularities: operations strongly depend on each other (in particular, insertions appear to make up in size for some of the deletions, while still introducing substantial change), and their prevalence also depends on the length of, and their position in, the utterance (longer utterances receive more operations; replacements preferentially target the interior of utterances, and insertions and deletions the second half of utterances). The behaviour of insertion

and deletion chunks, as well as replacements, was shown to be consistent with the biases identified in individual replacements in online quotations: the susceptibilities for being targeted by deletion or replacement, and appearing by insertion or replacement, closely complemented each other, in accordance with the hypothesis of an attractor at the lexical level; the overall evolution of the lexical makeup of utterances also reflected those biases by drifting in a specific direction on each lexical feature (corresponding to easier recall). More generally, we argued that the modeling approach provides crucial detail about the transformations, achieving a middle-ground between the focus on lexical features in individual word replacements and the wide-angle view of contrasts in the aggregated evolution of content along chains.

These analyses were made at the cost of several trade-offs. Transformations in the online data set were restricted to single-word replacements so that we could infer missing source-destination links between quotations, and lack of data meant that no analysis could be made of the context surrounding the quotations. The transmission chain experiments were led with an extremely simple read-and-rewrite task (though this was an intentional first step), which also did not open the analysis to the role of context in transformations and in the overall evolution of content. Nonetheless, these studies demonstrate that it is possible to decompose the transformations of utterances into combinations of smaller operations, and fully connect the behaviour of those operations with known effects in psycholinguistics, be it online (with a partial view of the process) or in controlled transmission chains (with a full view of the transformations). They further suggest that, due to cognitive biases in the way utterances and words are recalled, the evolution of short utterances like quotations could be subject to an attractor at the lexical level, making the words of utterances gradually easier to recall, on top of other changes in the actual content conveyed.

4.2.2 Challenges

However, these studies do not tell us the way utterances evolve semantically. Indeed, apart from the vector-based comparison of individual words for scoring matched and mismatched pairs in utterance alignments (an arguably simplistic approach to word comparison), none of the analyses we put forward have a grip on the meaning of the utterances, and much less on the change in meaning upon transformation. While it is noteworthy that it was still possible to extract reliable decompositions of the transformations without such information (as the manual evaluation of alignments attests), these analyses are blind to changes in the content circulated by the utterances.[2]

Let us show a few examples of the types of meaning change that were observed in the transmission chain experiments of the previous chapter.

Minor operations can change the function of a part of the utterance

Consider the following root utterance from in Experiment 2:

> "Can you think of anything else, Barbara, they might have told me about that party?" (4.1)

[2]We also explored the semantic distance traveled by words upon replacement, and the possible hyponym-hypernym relationships between replaced and replacing words, but did not present the analyses in the previous chapter as they provided no additional insight about the process.

The second part of this sentence is slightly misleading, and could be seen as a mild case of garden path sentence:[3] to "tell about" can be either transitive or intransitive, and while the final "that party" determines it as a transitive verb (for which it is the object), several subjects in the experiment rewrote the following sentence:

"Can you think of anything else, Barbara, they might have told me about **at** that party?" (4.2)

The added "at" turns "tell about" into an intransitive verb, and turns the final part of the sentence into an adverbial phrase of time, thus changing its function in the sentence. More importantly, the sentence in its new form implies that the speaker was at the party, whereas the original sentence implies the contrary (although one could imagine the speaker was present but does not remember the details of the party). There is therefore a substantial change in the high-level meaning of this utterance, through the addition of a single word which changes the function of part of the utterance.

Once this change has occurred, regularisations often happen in the rest of the branch, for instance removing "about" to turn the ending of the sentence into "told me at the/that party". Looking at the leaves of the seven branches this tree contains, 4 of them maintain the implication that neither the speaker nor Barbara were at the party, 2 imply that the speaker (and not Barbara) was at the party, and one implies that Barbara was at the party (and the speaker was not).

Minor operations can create an ambiguity triggering larger changes

As we discussed at the end of the previous chapter, minor changes can also lead to larger downstream consequences in surface representation (as well as in meaning). A second example of typographical error can be seen with the following sub-chain in Experiment 3 (putting aside the UK/US spelling change, "canceled" $\rightarrow$ "cancelled"):

"The charge of embezzlement against the artillery has been canceled." (4.3)

"The charge of embezzlement **again** the artillery has been cancelled." (4.4)

"The charge of embezzlement again, the charge has gone." (4.5)

In this case, the "against" $\rightarrow$ "again" replacement operated in the first transformation leads the following subject to interpret the sentence quite differently, making an larger change. This behaviour is far from systematic, as many times such small errors are corrected by later subjects. Consider for instance the following error made by a subject in Experiment 2:

"At least when they say they're going to have a war, they keep their word." (4.6)

"At least when they say they are going to have a war, they keep **there** word." (4.7)

[3] A garden path sentence is a sentence that misleads the reader into parsing its syntax one way, but necessitates a different structure to be understood once all the words have been read. A classic example is the sentence "The horse raced past the barn fell", which misleads the reader into interpreting "The horse" as the subject of "raced"; the correct parse corresponds to the meaning of "The horse that was raced past the barn fell", where "The horse" is the object of "raced". The difficulty comes from the fact that the search for the correct parse becomes necessary only once the reader has seen the final word, "fell".

The "their" → "there" replacement is maintained by the next subject, but then reverted by the one after that, thus coming back to the original sentence (aside from the change in contraction, "they're" → "they are").

Weak and strong pragmatics

The examples above, and the variability they illustrate, testify to the fact that different subjects can interpret the same utterance in strongly divergent ways. Subjects do not only differ on their performance in accurately reproducing the utterances presented, their productions also signal that different meanings can be constructed from the same root utterance (such differences accumulate, as was illustrated by the divergence of branches observed in the previous chapter). Part of this observation is commonplace, as the meaning of an utterance depends in obvious ways on the context in which it is produced or read, such as when deictics are used (words such as "today" or "here", which are context-bound by nature). However, most isolated utterances are under-specified in a way that makes them much more dependent on the context and on the interaction they appear in then what deictics suggest.

Consider once again utterance 4.1. With no further context, it is not clear what party the speaker is referring to, who were the participants, or why the speaker is asking about it. As the interpretations made by subjects in Experiment 2 illustrate, one could imagine that the speaker was at the party but does not recall its events, or that Barbara witnessed someone telling the speaker about the party, a telling that the speaker would not recall, and so on and so forth with other hypotheses. The sentence is originally extracted from a movie script, and in this case the sentence immediately preceding it in the script is enough to drastically reduce the possible interpretations:

> "I've spoken to the other children who were there that day. Can you think of anything else, Barbara, they might have told me about that party?" (4.8)

With this minimal context, it is now clear that the speaker was not at the party, but is asking Barbara to tell him or her something he or she already knows. To fully understand the utterances however, much more information on the interaction is needed: one must know that the utterances, extracted from the 1997 movie "The Devil's Advocate", are pronounced by a lawyer defending his client, a sexual abuser, while accusingly questionning Barbara, a victim of the abuser and witness in his trial.

This example illustrates what Scott-Phillips (2017) calls *strong* pragmatics. Contrary to *weak* pragmatics, which construes the context-dependence of meaning as a layer to be added on top of semantics, syntax, morphology, phonology and phonetics, strong pragmatics refers to the fact that all communication fundamentally depends on social cognition, which cuts through the other layers of linguistic analysis such as semantics and syntax. Indeed, what is communicated through the utterances discussed above, which can be rephrased as "tell me this thing I already know but that the audience does not", could have been conveyed through an entirely different set of sentences (that is different semantics, syntax, morphology, and so on) because it depends above all on the social cognition situation that participants find themselves in.

Examples of this phenomenon abund, and are not restricted to face-to-face interactions as depicted by films: no matter the type of mediation, any interaction is likely to feature strong pragmatics. Twitter conversations are a good case in point for online platforms. The short conversation reproduced below, for instance, illustrates the fact that the meaning as understood by participants is a construc-

tion depending on context, past history, and interaction dynamics.[4] It starts with the following tweet:

"We are all good-looking and ugly to someone else's eyes" (4.9)

This utterance seems a priori neutral, and is commonplace and consensual enough for it to be marked as favorite, retweeted and published anew regularly.[5] But as illustrated by the answers following it, the actual meaning exchanged in the conversation is not available to the non-interacting reader. A first answer is made in a humourous tone:

"but we're still ugly in the first place haha" (4.10)

Then, two replies later, the conversation ends:

"[laughing out loud,] true for some girls especially, I would say" (4.11)

Even after five replies, we cannot determine whether the meaning exchanged is about sexism and rejection, or simply a flimsy joke without consequence; yet when taken as cultural tokens, these two representations are diametrically opposed to each other. With no further information about the relationship between the participants, their past interactions or common history, and in spite of the conversation being entirely public, we cannot determine what the exchange is fundamentally about, or even decide what the initial tweet means to one participant or the other.

Summary of problems

Let us now return to our initial question, namely the identification of attractors in the evolution of linguistic content. As might be clear by now, this goal is challenging in at least three new and related ways. First, the importance of strong pragmatics renders it much more difficult to collect all the necessary data to understand the meaning that a subject attributes to an utterance, or what is exchanged in an interaction. Indeed, it is often necessary to rely on detailed information about the interactive situation to understand that meaning. Leaving aside the question of the theoretical and technical apparatus that would be required to quantitatively analyse such data, the situations in which an experimenter can have access to the whole interactive situation, and thus have access to meanings exchanged (i.e. determine the content of the representations that circulate), are extremely rare. In most cases, an experiment only gives access to artefacts that are part of a broader cycle of meaning creation.

Second, even when the interactive situation is available to observation, the meaning of an utterance is not reducible to a simple object, and remains a multi-scale (and inside each scale, multi-dimensional) target. Coming back once again to utterance 4.1 with its surrounding context, what aspect of the meaning should one focus on when examining its evolution to identify attractors? The presence

[4] The conversation is originally in French, and reads as follows: "On est tous le beau et le moche de quelqu'un" / "mais être moche c'est quand même la base ahah" / "[mort de rire,] pour certaines filles surtout, je pense".

[5] A simple search on Twitter using the original text in French shows that the utterance appears about once a month, with most instances retweeted several times.

of a request to publicise private information, the implication that Barbara is lying or holding back such information, the lower-level structure of the question? In other words, the goal of identifying attractors in the evolution of meaningful utterances is, at least in our current formulation, underspecified. This problem is not new, and might even not be a theoretical problem for Cultural Attraction Theory: behind the multi-dimensionality of meaning is the fact that culture itself is a multi-scale phenomenon (and multi-dimensional at each scale), difficult to characterise in simple mathematical terms. CAT works around this problem, as Sperber insists that it should not aim for a "grand unitary theory", and should rather generate useful domain-specific questions that depend on the matter at hand (Sperber 1996, 61, 83). Thus the empirical decision of which meaning level to focus on must be resolved by appealing to the importance of each level as individually observed.

Finally, a more important theoretical challenge comes from the fact that strong pragmatics puts an important part of the responsibility for meaning, that is for the content of a linguistic representation, in the interactive situation itself. If the meaning of an utterance is determined in great part by the interaction it features in, and if that meaning corresponds to the content of the linguistic representations whose epidemiology we wish to study, then how is it possible to identify two representations from different situations as being the same (or being close to each other)? To make progress in the epidemiology of meaning-bearing utterances, an approach to strong pragmatics must thus be able to relate meanings that come from different interactive situations, to some extent at least. Indeed, evaluating the evolution of representations requires us to be able to identify, if not the path taken by specific strands of representations which inherit from each other, at least the overall trajectory of a population of representations in a common state space. As a consequence, an approach to pragmatics useful to CAT must provide a way to declare meanings different, or identical, or evaluate the extent to which they differ, across situations (without which evolution can only be observed inside fixed interactive situations).

4.3 Approaches to meaning

If we thus broaden the scope of empirical studies of CAT to all interactions (face-to-face or digitally mediated, but not necessarily linguistic, and in any case beyond interaction-less transmission chains), as will eventually be necessary for strong pragmatics, we are faced with the concrete question of how to understand the way an agent (participant, subject, person, or non-human organism) extracts or constructs meaning in such an interaction. That is, which of the infinite possible meanings the agent selects (or constructs), and how that selection (or construction) operates. As we just saw, such meaning is highly dependent on the context and interaction the agent finds itself in, such that viable approaches to meaning will necessarily be coping with the complexity of possible interactive situations. This makes the picture considerably more complicated than when dealing with simple context-free utterances.

In this section, we present two prominent approaches to meaning and pragmatics, both of which can prove useful for further exploration of cultural evolution. The first, Relevance Theory, fleshes out the idea (first introduced by Grice 1989) that human communication is ostensive communication, based on the recognition of relevant communicative intentions. The second, the Enactive approach, starts from a more bare-bones level of description and proposes an understanding of how meaning emerges from the interaction of agents seen as dynamically coupled organisms. As we will see, both these theories provide (part of) an answer to how agents select, infer or construct subtly varied meanings in the course of an interaction, but they do so by starting from opposite ends. The first builds on a propositional notion of representations that are processed and combined in inference

processes, while the second starts from a representation-less description of organisms whose interaction and coupling with the environment endogenously generate (non-representational) meaning. The notions of meaning to which they arrive are quite disjoint, and have historically been considered in contradiction; indeed, we will then show how these differences can be grounds for a critique of CAT and other Darwin-inspired cultural evolution approaches. In spite of this, we will argue that both approaches to meaning could be productive guides for generating empirical questions and experiments regarding cultural evolution.

4.3.1 Relevance theory

Principles

As a general theory of communication, Relevance Theory has a very broad scope and relates to many areas of cognitive science. Sperber and Wilson (1995) and Wilson and Sperber (2002) provide detailed presentations of the full theory, and many publications in between and since then have fleshed out its relations to a number of neighbouring questions. Wilson and Sperber (2012), in particular, provides a thorough discussion of several linguistic phenomena based on Relevance Theory, as well as openings towards experimental and cultural evolution-related approaches to the question of meaning and relevance (for language evolution see in particular Sperber and Origgi 2012). Here we will restrict ourselves to a sorely abridged presentation of the already summarising Wilson and Sperber (2004), in the hopes that it will be enough for an approximate understanding of the principles underlying the theory and the explanations it provides.[6]

Relevance Theory (RT) opposes itself to the code model of linguistic communication, according to which a speaker's meaning is encoded in an utterance, passed on to the listener for instance by means of sound (the channel, or conduit), and then decoded by the listener to obtain the communicated meaning. By contrast, RT adopts an inferential model according to which an utterance does not encode a meaning per se, as the semantics of utterances provide only under-determined information about the speaker's meaning (as illustrated by the examples discussed above); instead, the inferential model considers that utterances provide evidence (and exactly the right amount of evidence) for the intended meaning to be inferred given the situational context. This model of communication was first elaborated by Grice, building on the fact that people who are communicating usually assume that what the other person is saying is meant to be understood given the context at hand; in other words, people take their interlocutors to be neither stupid nor adversarial, and assume (consciously or not) that what a speaker says is a signal for a meaning that the listener should be able to understand, through inference. Grice thus identified four general rules (maxims) that listeners generally assume their interlocutor will follow, and on which they rely to infer meaning: Quality (truthfulness), Quantity (informativeness), Relation (relevance), and Manner (clarity). RT agrees with the intuition behind Grice's observations (although it differs on exactly which listener expectations should be necessary), and fleshes out this general inferential model of communication in cognitively plausible terms.

RT proposes that inferential communication is based on a cost-reward comparison of possible conclusions that derive from a speaker's utterance. Indeed, a given utterance (or non-linguistic communicative act) in a given context can lead a listener to any number of conclusions about the world. Each

[6] Our presentation focuses on the founding principles of Relevance Theory. The remainder of Wilson and Sperber (2004) fleshes out the way the inferential procedure is applied to linguistic utterances, how the theory explains typical phenomena such as loose uses of language (e.g. the meaning of 'square' in expressions such as 'square face' or 'square mind'), irony, or poetry, and how it fits with the massive modularity of mind approach introduced in Sperber (1996), along with many detailed examples.

of those conclusions about the world can matter more or less to the listener (RT formulates this as the strength of the contextual effects created by the conclusion), and is also more or less costly to derive from the speaker's utterance and its context (processing cost in RT terminology). A conclusion that matters more to the listener achieves higher relevance, and conversely a higher processing cost will lower the relevance realised by a conclusion. These two dimensions let listeners order the conclusions that can be derived from a speaker's utterance based on their (context- and listener-dependent) relevance. For instance, Sperber or Wilson hearing that their train to work is one minute late is less relevant (because it matters less to them) than hearing that their train is late by a half hour. Conversely, a public announcement stating that their train is late provides a more relevant conclusion (because easier to derive) than the same conclusion derived through more deductive effort from bits of a conversation overheard between the people sitting next to them. A central claim of RT is that evolution has shaped human cognition in such a way that people automatically and easily perform this derivation and comparison process on all the stimuli they perceive, picking out those among the myriad available which maximise relevance. The Cognitive Principle of Relevance expresses this claim: "Human cognition tends to be geared to the maximization of relevance" (Wilson and Sperber 2004, 610).

Wilson and Sperber (2004) then define inferential communication as consisting of two elements. An *informative intention*, that is the intention of a speaker to inform an audience of something (more precisely, to make certain assumptions more, or less, manifest to the listener), and a *communicative intention*, that is the speaker's intention to inform the audience of their informative intention. In other words, inferential communication happens whenever the speaker says (or does) something in order to make her audience recognise that she wants to convey X. The meaning is successfully understood when the audience recognises the speaker's informative intention, that is when the audience recognises that the speaker wants to convey X (note that X itself might not be conveyed if the audience does not trust the speaker – the communication event is nonetheless successful, since the intention to convey X was recognised). Most often, the speaker does this by making an ostensive communication act (e.g. pointing, staring, or saying something that attracts the audience's attention) which signals to the audience that there is something worth processing to attend to. Indeed, ostensive stimuli create in the audience an automatic expectation for relevance, as the audience looks for the reason for which the speaker is attracting their attention. More precisely, RT posits that the audience automatically expects the stimulus to be *optimally* relevant; in the theory's terminology, this is formulated as the Communicative Principle of Relevance: "Every ostensive stimulus conveys a presumption of its own optimal relevance." This principle is the basis on which the audience's inferential process works: the speaker's ostensive stimulus signals something worth processing to reach a relevant conclusion (since she attracted their attention to process it), and it is also the stimulus that makes that relevant conclusion the easiest one to reach.

The authors discuss an example to illustrate this: we are at a table and my glass is empty, a fact that you might notice. If you do (without me communicating anything), one conclusion you could reach is that I might like a drink. If, however, I wave my glass at you, or say "My glass is empty" (ostensive stimulus attracting your attention to my empty glass), a relevant conclusion you would reach is that I want to communicate that I want a drink (and, if you trust me, conclude that I want a drink, although that is not necessary for the communication to succeed). RT thus proposes a procedure that can account for the way utterances are understood: when perceiving a stimulus (possibly ostensive), and given a certain context, compute the relevance of its conclusions (i.e. the strength of contextual effects pitted against the processing costs) following a path of least effort first (since the stimulus is expected to be optimal), and stop whenever you have reached your expected level of relevance. In other words, test hypotheses about the speaker's utterance such as possible disambiguations, resolution of entities and implicatures, and stop whenever the conclusions you

have reached seem relevant enough to you. The conclusion you have then obtained will be your assumption of the speaker's meaning for the context chosen at the beginning. Finally, note that this inference procedure can be operated in different possible contexts, as long as they are available given the memory constraints of the listener. The procedure thus also optimises on the context in which conclusions are drawn, and selects the context for which the final conclusion is most relevant.

Application

The framework provided by Relevance Theory is extremely rich, and has been the subject of extensive experimental exploration and validation (see Noveck and Sperber 2012; and Henst and Sperber 2012 for reviews). In particular, it brings direct insight to some of the limitations concerning meaning in the experimental approach of the previous chapter. The transmission chains we set up are a clear case of ostensive communication: we ostensively ask the subjects to direct their attention to the utterances they are asked to memorise and rewrite. However, the utterances are presented with no context other than the task itself, which frames the experiment as a memory exercise. There is no background information against which the subjects can evaluate the relevance of conclusions derived from the utterances, nor are the subjects involved in an activity that would make the conclusions matter in one way or another. Without an ecological activity to which the utterances can become relevant, the experiment has no control over the meanings that subjects will infer from the utterances, leaving the matter entirely under-specified. It is also easier to understand why subjects spontaneously wrote sentences directed to the experimenter such as "I can't remember": if the task does not create an ecological communicative activity that relates subjects to each other, and is instead (correctly) understood to be a memory task, the only valid interlocutor is the experimenter evaluating the memory performance. As a consequence, asking the subjects to keep in mind that their productions were sent to other subjects (as we did) was likely to create a slight discrepancy.

The relevance-theoretic approach also opens the door for the analysis of interaction, context, or past history in the evolution of meanings. Using carefully constructed contexts to orient what is available to the inferential process, it should be possible to greatly reduce the possible meanings interpreted by subjects, and thus explore the way interpretations evolve through chains of contextually-augmented utterances (or chains of constrained interactions). In practive however, such an implementation is likely to be extremely challenging (much more so than the procedure developed in the previous chapter): whatever the theory, it is still necessary to extract the basic propositions semantically encoded in the utterances typed in by subjects, determine the way subjects select contexts, and automatically or manually derive the possible conclusions that can be reached for a given utterance in a given context. The three tasks are far from trivial. Second, efficiently constraining the context in which an utterance is interpreted will likely be much more difficult than it sounds, as real life interpretation involves our own personal history, memory, preoccupations and any other pregnant contexts we can recruit during the process (see Sperber and Wilson 1995, secs 3.3–4, which discusses the ways contexts are chosen in the inferential process). It is doubtful that simply adding a few sentences around the target utterance would suffice. Instead, it could be necessary to create more encompassing situations such as a controlled video game where much larger parts of the many contexts available to subjects can be experimentally manipulated.

Meaning as indexed on knowledge optimisation

We have just seen that the notion of meaning provided by Relevance Theory is based on a maximisation of the relevance of conclusions derived by the listener. This account ultimately rests on the

three following cognitive mechanisms:

- Reconstruction of the logical form of an utterance, in order to start the inferential process (see Sperber and Wilson 1995, sec. 4.3, for details),
- Creation or selection of contexts inside which the inferential process operates (see Sperber and Wilson 1995, secs 3.3–4, for details),
- The inferential process itself, operated by what is hypothesised to be a special-purpose deductive device (see Sperber and Wilson 1995, secs 2.4–5, for details).

This account can also be formulated in terms of a system which optimises its representation of the world (without access to the truthfulness of what it perceives), and for which relevance indicates the path of strongest optimisation growth. To see this we must briefly return to the exact definition of relevance in the theory. Relevance in Sperber and Wilson (1995) is defined in the following three broad steps:

1. Define the *deductive device* that is used for inference: it is a mechanism for deriving conclusions from a set of premises P (usually coming from an utterance from a speaker) in a set of contextual assumptions C (Sperber and Wilson 1995, secs 2.4–5); in this device, all assumptions, premises and conclusions have a certain strength, corresponding to their level of accessibility (this is not a logical measure of confidence that can be quantified, but rather an ordinal property that can be used to compare assumptions between each other).
2. Define the notion of *contextual effect*: a set of premises P (coming from an utterance) in a set of contextual assumptions C generates a contextual effect if and only if the deductive device can derive conclusions from the combination of P and C that it cannot derive from P or C alone. Such conclusions can be new to C, can strengthen existing assumptions in C, or can weaken and even erase assumptions in C (Sperber and Wilson 1995, secs 2.6–7). The notion of contextual effect thus provides a indication of the strength of the relationship between an utterance and a set of contextual assumptions C.
3. Define the degree of *relevance* realised by conclusions derived by the deductive device from P and C: conclusions are more relevant if they have stronger contextual effects, and less relevant if they have higher processing costs (in terms of deductive steps involved).

With relevance thus defined, Relevance Theory's procedure for interpreting utterances and other stimuli can be seen as a tentative procedure to always improve the system's representation of the world: if the processing system has no access to the truthfulness of the stimuli and utterances it perceives, its best option is to process the incoming information in a way that maximises what it can infer about it, given its level of trust in the speaker. In other words, it will look for the set of assumptions that can most benefit from the new information. If the system functions with a deductive device such as the one defined above, this corresponds to finding the set of contextual assumptions C on which the new information (premises P) has the strongest contextual effect, with given processing cost constraints and for a given level of trust in the speaker (i.e., strength of the premises P); given the above definition of relevance, that is precisely the procedure to reach the most relevant conclusions. Under this description, higher relevance indicates a stronger update in contextual assumptions, thus a stronger update to the system's overall representation of the world. Thus Relevance Theory can be seen as indexing meaning on the increase in reliability of a listener's representation of the world. The Communicative Principle of Relevance then simply states that speakers know that listeners work by maximising inferred relevance, and will behave accordingly in order to be understood (i.e. by saying things which they know will trigger the right inferences for the listeners).

Whichever the formulation we choose, RT crucially relies on a deductive device which can represent

propositions extracted from utterances and process such propositions in order to derive new conclusions. Once the inference process is completed however, the meaning defined by RT is partially propositional, as it is made of a set of new assumptions that are made more less manifest (through a change of strength). Such meanings can thus be at least partially defined and used without reference to the situation they were deduced from, making them (partially) comparable to other meanings from other situations. The approach to pragmatics developed by RT is thus quite amenable to the study of cultural evolution (which is no surprise, given the authors).

One disadvantage of the relevance-theoretic approach is the high level at which it starts its description, making concrete implementations more challenging. As we indicated earlier, it relies on the reconstruction of the logical form of utterances, on a mechanism for the creation of contexts, and on the inferential process itself. Second, a theoretically important, but experimentally less important consequence of the theory's reliance on propositions and mental representations is that a full account of meaning will eventually require an account of the content of such representations (a problem known in philosophy of mind as the hard problem of content, Hutto and Myin 2013). We will now turn to a second approach to meaning, one that does not rely on propositions or even representations, and starts from a much lower level of description. Our hope is to convince the reader that this second approach, although less connected to the standard body of linguistic research, should also be a fruitful avenue to explore the effects of interaction in relation to meaning in cultural evolution.

4.3.2 The enactive approach

A non-computational metaphor of cognition

The enactive approach proposes a different foundational metaphor for the study of cognition. Indeed, Relevance Theory and Cultural Attraction Theory, along with our own experimental approach to their questions, mostly rely on a computational metaphor for the description of cognition: the mind is taken to be like a computer, that is an information-processing device which continuously receives stimuli, updates an internal representation of the world based on what it perceives, and acts based on its current representation and predictions given current stimuli. The human brain is our implementation of such an information-processing device. The approach we are interested in here is part of a range of approaches that question the utility of conceiving the mind as such an information-processing system,[7] and explore the extent to which parts of (or all) the metaphor can be relaxed or replaced by other paradigms (see Chemero and Silberstein 2008 for a review of the options available in the debate, and the fields corresponding to each choice). In this area, the enactive approach has the advantage of being extremely consistent in its rejection of the computational paradigm and of the idea that cognitive systems represent their environment, and is to our knowledge the only contender that has started developing a detailed non-computational approach to language itself. As we will see, instead of a computational paradigm, the Enactive approach proposes to base cognition on the dynamical coupling of organisms with their environment and with each other. Compared to Relevance Theory, it is located at the other end of the computational spectrum, and uses a simpler initial level of description. Nonetheless, it is concerned with questions common with RT: both aim to reach complete and plausible explanations of language and meaning. As Chemero and Silberstein (2008) argue however, these two types of approaches are not necessarily opposed, and could be usefully combined to form complementary explanations. Our goal here is to present the basic tenets of the enactive approach and show how, by starting with a different metaphor, it faces an orthogonal

[7] These approaches are sometimes collectively termed the "E turn", in reference to the many titles starting with the letter "e" (in particular, enactive, embedded, embodied and extended approaches to cognition).

set of problems compared to RT. In particular, the notion of meaning it develops seems more endogenous than that of RT (among other things, by being non-representational, it does not face the hard problem of content), but is currently much more low-level and not yet usable to fully comprehend actual linguistic interactions. What the theory currently provides can be seen as the explanation of preliminary steps common to language and less structured interactions, eventually to grow into a full theory of linguistic interactions (or, in enactive terms, "enlanguaged" interactions).

The first concrete articulation of this approach in cognitive science is usually attributed to Varela, Thompson, and Rosch (1991) who develop a view of cognition based on Merleau-Ponty's phenomenology. They propose to look at mind, cognition and meaning as fundamentally embodied and situated processes in which self-organisation plays a central role. Of course, nobody basing themselves on the computational paradigm would deny that what they talk about is ultimately grounded in physical embodied things; however the specificity of the enactive approach (and of other non-representational approaches with it) is that its explanations draw deeply on the embodiedness and situatedness of the processes, in that they create notions of cognition and meaning defined in terms of the coupling of physical systems, rather than in terms of symbolic processing. The initial formulation by Varela, Thompson, and Rosch (1991) led to many developments. We present here the main theory going under the name "enactive approach", and do so in four important conceptual stages which we believe roughly (though drastically) summarise what has been developed by Torrance (2006), Thompson (2007), De Jaegher and Di Paolo (2007), and Cuffari, Di Paolo, and De Jaegher (2015). While this will by no means do justice to the complete approach, we hope these stages will provide a clear-enough sketch of the dynamical and embodied account of cognition that the enactive approach develops and proposes to use instead of the computational metaphor of mind.

Stage 1: sensorimotor contingencies

The first stage is a reconceptualisation of the way an organism perceives its environment. This conceptualisation, known as the sensorimotor approach to perception (and thoroughly developed for vision by O'Regan and Noë 2001), essentially takes perception to be an exploratory activity based on a continuous perception-action loop. By contrast, the default approach to perception is to construe it as an inference problem: through its senses, an organism receives information about the world and attempts to reconstruct an internal representation of it, which is challenging because the information is degraded in a number of ways. Instead, the sensorimotor approach construes perception as the exploration of the regularities in the way stimulations change when the organism moves around or acts on its environment (or on an object). Rather than inferring and internally representing the properties and shape of an object that is being perceived (for instance), the sensorimotor approach construes an organism as exploring the changes it generates in the sensory stimulations when moving, thus making perception and action two parts of a common loop. As O'Regan and Noë (2001) put it: "seeing constitutes the ability to actively modify sensory impressions in certain law-obeying ways." An extreme example of such actively perceived properties is the softness of a sponge, which is felt by prodding and squeezing it but not through static contact (Myin 2003).

One of the strong motivations for this approach is that it provides an endogenous account of the feel of a perceptual modality (i.e. its perceptual consciousness), a longstanding problem in inferential approaches to perception. According to the sensorimotor approach, seeing and hearing feel differently (i.e. one can easily differentiate visual from auditory consciousness) not because they are processed by different parts of the brain, but because of the specific regularities with which stimulations are deformed in each sensory modality when the organism moves. Turning our head, for instance, generates a certain change of stimulation in vision, and a different change in hearing. The way each

modality sees its stimulation change with movement is referred to as its *sensorimotor contingencies*, and is directly tied to the type of perceptual consciousness the modality creates.

An interesting confirmation of this approach is found is experiments using "Tactile Visual Substitution systems", where blind people are equipped with a device that reproduces on their skin (through an array of stimulators) the luminance patterns captured by a camera. The subjects are then tested on their ability to recognise objects using this cutaneous stimulation, and are only able to do so if they actively control the movements of the camera itself (O'Regan and Noë 2001, 958). Furthermore, once they do control the camera, their sensations seem relatively close to actually seeing, because the sensorimotor contingencies are so similar: they begin to perceive objects as not on their skin but in front of them (in particular, they can be frightened by a zooming effect in the stimulations, which corresponds to an object approaching very fast), and the location of the stimulator array on the body becomes unimportant (subjects can easily transfer from stimulation on the back to on the forehead). Such experiments have contributed to showing that perception and action are two sides of the same dynamical interaction loop with the environment, and by generalising to other modalities, they suggest that sensorimotor contingencies provide an endogenous account of perceptual consciousness.

Stage 2: sense-making

The second stage extends this approach to life itself (here we follow De Jaegher and Di Paolo 2007; and Thompson 2007). In a nutshell, it can be seen as taking the reconceptualisation operated by the sensorimotor approach, which goes from a notion of perceptual consciousness based on inference to a notion made of sensorimotor contingencies arising in perception-action loops, and applying it to meaning in cognition: instead of being seen as the result of an inferential process, meaning will be seen as a property (or a regularity) of the dynamical interaction of an organism with its environment.

Let us make this step more precise. Inspired by the notion of autopoiesis developed by Maturana and Varela (1980), the enactive approach considers a living organism as an *autonomous system*, that is a network of processes with the following properties:

1. The system is self-produced and self-maintained. As a consequence the processes depend on each other for continued operation, that is, every process in the network is conditioned on the activity of one or several other processes of the network (a property called *operational closure*). A second consequence is that the network of processes acquires an identity (defined by its operational closure).
2. The system continually produces a boundary that distinguishes it from the environment (this need not be a physical boundary).
3. The system actively regulates its interaction with the environment in order to maintain its identity.

Crucially, the identity generated by operational closure is precarious: it disappears if some or all of the processes that make up the system cease. The system is thus in a permanent tension to regenerate the conditions for the continuation of its identity, and any interaction with the environment thus acquires an inherent value to the system since it can have positive or negative consequences on the continuity of the system's identity and autonomy. Since interactions with the environment are necessary for the network of processes to keep self-generating, the system is continuously regulating the strength of its coupling with the environment in order to maintain its identity. Interaction with the environment then becomes inherently meaningful to the system, and the enactive approach calls it "sense-making".

In this framework, cognition *is* precisely the sense-making activity, that is a system's actively regulating its coupling to the environment in order to maintain its identity. Notice how the enactive notion of meaning is defined in a parallel manner to the sensorimotor account of perceptual consciousness: instead of being inferred and represented, it is a property of the dynamics of the system's interaction with its environment.

Stage 3: participatory sense-making

The third stage extends the theory to interaction between two autonomous systems, and introduces the notion of an autonomy of the interaction itself. De Jaegher and Di Paolo (2007) develop this in two steps. First, they show that some interactions can only be explained at the level of the interaction itself, rather than at the level of participants. An interesting point in that direction has been made in an experiment by Auvray, Lenay, and Stewart (2009). The experiment involves two subjects who share a virtual line on which they each have a cursor. The line and cursors are all invisible, but the subjects receive haptic feedback whenever their cursor is overlapping with the other's cursor. Aside from the subject's cursors, two fixed obstacles are placed on the line (each is perceivable by one subject and not the other), and the cursor of each subject has a shadow that follows it at a fixed distance: when subject A's cursor touches the shadow of subject B's cursor, subject A receives haptic feedback but subject B does not (and vice-versa). The subjects are told about obstacles but not about shadows, and are tasked with clicking as much as possible on each other's cursors. Interestingly, they succeed in doing so, but not because they are able to distinguish between real cursor and shadow. The experiment shows instead that they are not able to make the distinction individually, but solve the task because the interaction of real cursors is more stable (and thus more frequent) than the interaction with a shadow: subjects individually fail the task while succeeding collectively, in a way that can only by understood because of the inherent (and unnoticed) stability of their interaction. The principle highlighted by this experiment is that of the stability of *perceptual crossings*: two organisms can have a dynamically stable interaction because they each look for a behaviour that they themselves create, without necessarily being aware of that fact (for instance mutual gaze of an infant and his mother, where the infant may not be aware that his mother maintains the gaze because he does too). Variations and detailed behaviours in this experimental paradigm have been extensively explored in this literature, providing further support for the results above (see for instance Bedia et al. 2014; Froese, Iizuka, and Ikegami 2014).

Second, De Jaegher and Di Paolo (2007) argue that such stable interactions can acquire an autonomy of their own. An example that most people have experienced in everyday life usefully illustrates their point: when trying to cross someone else in the corridor of a train, and moving to the side to avoid them, at times the other person spontaneously moves to the same side you did; when this happens, you and the other person enter an interaction which both are trying to break from the start: each one moves to one side, and the other does the same, until your movements desynchronise and the interaction breaks down. During the time it persisted however, the interaction acquired its own autonomy which constrained both you and the other person, as neither could break free from it.

This autonomy serves as the basis for defining sense-making at the level of the interaction itself: when two organisms interact while at the same time regulating their coupling to their environment (and respecting each other's autonomy), the interaction itself can spontaneously self-organise and become self-sustaining. Similarly to organisms, then, it acquires an identity of its own, and an interest in maintaining that identity: in that case, since the couplings of each organism to their environment and with each other have an impact on the continuation of the interaction, they become meaningful *to the interaction* which can then partly regulate them. A new sense-making activity thus appears

at the level of the interaction itself, a level that neither of the participants fully control, and which has the potential to create constraints on them. This notion, termed *participatory sense-making* (De Jaegher and Di Paolo 2007), has become a key building block of the enactive theory of interaction as it provides a well-founded and naturalised path to explaining emergent effects in interactions.

Stage 4: languaging

The fourth and final stage brings us to language. Relying on the concepts defined above, Cuffari, Di Paolo, and De Jaegher (2015) propose to see language as a specially structured pattern of participatory sense-making, governed by several levels of conventions interlocked with one another. More precisely, since a crucial feature of participatory sense-making is that the interaction itself acquires regularities that neither participant controls, it follows that interactions create some sort of tension between the way individual organisms regulate their autonomy and the regularities that the interaction may impose on them if they are to continue interacting. Consider once again the perceptual crossing experiment introduced above. In the initial setup by Auvray, Lenay, and Stewart (2009), subjects are not able to distinguish between the other's cursor and its shadow, such that prosociality is neither presupposed nor observed and yet the pair collectively succeeds in accomplishing the task (clicking more on the other's real cursor than on any other object on the shared line). They succeed because the interaction that appears when the two cursors cross each other is naturally stable: when the subjects cross each other, both are informed by haptic feedback, such that both come back on their steps to explore the object they just touched. The interaction thus leads both cursors to criss-cross each other for a small period of time, until one of them moves a bit too far and the stability breaks down; this kind of behaviour does not appear when a cursor touches a shadow, as in that case one of the two subjects is not informed of the encounter. Thus the stability of the interaction results from the way the spontaneous actions of the two subjects dynamically interlock and become coupled; this is a regularity at the level of the interaction that participants do not control, and do not even detect (recall that they fail to individually distinguish between the other's real cursor and their shadow).

Now suppose that subjects can be sensitive to that regularity: a tension appears between the naturally occurring stability, on one side, and the way subjects would like to act. Indeed, subjects become able to identify when they are in the course of a stable interaction and when they are not, but have a priori no way of influencing that interaction without breaking it, since its very existence relies on the naive behaviour described above: the only way to interact is by following the naive rules. Froese, Iizuka, and Ikegami (2014) created exactly that situation with two small changes to the perceptual crossing experiment: first, they allowed each participant a single click per session, making them much more conservative in their behaviours; second, they framed the experiment as a cooperative task where subjects should help each other in detecting each other's cursor (subjects are still not aware of the behaviour of the shadow cursor). In this situation, subjects become sensitive to interactive stabilities, and most importantly they manage to resolve the tension described above. Instead of both cursors permanently criss-crossing each other, a kind of turn-taking behaviour spontaneously appears where one subject stays still while the other criss-crosses it, then the roles are reversed and the first one criss-crosses the second one that is now staying still. Thus a new order of interactive regularity appears, built on the previous one: turn-taking in the perceptual crossing.[8]

Cuffari, Di Paolo, and De Jaegher (2015) generalise and recursively expand this kind of emergence of

[8]Interestingly, the authors also ask the subjects to give Perceptual Awareness Scale ratings for the moments at which they click, and find that such turn-taking episodes correspond to a mutually heightened perceptual awareness of the presence of an other.

a higher-order interactive norm. The level we just described corresponds to the emergence of what they call *co-defined social acts*. Co-defined social acts are like salutations, or acts of giving and receiving: they cannot be completed by one person alone. One person initiates the act (e.g. extending a tentative hand to be shaken in the case of a salutation, or holding out your keys to the person you want to give them to), but the other person must appropriately react to that initiation in order to complete the act (grasping the extended hand and shaking both together, for the case of salutation, or taking the keys offered in the case of giving and receiving). Otherwise the act fails and the interaction breaks down. On top of this level of normativity, another level can develop when social acts themselves serve to regulate other social acts (e.g. ostensively staring at your own extended hand to signal to your interlocutor that they should shake it), leading to yet another higher level of normativity. The expansion thus continues by building each level of normativity as the resolution of a tension between the types of individual and interactive autonomies that exist at the previous level. Cuffari, Di Paolo, and De Jaegher (2015) propose 8 levels of normativity,[9] each one corresponding to a new sensitivity of the interacting organisms to a regularity or constraint at the previous level. Often, the new regularity and its regulation by the participants appears at a different time scale, or in a different dimension than at the previous level. The authors thus propose that linguistic interactions (or languaging in enactive terms) can be understood through a gradual progression of interactive norms, tensions, and resolution by new norms, where each step accounts for additional aspects of full linguistic behaviour (words, for instance, then appear as 'patterns available for enacting certain forms of sense-making', Cuffari, Di Paolo, and De Jaegher 2015, 32). To our knowledge, the higher levels of this expansion have not yet been empirically validated in the manner described above for the first two levels. The theory nonetheless proposes a clear roadmap for constructing an explanation of language, meaning, and linguistic interactions which is fully grounded in the dynamics of interaction between participants. More precisely, the explanation is in terms of recursive regularities and conventions in social interactions. Similarly to how the sensorimotor approach to perception accounts for visual consciousness in terms of regularities in the perception-action loop, the enactive approach accounts for meaning itself as being a combination of aspects of the regularities in social interactions.

Discussion

A number of points can be noted about the approach we just outlined. Overall, the approach strongly reflects the intuitive idea hinted to in our previous discussion of meaning changes in transmission chains (Section 4.2.1): in meaningful interactions, "everything matters". More precisely, anything *can* matter: any seemingly minor detail of the dynamics of an interaction may (or may not) become extremely important if for some reason the participants are sensitive to it in one way or another, and rely on it for instance to resolve a tension. In particular, the simplifications that can occur when symbolically encoding utterances for use in Relevance Theory can neglect aspects of meaningful interactions which turn out to be essential ingredients. A reconstruction that starts from simpler interactions, such as the one encouraged by the enactive standpoint, is more likely to pick up on such ingredients. As noted previously, the problem here is the extreme complexity of interactions and contextual situations, which the enactive approach tackles by starting from simpler (but, crucially, always meaningful) interactions, and by using the language and tools of dynamical systems theory (see for instance Beer 2000; 2014). A second point related to the "anything can matter" intuition is that the coordination of interacting organisms that is necessary to achieve dynamical coupling can

[9] The levels are Participatory sense-making (which we started with), Social Agency (e.g. turn-taking in perceptual crossing), Coordination of Social Acts (e.g. giving and receiving), Normativity of Social Acts, Community of Interactors, Mutual Recognition and Dialogical Structure, Participation Genres, and finally Languaging.

rely on a diversity of dimensions: while interacting organisms need to have comparable dynamics to make it possible for an actual coupling to emerge in their interaction, different dimensions of the dynamics are eligible to that role at different levels of interactive normativity. For instance, coupling in perceptual crossing experiments is likely to emerge only when the subjects make movements of comparable magnitude at the sub-second timescale. Turn-taking on the other hand, is likely to require the subjects to have comparable behaviour at the timescale of a few seconds (the duration of a turn in turn-taking). The poorer the match in dynamics in a given dimension at a given scale, the more difficult it should be for a coupling to appear in that dimension at that scale. Higher levels of interactive normativity will likely involve yet larger timescales (a point that could be related to multi-scale complexity matching in conversations, Abney et al. 2014).

Dale et al. (2014) provide a review of the empirical work that has already been done in (not exclusively enactive) dynamical approaches to interaction. At the level of simple interactions, promising applications have been made through software implementations: Botelho et al. (2015), for instance, develop software agents endowed with inherent goals, such that they could have a very simple but endogenous notion of meaning; Froese, Gershenson, and Rosenblueth (2013) further illustrate the dynamical coupling of maximally simple software agents that manage to collectively solve a task precisely (and only) thanks to their coupling (similar to what the perceptual crossing paradigm explores with human subjects). Now as regards language itself, the approach evidently needs to be much further developed and empirically explored. The problem seems tractable however, as the mysterious aspects of meaningful interaction are already part of the simpler levels of interaction that the current theory convincingly accounts for. We also note that the general approach of expanding normativity levels on top of the previous levels is promising in yet another aspect: on the enactive account, the grammar of a particular language as it is classically understood (i.e. structured utterances with phonetics, phonology, morphology and syntax in English or Spanish for instance – leaving aside semantics and pragmatics since they are closer to meaning, which is accounted for across levels) could correspond to a particular expansion of normativity levels that differs from that of another language. In particular, if the grammar-related normativity levels differ, the levels above and below will also, a fact that could correspond to the different types of agency, and of interaction, afforded by different languages. Any speaker who is fluent enough in more than one language has experienced the change in the way social interactions feel and unfold when switching from one language to another (the effect is strongest when using different languages with the same person). English and Spanish for instance afford different types of interactions, as one constrains meanings differently than the other, each language implicitly expressing a particular set of relationships and making it necessary to spell out other parts more explicitly. In the words of Evans (2011), who details this link between a language and social interaction: "Languages differ not so much in what you can say as in what you must say" (2011, 70). The enactive account of social interaction provides a natural continuity between a particular language and the type of agency that goes with it, and one can expect this type of phenomenon to be explicable once the theory has been further fleshed out.

4.3.3 Applied to cultural attraction

Compared to Relevance Theory, the Enactive approach starts from radically simpler types of interaction, and relies neither on a capacity for symbolic processing nor on internal representations of the environment. This makes the approach slightly more involved to present than Relevance Theory. In spite of their differences, both approaches provide a notion of relevance or value to the individual, which is then used to ground a notion of meaning (be it linguistic or not). Unlike RT, the enactive notion of meaning exists even in the most basic types of interaction an organism can have with its environment.

The two approaches can be seen as starting from different descriptive levels and building what is missing for an account of meaning (though this characterisation does not exhaust their differences): RT starts with representations and symbolic processing, which are implementable in bodies (by following the model of computers) and come already structured in similar ways to language, and constructs on top of those a notion of meaning, with a corresponding interpretation procedure (inference), that is subtle and flexible enough to apply to fuzzy cases like poetry or loose language. The Enactive approach starts from an embodied and dynamic notion of meaning, which by definition exists for all organisms and need not be inferred, but comes without any particular structure; this approach then builds its account of language by detailing how interaction (and thus meaning) can become increasingly structured through a series of emerging levels of interactive normativity.[10] Unlike RT, it does not (yet) provide a clear way to compare meanings across interactive situations, as its very definition of meaning is intrinsically related to the subtle differences of dynamics in different interations. In RT, that work is theoretically done by the notion of representation (or more precisely, the informative intention that is inferred and represented), although in practice much remains to be defined to directly compare representations between each other. In its current state, the Enactive approach is applicable to much simpler interactions than those tackled by RT, as more experimental work is needed to confirm the higher-level details.

This finally brings us back to the usefulness of both these approaches for CAT. Without he does not flesh out the connection, Sperber (1996) refers repeatedly to RT as an important tool to understand the role of interpretation in the epidemiology of representations; indeed, the information-processing account provided by RT fits well with the representational foundation of CAT. But the fact that CAT is formulated in terms of representations is not necessarily a theoretical constraint to integrate it with an enactive approach to meaning: similarly to what we argued above for RT, information-processing can be seen as a particular starting point for the description with no strong import for the ontology of what is being transmitted or evolved,[11] and we see no reason to consider a priori that the two approaches would be incompatible.

In current practice however, the fact that CAT gives so much autonomy to a contentful notion of representation encourages empirical approaches to use a code model and give context a minimal role (i.e. a representation is an encoding of some content). Most applications thus strip down the context-dependence of CAT, thereby obviating the need for Relevance Theory in the first place. By not seeing that the contents of a representation are inherently contextual and constructed in interaction, one reduces the cultural evolution process to a simple accumulation of interpretations that do not rely on context: such a simple model depicts representations as being straightforwardly interpreted into mental versions (with some degree of transformation through reconstruction) then produced anew as public versions (again with some degree of transformation), and the circulation and tranformation of representations creates a dynamical system with attractors. As we just saw, both the Enactive approach and RT show that taking context into account drastically changes the picture. The fact that it is possible to easily degrade CAT into a code model theory also suggests that integrating it with an approach to meaning might require some amendments. We thus see great value in contrasting and possibly combining the two approaches to replace the naive code model that is used in many applications of CAT. In practice, attempting to apply these two approaches will also let us explore if

[10] In doing so, it also provides a partial account of the phenomenology of meaning, something that RT can only provide through a definition of the exact content of representations, and a mechanism for how they represent that content. For more details, see again Hutto and Myin (2013) on the hard problem of content. Harvey (2015) also provides a useful discussion of that question.

[11] Indeed, Sperber (1996, 135, footnote 40) explains that his view of the content of concepts (which we can assume to be close enough to the content of a representation) is influenced by Millikan (1984), which is quite close to non-representational views of content (Harvey 2015).

the theory must be amended in order to fully take the situatedness of meaning into account.[12]

4.4 Empirical speculations

Following Chemero and Silberstein (2008) and Beer and Williams (2015), we believe that contrasting and combining practical uses of the information-processing and dynamical approaches is the best way to move forward in debates that otherwise turn into scholasticism. This section proposes a few avenues to do so in the context of meaning-related cultural evolution. The first point we discuss is an additional method that is directly applicable to the previous chapter's data. The second, third and fourth points discuss approaches similar to the paradigm presented in the previous chapter, which work with ecological material; these approaches must deal with the full complexity of language. The final point discusses an approach similar to the perceptual crossing paradigm, which works with much simpler interaction situations; here the dimensionality of behaviours is lower, such that it should be easier to understand complex dynamical patterns.

4.4.1 Hand-coded meaning classes

Our first point is an analysis method that combines manual and computational steps. It is directly applicable to the data sets of the previous chapter without necessarily adding a notion of context, but is also usable for analysing the data that would be generated by the approaches we delineate further down.

Its first step is to devise a manual measure of the similarity between random pairs of utterances in a tree; given such a manual measure, a dimensionality reduction method can identify classes of meanings in the tree, and attribute each utterance of the tree to a particular meaning class. The operation makes the evolution much more manageable, as each utterance is now assigned to one of a finite number of meaning classes. Since such classes can be approximately represented in a 2-dimensional space (a result of the dimensionality reduction based on similarity measures), it then becomes possible to plot the evolution happening in each tree and observe when a transformation creates a transition from one class to another. The transition probabilities between each meaning class also give an idea of the types of transformations that happen most often at the level of meanings, as encoded in the first step.

The main challenge for this method is the first step, that is gathering manual measures of similarity between utterances in a tree. One reliable approach to this problem is to gather similarity ratings in a triad comparison task (Romney et al. 1996): manual coders are presented with three utterances, and must decide which of the three is most different in meaning to the other two (in which case

[12] Anthropologist Tim Ingold, for instance, has argued at length that the whole project of a Darwin-inspired theory of cultural change can only make sense if one accepts a misguided dualism between context-free information on one side and material implementation on the other, a dualism that is encouraged by the computational metaphor of cognition (Ingold 1998; 2001; 2004; 2007). The requirement of being able to compare meanings across situations (so as to analyse their evolution), in particular, would rest on being able to identify information that exists independently of the situation in which it emerged. The alternative he proposes is heavily inspired by dynamical approaches to cognition (such as the Enactive approach) and a developmental focus on biology and heredity. While we agree that computational approaches to cognition are prone to such dualism, information-processing analyses can also be used as a tool to combine with dynamical approaches, without taking the computational metaphor too seriously. Moreover, we see no reason to believe that dynamical approaches could not be compatible with Darwin-inspired approaches to cultural change. The question of how to relate meanings from different interactive situations is, we believe, better tackled through further formalisation and empirical investigation of the available accounts of meaning.

the other two are considered to have some degree of similarity). While the complete set of triad comparisons grows as n^3 (n being the number of utterances in a tree) and is thus too expensive for trees containing 71 utterances (such as in Experiment 3 in the previous chapter), it is not necessary to rate all the combinations in order to obtain a reliable measure. Indeed, in the complete set of triad comparisons, each pair of utterances appears $n - 2$ times. The so-called Balanced-Incomplete Block Design (Weller and Romney 1988, 49–55) compares triads without repeating pairs more than a fixed number of times, in which case the number of triads grows as n^2.[13] Triad comparisons can be obtained in large quantities by outsourcing the task on platforms such as Amazon Mechanical Turk, or can alternatively be integrated in the online transmission chain platform.[14]

The modified Correspondence Analysis used by Romney et al. (1996) on this type of data yields an n-dimensional representation of the utterances, where the first dimension distinguishes the utterances the most, and the last dimension the least. Utterances can then be plotted on the first two dimensions, grouped into classes by a clustering method, and the evolution along the branches thus reduced to transitions between meaning classes. Such a measure would provide an interesting first insight into the meaningful transformations operated by the subjects.

4.4.2 Minimal interaction and context

The transmission chain paradigm can also be tweaked in at least two ways to begin exploring the role of context and interaction. One area to explore is to consider utterances not in isolation, but as part of a contextual paragraph. The surrounding paragraph could provide enough background for subjects to have a feel of what it amounts to pronounce that utterance in context. The task can then be framed as role-playing: subjects would be asked to imagine the scene depicted by the paragraph they just read, imagine themselves as the person pronouncing the last utterance (or an utterance in the middle, highlighted as they were reading), and rewrite it. The larger the context, the more constrained we expect interpretation to be, thus the more limited we expect transformations to be.

A second change, which can be combined with the first, is to introduce minimal forms of live communication in the task so as to embed the subjects in an actual communicative task. For instance, the transmission step could be turned into a minimal interaction where the first subject proposes an utterance (that is, their memorisation of what they read) and the second can either accept it or ask for a better reformulation (up to a maximum number of times), in which case the first subject must rewrite their proposal. The interaction requires a fine balance of bonuses and rejection penalties so that the functionality is not abused, but it gives subjects an interest in the task: a receiving subject is encouraged to ask for a consistent utterance, because they will later be asked to pass on their own memorisation of that utterance to another subject. Combining such minimal interactions with contextual paragraphs could further ensure that subjects do not change to a completely different utterance in order to see their proposal accepted: if the receiver is presented with the contextual

[13] A design where each pair appears once for $n = 70$ does not exist, however designs exist for $n = 69$ or $n = 73$, and can be created by removing utterances or adding dummy utterances. $n = 69$ yields 805 triad comparisons, and $n = 73$ yields 876 (see Weller and Romney 1988 for the detailed computation). More reliable measures can be obtained by having each pair appear twice, which doubles the number of triad comparisons.

[14] Consider for instance an experiment similar to Experiment 3 of the previous chapter, with 71 utterances per tree (one root, and 7 branches of 10 reformulations), and a targeted 70 subjects in total. In a given tree, the number of triad comparisons with each pair appearing exactly once is at most 876 (see previous footnote), which can be distributed as an average 12.51 comparisons per subject. Counting about 20 seconds per comparison, the added comparisons thus require each subject to spend, on average, an additional 4m10s on each tree. Given that individual transformations took an average 1m12s in Experiment 3, that is subjects spend on average 1m12s on each tree in Experiment 3, this conservative estimate multiplies the total cost of the experiment by about 4.5. Collecting these triad comparisons for all the trees is thus quite costly, but doing it for one or two trees only is very affordable.

paragraph when deciding to accept or reject the proposed utterance, then they will expect the two to be consistent with each other.

While these changes allow us to introduce minimal forms of context and interaction in a transmission chain, the paradigm dissociates the complexity of one communication means, namely the utterances (which have highly complex linguistic structure), from the complexity of another communication means, namely the interaction itself (which is minimal). This is quite compatible with Relevance Theory, but the result is still somewhat unecological: if context and interactive situation are crucial to the way meaning is understood (according to both Relevance Theory and the Enactive approach), it would be more natural to match the complexity allowed by the interactive situation with the complexity of other communication means that are provided to the subjects. At the very least, it seems necessary to have an interactive situation at least as versatile as the productions that are asked of the subjects. Thus, asking subjects to write complex utterances while constraining their non-verbal interaction to a binary acceptance-rejection outcome still only makes sense in a code model theory, where complex meanings can be understood in spite of the interactive situation being extremely simple. To create situations where the meshing of context, interaction and meaning can be understood, then, it is necessary to match the complexity of all the communicative means provided to the subjects (interactive, verbal, or other). In what follows we discuss possibilities for analysing situations where the complexity is close to that of real life. We then move on to situations where both the interaction situation and the possible meanings have a much reduced complexity.

4.4.3 Fully measured contexts

In some cases, it is possible to fully measure all observables in an unconstrained interactive situation. The first way of doing this is to have enough sensors in an *in vivo* situation: the Human Speechome Project (Roy et al. 2006), for instance, equips a family's house with wide-angle ceiling cameras and microphones that map the entire space of the house, and records the quasi-totality of what happens in the first three years of the life of a newborn in the family (with some privacy controls). Everything the child hears and sees, all the interactions she is involved in are taped in a way that makes it possible to reconstitute the detailed movements, vocal productions, and relevant interactive features (such as gaze) of participants. The analysis of such rich data is humongous and involves many novel semi-automated coding techniques (as the dozens of data-analysis publications related to the project[15] attest to), but gives access to the full detail of interactions that a child is exposed to in her house. Roy et al. (2015), for instance, use the data to explore the spatial and linguistic contexts in which a child is exposed to a word, and relate them to the context in which she first produces said word. Such longitudinal multi-modal data is exactly the type of measurement that is necessary to fully understand the meaning that is produced in an interaction, explore when that meaning is reproduced, and see how it changes through time.

The second way is to create an *in vitro* interactive situation that is as encompassing as possible, and measure every possible aspect of the interaction in it. In the laboratory, this corresponds to the approach taken by Moussaïd, Brighton, and Gaissmaier (2015), who taped unconstrained pairwise conversations chained one after the other, hand-coded them for specific features, and then measured the evolution of such features along chains. Another approach consists in taking advantage of online video games that create complete worlds in which the interactions of players can be measured in multiple modalities (gaming behaviours and parallel verbal conversations), as has been used for instance in the study of language learning (Zheng, Newgarden, and Young 2012). As developing

[15] See https://www.media.mit.edu/cogmac/projects/hsp.html for a full list.

such a game for experimental purposes involves a substantial investment however, researchers must rely on existing gaming environments and communities to access the data.

Fully measuring unconstrained interactions is thus one way of matching the complexity of the different communicative means available to the subjects. However, the problem with this approach for cultural evolution is precisely the increased complexity of context and interaction: existing works use important manual or semi-manual steps in their analyses, such that their access to meaning in the data is in large part interpretive. Using the relevance-theoretic approach to delve into the detail of how meaning is inferred in such situations is likely to become quickly unmanageable. We therefore turn to the other end of the spectrum: approaches which, by relying on the Enactive account for which simpler meanings can exist without the requirement of symbolic processing, aims to lower the complexity of interactions and possible meanings.

4.4.4 Preliminary enactive steps to language

The final approach we discuus aims to take advantage of the empirical appeal of the Enactive approach: using paradigms similar to that of perceptual crossing, one might be able to create interactions that have meaning to the participants without the need to use complex linguistic material. In such a situation, participants would interact through a combination of low-dimensional channels in a task that has inherent meaning to them, and both the interaction and the meaning emerging from it would have extremely low complexity. Such an approach would not enable participants to have everyday life conversations, but it would let the experimenter introduce complexity in experimental tasks little by little, with the end goal of creating situations whose complexity can approach that of actual linguistic interactions.

In its current state however, the Enactive approach does not yet provide a full account of language. The phenomena it does account for, such as sensorimotor contingencies and perceptual crossings, are better described as potential preliminary steps to language: they are elements that can be combined and fleshed out in order to develop an explanation of language.

The theory of languaging developed by Cuffari, Di Paolo, and De Jaegher (2015) proposes a way to do so: it progresses in a series of normative levels of interaction, each one appearing as a resolution of a tension at the previous level. The first normativity level is that of Participatory sense-making, which the perceptual crossing stability (Auvray, Lenay, and Stewart 2009) illustrates: two participants are coupled in an interaction that neither of them controls, or even detects. The second normativity level is that of Social Agency, illustrated by the turn-taking phenomenon that appears in perceptual crossing when participants are sensitive to an interactive stability (Froese, Iizuka, and Ikegami 2014). As we mentioned earlier, this phenomenon can be seen as a co-defined social act, that is an act that is initiated by one person (who makes a *partial* act), and must be completed by the other person recognising it and responding appropriately (e.g. giving and receiving). The authors propose another 6 conceptual stages which together would provide a first account of the languaging behaviour of human beings. While these stages are proposed as conceptual steps, and not as causal steps that would correspond to the actual progression of communicative complexities found in other species (or through evolution), an important milestone here would be to empirically validate the remaining six stages in a similar fashion to the first two. If successful, completing such a validation would provide a more comprehensive view of the enactive language account, and would open the way for integrations with cultural evolution theories.

The third normativity level, for instance, is that of Coordination of Social Acts; it emerges when one social act is used to regulate another social act. The authors' example of this normativity level is a

japanese woman extending her personal card to a westerner. The woman offers her card by holding it with both hands, which is the polite way in Japan. The polite way of responding to this partial act in Japan is to accept the card with both hands too. However, a westerner would spontaneously take the card with only one hand, a response that the woman would not consider appropriate (indeed, it amounts to answering with part of one's body to a partial act that involves the whole body). In response to this, the japanese woman could hold on to her card, and ostensively stare at the westerner's other hand, thus initiating a second partial act to regulate the first. The westerner, understanding the regulatory stare, could nod (thus completing the regulatory act) and accept the card with both hands (thus completing the initial act of giving).

This type of regulatory social act can appear when a second channel and a repertoire of acts are available to regulate what happens in a primary channel. To empirically validate this normativity level, then, one could imagine setting up two parallel perceptual crossing tasks: the first, more complex, would involve cursors on a shared surface instead of a line; the second, simpler, would be the classical uni-dimensional perceptual crossing setup. The reasoning is that the simpler channel could emerge as a regulatory aid for a new type of social act to emerge in the more complex channel. This simple extension of perceptual crossing is most likely not the best, but it illustrates the type of experiment that one could devise in order to observe the emergence of the third stage.

In the process of validating the third normativity level, several conventions will appear that depend on the history of interactions between agents and in a community. For instance, one repertoire of social acts will develop in one community of frequently-interacting participants, and a second repertoire will develop in another. Participants from different communities that are brought together will be forced to adjust, expand, or develop anew their repertoire of social acts. This setting opens a number of questions about the evolution of such minimal social acts in communities. In order to explore the integration of this approach with CAT, it will be necessary to empirically flesh out the full range of normativity levels in this manner. While the task is not easy, we believe that such a project has great potential to possibly complement a relevance-theoretic approach to meaning, on one side, and to identify incompatibilities and possibilities for integration with the current theories of cultural evolution, on the other.

4.5 Conclusion

In this chapter, we took a more in depth look at what would be required in order to understand the transformation of the meaning of utterances in short-term cultural evolution, be it in online quotations or in artificial transmission chains. We first discussed a number of examples of meaning change that appeared in the experiments of the previous chapter. We then presented two radically different, but both prominent, approaches to meaning in language: Relevance Theory, which bases its account on the inference of relevant conclusions given the context in which an utterance is pronounced; and the Enactive approach, which bases its account on the agent's sense-making activity by which it maintains its identity as an agent. While the two approaches sit at opposite ends of the computational-dynamical spectrum in cognitive science, both base their account of meaning on some notion of the relevance that interactions have for agents (though those notions are very different for the two theories). Notwithstanding the simpler integration of RT with Cultural Attraction Theory, we proposed that both RT and the Enactive approach could be valuable paths to explore in the study of meaning in this context. Finally, then, we extended a few speculations as to how these approaches could inspire future work aimed at understanding the way meanings evolve along with cultural evolution.

Our stance may be taken as very ecumenical, too much so maybe. How could it make sense to try and combine approaches that rely on bodies of philosophical works that continuously butcher each other? In their current state, neither Relevance Theory nor the Enactive approach provide complete accounts of what linguistic meaning is or how we understand it, but both provide extremely valuable steps to begin understanding the phenomenon. Our belief is that philosophical butchery is a healthy and necessary activity, and empirical combination and contrasting of the outcomes is too. While the two approaches clearly have disagreements on starting points and approaches to cognition overall, it is not said that different approaches to meaning, when fully formalised at the right level, could not be related to one another. In the meantime, both can inspire valuable empirical work and the design of new methods, be it to contrast or to combine them, all of which contribute to better understanding the issues at stake.

General conclusion

References

Abney, Drew H., Alexandra Paxton, Rick Dale, and Christopher T. Kello. 2014. 'Complexity Matching in Dyadic Conversation'. *Journal of Experimental Psychology: General* 143 (6): 2304–15. doi:10.1037/xge0000021.

Acerbi, Alberto. 2016. 'A Cultural Evolution Approach to Digital Media'. *Frontiers in Human Neuroscience* 10 (December). doi:10.3389/fnhum.2016.00636.

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.

Acerbi, Alberto, and Jamshid J. Tehrani. 2017. 'Did Einstein Really Say That? Testing Content Versus Context in the Cultural Selection of Quotations'. *SocArXiv Preprints*, April. https://osf.io/preprints/socarxiv/x8db2.

Acerbi, Alberto, and Claudio Tennie. 2016. 'The Role of Redundant Information in Cultural Transmission and Cultural Stabilization'. *Journal of Comparative Psychology* 130 (1): 62–70. doi:10.1037/a0040094.

Adamic, Lada A., Thomas M. Lento, Eytan Adar, and Pauline C. Ng. 2016. 'Information Evolution in Social Networks'. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 473–82. New York, NY, USA: ACM. doi:10.1145/2835776.2835827.

Althoff, Tim, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. 'How to Ask for a Favor: A Case Study on the Success of Altruistic Requests'. *arXiv:1405.3282 [Physics]*, May. http://arxiv.org/abs/1405.3282.

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, USA: Curran Associates, Inc. http://papers.nips.cc/paper/4761-human-memory-search-as-a-random-walk-in-a-semantic-network.pdf.

Auvray, Malika, Charles Lenay, and John Stewart. 2009. 'Perceptual Interactions in a Minimalist Virtual Environment'. *New Ideas in Psychology* 27 (1): 32–47. doi:10.1016/j.newideapsych.2007.12.002.

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.

Bakshy, Eytan, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. 'Everyone's an Influencer: Quantifying Influence on Twitter'. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, edited by Irwin King, Wolfgang Nejdl, and Hang Li, 65–74. WSDM '11. New York, NY, USA: ACM. doi:10.1145/1935826.1935845.

Bakshy, Eytan, Brian Karrer, and Lada A. Adamic. 2009. 'Social Influence and the Diffusion of User-Created Content'. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, 325–34. New York, NY, USA: ACM. doi:10.1145/1566374.1566421.

Bangerter, Adrian. 2000. 'Transformation Between Scientific and Social Representations of Conception: The Method of Serial Reproduction'. *British Journal of Social Psychology* 39 (4): 521–35. doi:10.1348/014466600164615.

Barrett, Justin L., and Melanie A. Nyhof. 2001. 'Spreading Non-Natural Concepts: The Role of Intuitive Conceptual Structures in Memory and Transmission of Cultural Materials'. *Journal of Cognition and Culture* 1 (1): 69–100. doi:10.1163/156853701300063589.

Bartlett, Sir Frederic Charles. 1995. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bastian, Mikaël, Sébastien Lerique, Vincent Adam, Michael S. Franklin, Jonathan W. Schooler, and Jérôme Sackur. 2017. 'Language Facilitates Introspection: Verbal Mind-Wandering Has Privileged Access to Consciousness'. *Consciousness and Cognition* 49 (March): 86–97. doi:10.1016/j.concog.2017.01.002.

Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber. 2013. 'A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice'. *Behavioral and Brain Sciences* 36 (1): 59–78.

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.

Bebbington, Keely, Colin MacLeod, T. Mark Ellison, and Nicolas Fay. 2017. 'The Sky Is Falling: Evidence of a Negativity Bias in the Social Transmission of Information'. *Evolution and Human Behavior* 38 (1): 92–101. doi:10.1016/j.evolhumbehav.2016.07.004.

Bedia, Manuel G., Miguel Aguilera, Tomás Gómez, David G. Larrode, and Francisco Seron. 2014. 'Quantifying Long-Range Correlations and 1/F Patterns in a Minimal Experiment of Social Interaction'. *Frontiers in Psychology* 5. doi:10.3389/fpsyg.2014.01281.

Beer, Randall D. 2000. 'Dynamical Approaches to Cognitive Science'. *Trends in Cognitive Sciences* 4 (3): 91–99. doi:10.1016/S1364-6613(99)01440-0.

———. 2014. 'Dynamical Analysis of Evolved Agents: A Primer'. In *The Horizons of Evolutionary Robotics*, edited by Patricia A. Vargas, Ezequiel A. Di Paolo, Inman Harvey, and Phil Husbands, 65–76. Cambridge, Mass. ; London: MIT Press.

Beer, Randall D., and Paul L. Williams. 2015. 'Information Processing and Dynamics in Minimally Cognitive Agents'. *Cognitive Science* 39 (1): 1–38. doi:10.1111/cogs.12142.

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 *Dar-*

winizing Culture: The Status of Memetics as a Science, edited by Robert Aunger, 189–204. Oxford, NY: Oxford University Press.

Botelho, Luis, Luis Nunes, Ricardo Ribeiro, and Rui J. Lopes. 2015. 'Software Agents with Concerns of Their Own'. *arXiv:1511.03958 [Cs]*, November. http://arxiv.org/abs/1511.03958.

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.

Boyer, Pascal. 2001. *Religion Explained: The Evolutionary Origins of Religious Thought*. Basic Books.

Brysbaert, Marc, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. 'How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age'. *Frontiers in Psychology* 7 (July). doi:10.3389/fpsyg.2016.01116.

Caldwell, Christine A., and Ailsa E. Millen. 2008a. 'Experimental Models for Testing Hypotheses About Cumulative Cultural Evolution'. *Evolution and Human Behavior* 29 (3): 165–71. doi:10.1016/j.evolhumbehav.2007.12.001.

———. 2008b. 'Studying Cumulative Cultural Evolution in the Laboratory'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1509): 3529–39. doi:10.1098/rstb.2008.0133.

Caldwell, Christine A., and Kenny Smith. 2012. 'Cultural Evolution and Perpetuation of Arbitrary Communicative Conventions in Experimental Microsocieties'. *PLOS ONE* 7 (8): e43807. doi:10.1371/journal.pone.0043807.

Carr, Jon W., Kenny Smith, Hannah Cornish, and Simon Kirby. 2017. 'The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World'. *Cognitive Science* 41 (4): 892–923. doi:10.1111/cogs.12371.

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.

Chemero, Tony, and Michael Silberstein. 2008. 'After the Philosophy of Mind: Replacing Scholasticism with Science'. *Philosophy of Science* 75 (1): 1–27.

Christie, Tom, and Django REST framework contributors. 2017. 'Django REST Framework'. http://www.django-rest-framework.org/.

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 Systematically 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 Diffusion of Counter-Intuitive Beliefs'. *Manuscript Submitted for Publication*. https://sites.google.com/site/hugomercier/Argumentation%20and%20counter-intuitive%20beliefs.pdf.

Clark, Andy, and David Chalmers. 1998. 'The Extended Mind'. *Analysis* 58 (1): 7–19. http://www.jstor.org.ins2i.bib.cnrs.fr/stable/3328150.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. 'Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics'. *Bioinformatics* 25 (11): 1422–3. doi:10.1093/bioinformatics/btp163.

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.

Cointet, Jean-Philippe, and Camille Roth. 2007. 'How Realistic Should Knowledge Diffusion Models Be?' *Journal of Artificial Societies and Social Simulation* 10 (3): 5. http://jasss.soc.surrey.ac.uk/10/3/5.html.

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.

Croft, William. 2013. 'An Evolutionary Model of Language Change and Language Structure'. In *Explaining Language Change: An Evolutionary Approach*, Draft 2nd edition (revised). http://www.unm.edu/~wcroft/Papers/ELC2-Chap02.pdf.

Cuffari, Elena Clare, Ezequiel Di Paolo, and Hanne De Jaegher. 2015. 'From Participatory Sense-Making to Language: There and Back Again'. *Phenomenology and the Cognitive Sciences* 14 (4): 1089–1125. doi:10.1007/s11097-014-9404-9.

Czaplicki, Evan, and Elm contributors. 2017. 'Elm: A Delightful Language for Reliable Webapps'. http://elm-lang.org/.

Dale, Rick, Riccardo Fusaroli, Nicholas D. Duran, and Daniel C. Richardson. 2014. 'The Self-Organization of Human Interaction'. In *The Psychology of Learning and Motivation. Volume 59 Volume 59*, edited by Brian H Ross, 43–96. Amsterdam: Academic Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=585381.

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 Association 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.

Danescu-Niculescu-Mizil, Cristian, Lillian Lee, Bo Pang, and Jon Kleinberg. 2011. 'Echoes of Power: Language Effects and Power Differences in Social Interaction'. *arXiv:1112.3670 [Physics]*, December. http://arxiv.org/abs/1112.3670.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. 'A Computational Approach to Politeness with Application to Social Factors'.

arXiv:1306.6078 [Physics], June. http://arxiv.org/abs/1306.6078.

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

De Jaegher, Hanne, and Ezequiel Di Paolo. 2007. 'Participatory Sense-Making'. *Phenomenology and the Cognitive Sciences* 6 (4): 485–507. doi:10.1007/s11097-007-9076-9.

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.

Di Paolo, Ezequiel A. 2005. 'Autopoiesis, Adaptivity, Teleology, Agency'. *Phenomenology and the Cognitive Sciences* 4 (4): 429–52. doi:10.1007/s11097-005-9002-y.

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

Ember.js contributors. 2017. 'Ember.js: A Framework for Creating Ambitious Web Applications.' https://emberjs.com/.

Eriksson, Kimmo, and Julie C. Coultas. 2012. 'The Advantage of Multiple Cultural Parents in the Cultural Transmission of Stories'. *Evolution and Human Behavior* 33 (4): 251–59. doi:10.1016/j.evolhumbehav.2011.10.002.

———. 2014. 'Corpses, Maggots, Poodles and Rats: Emotional Selection Operating in Three Phases of Cultural Transmission of Urban Legends'. *Journal of Cognition and Culture* 14 (1): 1–26. doi:10.1163/15685373-12342107.

Evans, Nicholas. 2011. 'Your Mind in Mine: Social Cognition in Grammar'. In *Dying Words: Endangered Languages and What They Have to Tell Us*, 69–80. New York, NY, USA: John Wiley & Sons.

Fay, Nicolas, Simon Garrod, Leo Roberts, and Nik Swoboda. 2010. 'The Interactive Evolution of Human Communication Systems'. *Cognitive Science* 34 (3): 351–86. doi:10.1111/j.1551-6709.2009.01090.x.

Fénéon, Félix, and Luc Sante. 2007. *Novels in Three Lines*. New York, NY, USA: New York Review Books.

Fisher, Ronald Aylmer. 1930. *The Genetical Theory of Natural Selection*. Oxford, UK: Clarendon Press. http://www.biodiversitylibrary.org/title/27468.

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.

Froese, Tom, Carlos Gershenson, and David A. Rosenblueth. 2013. 'The Dynamically Extended Mind'. In *2013 IEEE Congress on Evolutionary Computation*, 1419–26. doi:10.1109/CEC.2013.6557730.

Froese, Tom, Hiroyuki Iizuka, and Takashi Ikegami. 2014. 'Embodied Social Interaction Constitutes Social Cognition in Pairs of Humans: A Minimalist Virtual Reality Experiment'. *Scientific Reports* 4

(January): 3672. doi:10.1038/srep03672.

Fuentes, Agustín. 2006. 'Evolution Is Important but It Is Not Simple: Defining Cultural Traits and Incorporating Complex Evolutionary Theory'. *Behavioral and Brain Sciences* 29 (4): 354–55.

———. 2009. 'A New Synthesis. Resituating Approaches to the Evolution of Human Behaviour'. *Anthropology Today* 25 (3): 12–17.

Futuyma, Douglas J. 2005. *Evolution*. Sunderland, MA, USA: Sinauer Associates.

Galantucci, Bruno. 2005. 'An Experimental Study of the Emergence of Human Communication Systems'. *Cognitive Science* 29 (5): 737–67. doi:10.1207/s15516709cog0000_34.

Galantucci, Bruno, Simon Garrod, and Gareth Roberts. 2012. 'Experimental Semiotics'. *Language and Linguistics Compass* 6 (8): 477–93. doi:10.1002/lnc3.351.

Garlock, Victoria M., Amanda C. Walley, and Jamie L. Metsala. 2001. 'Age-of-Acquisition, Word Frequency, 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.

Garrod, Simon, Nicolas Fay, John Lee, Jon Oberlander, and Tracy MacLeod. 2007. 'Foundations of Representation: Where Might Graphical Symbol Systems Come from?' *Cognitive Science* 31 (6): 961–87. doi:10.1080/03640210701703659.

Gauld, Alan, and Geoffrey M. Stephenson. 1967. 'Some Experiments Relating to Bartlett's Theory of Remembering'. *British Journal of Psychology* 58 (1): 39–49. doi:10.1111/j.2044-8295.1967.tb01054.x.

Gelder, Tim van. 1998. 'The Dynamical Hypothesis in Cognitive Science'. *Behavioral and Brain Sciences* 21 (5): 615–28. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/the-dynamical-hypothesis-in-cognitive-science/C121F1B65A534F3E7A27075EE489AD1E.

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.

Goh, K.-I., and A.-L. Barabási. 2008. 'Burstiness and Memory in Complex Systems'. *EPL (Europhysics Letters)* 81 (4): 48002. doi:10.1209/0295-5075/81/48002.

Goodman, Leo A. 1965. 'On Simultaneous Confidence Intervals for Multinomial Proportions'. *Technometrics* 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.

Grice, Herbert Paul. 1989. *Studies in the Way of Words*. Cambridge, MA, USA: Harvard University Press.

Griffiths, Paul E., and Russell D. Gray. 2005. 'Discussion: Three Ways to Misunderstand Developmental Systems Theory'. *Biology and Philosophy* 20 (2): 417–25. doi:10.1007/s10539-004-0758-1.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy: An Introduction*. http://dx.doi.org/10.1017/CBO9780511744082.

Griffiths, Thomas L., and Michael L. Kalish. 2007. 'Language Evolution by Iterated Learning with

Bayesian Agents'. *Cognitive Science* 31 (3): 441–80. doi:10.1080/15326900701326576.

Griffiths, Thomas L., Brian R. Christian, and Michael L. Kalish. 2008. 'Using Category Structures to Test Iterated Learning as a Method for Identifying Inductive Biases'. *Cognitive Science* 32 (1): 68–107. doi:10.1080/03640210701801974.

Griffiths, Thomas L., Michael L. Kalish, and Stephan Lewandowsky. 2008. 'Theoretical and Empirical Evidence for the Impact of Inductive Biases on Cultural Evolution'. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363 (1509): 3503–14. doi:10.1098/rstb.2008.0146.

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.

Haldane, John Burdon Sanderson. 1932. *The Causes of Evolution*. London, UK; New York, NY, USA; Toronto, Canada: Longmans Green & Co.

Hall, K. R. L. 1950. 'The Effect of Names and Titles Upon the Serial Reproduction of Pictorial and Verbal Material'. *British Journal of Psychology. General Section* 41 (3): 109–21. doi:10.1111/j.2044-8295.1950.tb00269.x.

Harvey, Matthew I. 2015. 'Content in Languaging: Why Radical Enactivism Is Incompatible with Representational Theories of Language'. *Language Sciences* 48 (March): 90–129. doi:10.1016/j.langsci.2014.12.004.

Heath, Chip, Chris Bell, and Emily Sternberg. 2001. 'Emotional Selection in Memes: The Case of Urban Legends'. *Journal of Personality and Social Psychology* 81 (6): 1028–41. doi:10.1037/0022-3514.81.6.1028.

Henst, Jean-Baptiste van der, and Dan Sperber. 2012. 'Testing the Cognitive and Communicative Principles of Relevance'. In *Meaning and Relevance*, edited by Deirdre Wilson and Dan Sperber, 279–306. Cambridge, UK: Cambridge University Press.

Heuven, Walter J. B. van, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. 'SUBTLEX-UK: A New and Improved Word Frequency Database for British English'. *The Quarterly Journal of Experimental Psychology* 67 (6): 1176–90. doi:10.1080/17470218.2013.850521.

Hofmann, Heike, Karen Kafadar, and Hadley Wickham. 2011. 'Letter-Value Plots: Boxplots for Large Data'. had.co.nz.

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.

Huber, Peter J. 1981. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. New York, NY, USA: Wiley. http://public.eblib.com/choice/publicfullrecord.aspx?p=224924.

Hunter, John D. 2007. 'Matplotlib: A 2d Graphics Environment'. *Computing in Science & Engineering*

9 (3): 90–95. doi:10.1109/MCSE.2007.55.

Hutto, Daniel D, and Erik Myin. 2013. *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, Mass.: MIT Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=3339551.

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. http://www.jstor.org/stable/25842543.

———. 2001. 'From the Transmission of Representations to the Education of Attention'. In *The Debated Mind: Evolutionary Psychology Versus Ethnography*, edited by Harvey Whitehouse, 113–53. Oxford & New York: Berg. http://lchc.ucsd.edu/MCA/Paper/ingold/ingold1.htm.

———. 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.

Jablonka, Eva M. 2001. 'The Systems of Inheritance'. In *Cycles of Contingency: Developmental Systems and Evolution*, edited by Susan Oyama, Paul Griffiths, and Russell D Gray, 99–116. Cambridge, Mass.: MIT Press.

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.

Jo, Hang-Hyun, Márton Karsai, János Kertész, and Kimmo Kaski. 2012. 'Circadian Pattern and Burstiness in Mobile Phone Communication'. *New Journal of Physics* 14 (1): 013055. doi:10.1088/1367-2630/14/1/013055.

Kalish, Michael L., Thomas L. Griffiths, and Stephan Lewandowsky. 2007. 'Iterated Learning: Intergenerational Knowledge Transmission Reveals Inductive Biases'. *Psychonomic Bulletin & Review* 14 (2): 288–94. doi:10.3758/BF03194066.

Kashima, Yoshihisa. 2000a. 'Maintaining Cultural Stereotypes in the Serial Reproduction of Narratives'. *Personality and Social Psychology Bulletin* 26 (5): 594–604. doi:10.1177/0146167200267007.

———. 2000b. 'Recovering Bartlett's Social Psychology of Cultural Dynamics'. *European Journal of Social Psychology* 30 (3): 383–403. doi:10.1002/(SICI)1099-0992(200005/06)30:3<383::AID-EJSP996>3.0.CO;2-C.

Keuleers, Emmanuel, and Marc Brysbaert. 2010. 'Wuggy: A Multilingual Pseudoword Generator'. *Behavior Research Methods* 42 (3): 627–33. doi:10.3758/BRM.42.3.627.

Keuleers, Emmanuel, Michaël Stevens, Paweł Mandera, and Marc Brysbaert. 2015. 'Word Knowledge in the Crowd: Measuring Vocabulary Size and Word Prevalence in a Massive Online Experiment'. *The Quarterly Journal of Experimental Psychology* 68 (8): 1665–92. doi:10.1080/17470218.2015.1022560.

Killingsworth, Matthew A., and Daniel T. Gilbert. 2010. 'A Wandering Mind Is an Unhappy Mind'. *Science* 330 (6006): 932–32. doi:10.1126/science.1192439.

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.

Kirby, Simon, Mike Dowman, and Thomas L. Griffiths. 2007. 'Innateness and Culture in

the Evolution of Language'. *Proceedings of the National Academy of Sciences* 104 (12): 5241–5. doi:10.1073/pnas.0608222104.

Kirby, Simon, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. 'Compression and Communication in the Cultural Evolution of Linguistic Structure'. *Cognition* 141 (August): 87–102. doi:10.1016/j.cognition.2015.03.016.

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

Krosnick, Jon A. 2000. 'The Threat of Satisficing in Surveys: The Shortcuts Respondents Take in Answering Questions'. *Survey Methods Newsletter* 20 (1): 4–8.

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.

Laland, Kevin, Tobias Uller, Marc Feldman, Kim Sterelny, Gerd B. Müller, Armin Moczek, Eva Jablonka, et al. 2014. 'Does Evolutionary Theory Need a Rethink?' *Nature News* 514 (7521): 161. doi:10.1038/514161a.

Lauf, Aurelien, Mathieu Valette, and Leila Khouas. 2013. 'Analyzing Variation Patterns in Quotes over Time'. *Research in Computing Science* 70: 223–32. http://www.micai.org/rcs/2013_70/Analyzing%20Variation%20Patterns%20In%20Quotes%20Over%20Time.html.

Leeuw, Edith D. de, Joop J. Hox, and Don A. Dillman. 2008. *International Handbook of Survey Methodology*. New York, NY, USA ; London, UK: Taylor & Francis.

Leskovec, Jure, and Andrej Krevl. 2014. 'SNAP Datasets: Stanford Large Network Dataset Collection', June. http://snap.stanford.edu/data.

Leskovec, Jure, and Rok Sosič. 2016. 'SNAP: A General-Purpose Network Analysis and Graph-Mining Library'. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8 (1): 1.

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.

Lewens, Tim. 2012. 'Cultural Evolution: Integration and Skepticism'. In *The Oxford Handbook of Philosophy of Social Science*, edited by Harold Kincaid, 458–80. Oxford; New York: Oxford University Press.

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. 'Evolu-

tion of Music by Public Choice'. *Proceedings of the National Academy of Sciences* 109 (30): 12081–6. doi:10.1073/pnas.1203182109.

Mackerron, G., and S. Mourato. 2013. 'Happiness Is Greater in Natural Environments'. *Global Environmental Change*. http://www.sciencedirect.com/science/article/pii/S0959378013000575.

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.

Maturana, Humberto R, and Francisco J Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Holland; Boston: D. Reidel Pub. Co.

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

Maxwell, R. S. 1936. 'Remembering in Different Social Groups'. *British Journal of Psychology. General Section* 27 (1): 30–40. doi:10.1111/j.2044-8295.1936.tb00814.x.

McGraw, John J., Sebastian Wallot, Panagiotis Mitkidis, and Andreas Roepstorff. 2014. 'Culture's Building Blocks: Investigating Cultural Evolution in a LEGO Construction Task'. *Frontiers in Psychology* 5 (September). doi:10.3389/fpsyg.2014.01017.

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.

Menary, Richard. 2010. *The Extended Mind*. Cambridge, MA, USA: MIT Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=3339152.

Mercier, Hugo, and Dan Sperber. 2011. 'Why Do Humans Reason? Arguments for an Argumentative Theory'. *Behavioral and Brain Sciences* 34 (2): 57–74. doi:10.1017/S0140525X10000968.

Mesoudi, Alex, and Andrew Whiten. 2004. 'The Hierarchical Transformation of Event Knowledge in Human Cultural Transmission'. *Journal of Cognition and Culture* 4 (1): 1–24. doi:10.1163/156853704323074732.

———. 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 Robin Dunbar. 2006. 'A Bias for Social Information in Human Cultural Transmission'. *British Journal of Psychology* 97 (3): 405–23. doi:10.1348/000712605X85871.

Mesoudi, Alex, Andrew Whiten, and Kevin N. Laland. 2006. 'Towards a Unified Science of Cultural Evolution'. *Behavioral and Brain Sciences* 29 (4): 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.

Miller, Geoffrey. 2012. 'The Smartphone Psychology Manifesto'. *Perspectives on Psychological Science* 7

(3): 221–37. doi:10.1177/1745691612441215.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, Mass.: MIT Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=49593.

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.

Mitkidis, Panagiotis, John J. McGraw, Andreas Roepstorff, and Sebastian Wallot. 2015. 'Building Trust: Heart Rate Synchrony and Arousal During Joint Action Increased by Public Goods Game'. *Physiology & Behavior* 149 (October): 101–6. doi:10.1016/j.physbeh.2015.05.033.

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.

———. 2016. *How Traditions Live and Die*. New York: Oxford University Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1074945.

Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. 'Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and Its Application to Bible Reuse'. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by Jian Su, Xavier Carreras, and Kevin Duh, 1849–59. Austin, TX: Association for Computational Linguistics.

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.

Myin, Erik. 2003. 'An Account of Color Without a Subject?' *Behavioral and Brain Sciences* 26 (1): 42–43.

Needleman, Saul B., and Christian D. Wunsch. 1970. 'A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins'. *Journal of Molecular Biology* 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4.

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.

Niculae, Vlad, and Cristian Danescu-Niculescu-Mizil. 2016. 'Conversational Markers of Constructive Discussions'. *arXiv:1604.07407 [Physics, Stat]*, April. http://arxiv.org/abs/1604.07407.

Norenzayan, Ara, Scott Atran, Jason Faulkner, and Mark Schaller. 2006. 'Memory and Mystery: The Cultural Selection of Minimally Counterintuitive Narratives'. *Cognitive Science* 30 (3): 531–53. doi:10.1207/s15516709cog0000_68.

Northway, Mary L. 1936. 'The Influence of Age and Social Group on Children's Remembering'. *British Journal of Psychology. General Section* 27 (1): 11–29. doi:10.1111/j.2044-8295.1936.tb00813.x.

Noveck, Ira, and Dan Sperber. 2012. 'The Why and How of Experimental Pragmatics: The Case of "Scalar Inferences"'. In *Meaning and Relevance*, edited by Deirdre Wilson and Dan Sperber, 307–30. Cambridge, UK: Cambridge University Press.

O'Brien, Michael J., and Kevin N. Laland. 2012. 'Genes, Culture, and Agriculture: An Example of Human Niche Construction'. *Current Anthropology* 53 (4): 434–70. doi:10.1086/666585.

O'Regan, J. Kevin, and Alva Noë. 2001. 'A Sensorimotor Account of Vision and Visual Consciousness'. *Behavioral and Brain Sciences* 24 (5): 939–73. doi:10.1017/S0140525X01000115.

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.

Oyama, Susan. 2000. *The Ontogeny of Information: Developmental Systems and Evolution*. Durham, NC: Duke University Press.

Oyama, Susan, Paul Griffiths, and Russell D Gray, eds. 2001. *Cycles of Contingency: Developmental Systems and Evolution*. Cambridge, Mass.: MIT Press.

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.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. 'GloVe: Global Vectors for Word Representation'. In, 1532–43. http://www.aclweb.org/anthology/D14-1162.

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.

Perfors, Amy, and Daniel J. Navarro. 2014. 'Language Evolution Can Be Shaped by the Structure of the World'. *Cognitive Science* 38 (4): 775–93. doi:10.1111/cogs.12102.

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.

Purzycki, Benjamin Grant, and Aiyana K. Willard. 2016. 'MCI Theory: A Critical Discussion'. *Religion, Brain & Behavior* 6 (3): 207–48. doi:10.1080/2153599X.2015.1024915.

Rayner, Keith, Timothy J. Slattery, and Nathalie N. Bélanger. 2010. 'Eye Movements, the Perceptual Span, and Reading Speed'. *Psychonomic Bulletin & Review* 17 (6): 834–39. doi:10.3758/PBR.17.6.834.

Reali, Florencia, and Thomas L. Griffiths. 2009. 'The Evolution of Frequency Distributions: Relating Regularization to Inductive Biases Through Iterated Learning'. *Cognition* 111 (3): 317–28. doi:10.1016/j.cognition.2009.02.012.

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.

Risjord, Mark. 2012. 'Models of Culture'. In *The Oxford Handbook of Philosophy of Social Science*, edited by Harold Kincaid, 387–408. Oxford; New York: Oxford University Press.

Roberts, Gareth, and Bruno Galantucci. 2017. 'Investigating Meaning in Experimental Semiotics'. *Psychology of Language and Communication* 20 (2): 130–53. doi:10.1515/plc-2016-0008.

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.

Rogers, Everett M. 2005. *Diffusion of Innovations*. New York, NY: Free Press.

Romney, A. K., J. P. Boyd, C. C. Moore, W. H. Batchelder, and T. J. Brazill. 1996. 'Culture as Shared Cognitive Representations'. *Proceedings of the National Academy of Sciences* 93 (10): 4699–4705. http://www.pnas.org/content/93/10/4699.

Roy, Brandon C., Michael C. Frank, Philip DeCamp, Matthew Miller, and Deb Roy. 2015. 'Predicting the Birth of a Spoken Word'. *Proceedings of the National Academy of Sciences* 112 (41): 12663–8. doi:10.1073/pnas.1419773112.

Roy, Deb, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, et al. 2006. 'The Human Speechome Project'. In *Proceedings of the 28th Annual Cognitive Science Conference*, edited by Ron Sun and Naomi Miyake, 2059–64. Austin, TX, USA: Cognitive Science Society.

Ruan, Zhongyuan, Gerardo Iñiguez, Márton Karsai, and János Kertész. 2015. 'Kinetics of Social Contagion'. *Physical Review Letters* 115 (21): 218702. doi:10.1103/PhysRevLett.115.218702.

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.

Scott-Phillips, Thomas C. 2017. 'Pragmatics and the Aims of Language Evolution'. *Psychonomic Bulletin & Review* 24 (1): 186–89. doi:10.3758/s13423-016-1061-2.

Scott-Phillips, Thomas C., and Simon Kirby. 2010. 'Language Evolution in the Laboratory'. *Trends in Cognitive Sciences* 14 (9): 411–17. doi:10.1016/j.tics.2010.06.006.

Scott-Phillips, Thomas C., Simon Kirby, and Graham R. S. Ritchie. 2009. 'Signalling Signalhood and

the Emergence of Communication'. *Cognition* 113 (2): 226–33. doi:10.1016/j.cognition.2009.08.009.

Scott-Phillips, Thomas C., Kevin N. Laland, David M. Shuker, Thomas E. Dickins, and Stuart A. West. 2014. 'The Niche Construction Perspective: A Critical Appraisal'. *Evolution* 68 (5): 1231–43. doi:10.1111/evo.12332.

Silvey, Catriona, Simon Kirby, and Kenny Smith. 2015. 'Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions'. *Cognitive Science* 39 (1): 212–26. doi:10.1111/cogs.12150.

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.

Smith, Kenny, and Elizabeth Wonnacott. 2010. 'Eliminating Unpredictable Variation Through Iterated Learning'. *Cognition* 116 (3): 444–49. doi:10.1016/j.cognition.2010.06.004.

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

Sperber, Dan, and Gloria Origgi. 2012. 'A Pragmatic Perspective on the Evolution of Language'. In *Meaning and Relevance*, edited by Deirdre Wilson and Dan Sperber, 331–38. Cambridge, UK: Cambridge University Press.

Sperber, Dan, and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*. Oxford UK & Cambridge USA: Blackwell.

Sterelny, Kim. 2001. 'Niche Construction, Developmental Systems, and the Extended Replicator'. In *Cycles of Contingency: Developmental Systems and Evolution*, edited by Susan Oyama, Paul Griffiths, and Russell D Gray, 333–49. Cambridge, Mass.: MIT Press.

———. 2010. 'Minds: Extended or Scaffolded?' *Phenomenology and the Cognitive Sciences* 9 (4): 465–81. doi:10.1007/s11097-010-9174-y.

———. 2012. *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA, USA; London, UK: The MIT Press.

———. 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.

Stotz, Karola. 2010. 'Human Nature and Cognitive–Developmental Niche Construction'. *Phenomenology and the Cognitive Sciences* 9 (4): 483–501. doi:10.1007/s11097-010-9178-7.

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.

Tamariz, Mónica, and Simon Kirby. 2015. 'Culture: Copying, Compression, and Conventionality'. *Cognitive Science* 39 (1): 171–83. doi:10.1111/cogs.12144.

Tamariz, Monica, T. Mark Ellison, Dale J. Barr, and Nicolas Fay. 2014. 'Cultural Selection Drives the Evolution of Human Communication Systems'. *Proceedings of the Royal Society of London B: Biological*

Sciences 281 (1788): 20140488. doi:10.1098/rspb.2014.0488.

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

Torrance, Steve. 2006. 'In Search of the Enactive: Introduction to Special Issue on Enactive Experience'. *Phenomenology and the Cognitive Sciences* 4 (4): 357–68. doi:10.1007/s11097-005-9004-9.

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.

Varela, Francisco, Evan Thompson, and Eleanor Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Verhoef, Tessa, Simon Kirby, and Bart de Boer. 2014. 'Emergence of Combinatorial Structure and Economy Through Iterated Learning with Continuous Acoustic Signals'. *Journal of Phonetics* 43 (March): 57–68. doi:10.1016/j.wocn.2014.02.005.

Wallot, Sebastian, Panagiotis Mitkidis, John J. McGraw, and Andreas Roepstorff. 2016. 'Beyond Synchrony: Joint Action in a Complex Production Task Reveals Beneficial Effects of Decreased Interpersonal Synchrony'. *PLOS ONE* 11 (12): e0168306. doi:10.1371/journal.pone.0168306.

Walt, Stéfan van der, S. Chris Colbert, and Gaël Varoquaux. 2011. 'The NumPy Array: A Structure for Efficient Numerical Computation'. *Computing in Science & Engineering* 13 (2): 22–30. doi:10.1109/MCSE.2011.37.

Ward, T. H. G. 1949. 'An Experiment on Serial Reproduction with Special Reference to the Changes in the Design of Early Coin Types'. *British Journal of Psychology. General Section* 39 (3): 142–47. doi:10.1111/j.2044-8295.1949.tb00213.x.

Watts, Duncan J. 2002. 'A Simple Model of Global Cascades on Random Networks'. *Proceedings of the National Academy of Sciences* 99 (9): 5766–71. doi:10.1073/pnas.082090499.

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.

Weller, Susan C, and Antone K Romney. 1988. *Systematic Data Collection*. Newbury Park, CA, USA: Sage Publications.

Weng, L., A. Flammini, A. Vespignani, and F. Menczer. 2012. 'Competition Among Memes in a World with Limited Attention'. *Scientific Reports* 2 (March). doi:10.1038/srep00335.

Weng, Lilian, Márton Karsai, Nicola Perra, Filippo Menczer, and Alessandro Flammini. 2015. 'Attention on Weak Ties in Social and Communication Networks'. *arXiv:1505.02399 [Physics]*, May. http://arxiv.org/abs/1505.02399.

Whiten, Andrew, Christine A Caldwell, and Alex Mesoudi. 2016. 'Cultural Diffusion in Humans and Other Animals'. *Current Opinion in Psychology*, Culture, 8 (April): 15–21. doi:10.1016/j.copsyc.2015.09.002.

Wilson, Deirdre, and Dan Sperber. 2002. 'Truthfulness and Relevance'. *Mind* 111 (443): 583–632.

doi:10.1093/mind/111.443.583.

———. 2004. 'Relevance Theory'. In *The Handbook of Pragmatics*, edited by Laurence R Horn and Gregory Ward, 607–32. Oxford, UK: Blackwell.

———. 2012. *Meaning and Relevance*. Cambridge, UK: Cambridge University Press.

Wimsatt, William C., and James R. Griesemer. 2007. 'Reproducing Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution'. In *Integrating Evolution and Development: From Theory to Practice*, edited by Roger Sansom and Robert N Brandon, 227–323. Cambridge, MA, USA: MIT Press.

Winters, James, Simon Kirby, and Kenny Smith. 2015. 'Languages Adapt to Their Contextual Niche'. *Language and Cognition* 7 (3): 415–49. doi:10.1017/langcog.2014.35.

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

Xu, Jing, Mike Dowman, and Thomas L. Griffiths. 2013. 'Cultural Transmission Results in Convergence Towards Colour Term Universals'. *Proc. R. Soc. B* 280 (1758): 20123073. doi:10.1098/rspb.2012.3073.

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.

Zheng, Dongping, Kristi Newgarden, and Michael F. Young. 2012. 'Multimodal Analysis of Language Learning in World of Warcraft Play: Languaging as Values-Realizing'. *ReCALL* 24 (3): 339–60. doi:10.1017/S0958344012000183.