Inherent and probabilistic naturalness

Luca Gasparri | luca.gasparri@cnrs.fr

AAM, October 2023. Published in *Philosophical Studies*, https://doi.org/10.1007/s11098-023-02070-x

Abstract: Standard accounts hold that regularities of behavior must be arbitrary to constitute a convention. Yet, there is growing consensus that conventionality is a graded phenomenon, and that conventions can be more or less natural. I develop an account of natural conventions that distinguishes two basic dimensions of conventional naturalness: a probabilistic dimension and an inherent one. A convention is probabilistically natural if it is likely to emerge in a population of agents, and inherently natural if its content is a regularity that scores high on relevant measures for naturalness. I motivate the proposal on conceptual grounds and then showcase its descriptive benefits by discussing two case studies in language: the tendency towards word-length optimality and the prevalence of shape opacity in spoken language vocabularies.

Keywords: conventions; naturalness; probability; arbitrariness; language

Acknowledgments: Thanks to Guido Andreolli, Piera Filippi, Hans-Johann Glock, Jeremy Kuhn, Christopher Piñón, Ronald Planer, Stuart K. Watson, Markus Wild and Bodo Winter for discussions on conventions and arbitrariness as I was developing the ideas presented in this paper. Thanks also to Guido Andreolli, Hans-Johann Glock and Gerhard Schaden for feedback on an earlier draft of the paper. Finally, thanks to two anonymous reviewers at this journal for very helpful comments on the submitted manuscript. All errors are mine.

1. Introduction

In English, sentences are written from left to right. In Standard Arabic, sentences are written from right to left. In some Western countries, people dress in black tie at cocktail parties. In parts of South Asia, the prevalent formal attire for men is the *sherwani*. Writing sentences from left to right and dressing in black tie at cocktail parties are examples of conventions. Minimally, a convention is a regular pattern of behavior a population of agents selects from among a pool of alternative regularities to address a coordination problem. On the classical analysis stemming from Lewis (1969), regularities of behavior must satisfy some individually necessary and jointly sufficient conditions to qualify as conventions. One such condition is that for a regularity R to constitute a convention, R must be *arbitrary*: the coordination endeavor that led to R must have had the possibility of giving rise to an alternative outcome.[1]

Think of regularities of behavior as solutions to games of coordination played by populations of infallible, ideally rational agents. Then imagine two games, G and G*, which differ solely in the following: G has exactly one optimal solution, whereas G* has $n > 1$ optimal solutions.[2] Game G has exactly one optimal solution and the players, being infallible and ideally rational, will inevitably select that solution. By contrast, G* has n optimal solutions. These will attract player preferences with equal strength, and their competition will have to be resolved on the basis of discretion or chance. Since, under the assumptions we have made about the players, the probability of the only option in the pool of optimal solutions to G is a perfect 1, G cannot lead to a convention. By contrast, G* will lead to a convention, since whatever regularity will end up emerging among the n providing an optimal solution to the game, it could have been arbitrarily replaced by $n - 1$ equally good outcomes.

Taken without further qualification, the idea that conventions must be arbitrary appears to invite us to think of conventionality as a dichotomous affair. Either a regularity is arbitrary, in which case it can be a convention, or it is not arbitrary, in which case it cannot be a convention. However, the matter is clearly more complex. Take the toy examples above and picture game G* under distant values of n. If $n = 2$ (e.g., a driving left vs. driving right scenario), each option in the pool of optimal solutions to the game has a probability of 0.5 and 1 competitor. If $n = 100$

[1] I stipulate that a regularity of behavior R is arbitrary if and only if R could (rationally) have been otherwise, and will use “arbitrariness” and “arbitrary” accordingly in the paper. Note that, according to Lewis, arbitrariness is just one among several conditions that regularities of behavior must fulfill to qualify as conventions for a population of agents. Other conditions include common knowledge and the expectation of conformity.

[2] Readers familiar with game theory should feel free to replace “optimal solutions” with “Nash equilibria”.

(e.g., choosing the same number among the first 100 primes), each optimal option has a probability of 0.01 and 99 competitors. In both cases the outcome of G* will be arbitrary, since in both cases the *n* of options the players could have selected while remaining ideally rational is greater than 1. However, under *n* = 100 there is a significantly greater number of ways G* could have ended while still yielding an optimal equilibrium.

This suggests a simple amendment: while a shared pattern of behavior has to be arbitrary to *some* extent to constitute a convention, it can be *more or less* so. Acknowledging that regularities that could not have been otherwise cannot be conventions does not mean that we have to treat conventionality as an all-or-none phenomenon. We can still differentiate between conventions that could *hardly* or *easily* have been otherwise, for instance. Thus, we should think of conventionality as a graded affair. In some recent work, this provision has been discussed under the heading of the claim that conventions can be more or less *natural*. However, parties to this emerging consensus have interpreted the notion of a "natural convention" in different ways. Which raises the question of what exact property, or collection thereof, we should understand "being a natural convention" to entail.

The goal of this paper is to develop an account of conventional naturalness that can lay claim to independent plausibility and streamlines the recent literature on the topic. The core claim I will defend is that an account of natural conventions should distinguish two basic dimensions of conventional naturalness: a *probabilistic* dimension, tracking the likelihood that a regularity of behavior becomes a convention at a population of agents, and an *inherent* dimension, tracking relevant properties of the regularity of behavior in play.

The plan is as follows. Section 2 reviews some accounts of natural conventions. Section 3 identifies some junctures of conceptual instability in the theoretical landscape. Section 4 introduces the two-dimensional account and argues that it stabilizes the theoretical landscape. Section 5 applies the account to two case studies in language: the tendency towards word-length optimality and the prevalence of shape opacity in spoken language vocabularies. Section 6 recapitulates and offers some concluding remarks.

2. Three pleas for natural conventions

Let us start by unpacking how the notion of a "natural convention" has been interpreted in some recent pleas for a graded view of conventionality. I will review *three* works on the topic: Cumming, Greenberg & Kelly (2017) (henceforth, CG&K), Simons & Zollman (2019) (henceforth, S&Z), and O'Connor (2021) (henceforth, OC).

CG&K focus on narrative coherence in film, and appeal to a notion of "natural convention" as part of their case for a semantic view of film. Films consist of series of discrete representational units, their shots, distributed on a linear temporal sequence. Film content is determined by the representational content of the shots, and by the way the shots are arranged on the film's sequence. In principle, each shot could contribute a self-standing chunk of content bearing no overt or implied relationship with that of neighboring shots. But this is not how films work. Filmmakers employ a variety of techniques (most notably, montage) to arrange shot sequences so as to convey coherent *stories*.

Consider a simple POV transition featuring an initial shot representing a character's face or eyes, immediately followed by a shot representing an object. The transition triggers an inference that the object is the content of the character's visual experience, even if this relation is not overtly depicted in the sequence. This technique, CG&K observe, is a conventional solution to the problem of representing that something is seen or observed by a character. Filmmakers could have converged on other methods of representation to trigger the desired inference. Yet, the convention makes psychological sense: it triggers the inference by piggybacking on our pre-reflexive disposition to transition "from noticing a glance to looking at the object of that glance" (CG&K: 1–2).[3] According to CG&K, who build on Metz (1974), Eco (1976), and Bordwell

[3] For a more in-depth description of the workings of the POV convention, see Cumming et al. (2021).

(2007), editing patterns like the POV transition should therefore be viewed as "natural conventions": conventional patterns of representation that warrant an ascription of naturalness because they align with relevant dispositions in the relevant population of agents (here, the psychology of film viewers). As a result of their concordance with widespread psychological or cultural inheritance, these "natural" conventions have high chances of emerging, and can be easily interpreted even without prior exposure to the regularity.[4]

Next, S&Z. S&Z develop a notion of a "natural convention" as part of an argument that the conventions of indirect speech (in particular, requests) should be viewed as "highly natural". S&Z note that conventions can differ from one another not only in terms of how widespread they are in a population, but also in relation to the properties of the coordination problem they address, in relation to how well they address it, and in relation to how they emerged. They then argue that these differences reveal the existence of a continuum of naturalness within the realm of the conventional. Building on the Lewisian analysis, on Morgan (1978), Clark (1996), Millikan (2005), and again Bordwell (2007), S&Z identify three criteria for conventional naturalness.[5] The first is QUALITY: a convention K is more natural than a convention K* if K yields better rewards than K*. The second is STABILITY: a convention K is more natural than a convention K* if K is more stable than K*. The third is AVAILABILITY: a convention K is more natural than a convention K* if K has higher chances of emerging than K*.

QUALITY echoes a conceptual wisdom implicit in the Lewisian analysis. In addressing a coordination problem that allows for multiple solutions, any regularity of behavior that solves

[4] "Given much common psychological and cultural inheritance, certain regularities are natural for beings like us to follow. With little or no explicit learning, these will tend to be the ones that become entrenched as conventions. Indeed, because of this feature, natural conventions may be seamlessly grasped by viewers with no prior exposure, even in the course of interpreting a film" (CG&K: 6).

[5] "[T]hree ways that conventions might vary" and be "continuous with non-conventional behavioral regularities in important ways" (S&Z: 6).

the problem is a possible convention. However, the regularities involved may have varying degrees of optimality. Optimally rewarding regularities are “natural” because their conventionalization can be rationally explained in terms of preference for higher payoffs. Regarding STABILITY, the idea is to classify conventions as “natural” on condition that they exhibit a high resistance to perturbations. Patterns of behavior that reward optimally but fall apart if a small portion of the population deviates from them are unlikely to become entrenched as conventions. So stable regularities are more “natural” than their less resilient competitors. As for AVAILABILITY, the thought is to regard conventions as more or less “natural” based simply on their probability of emergence. S&Z exemplify the criterion with Schelling’s (1960) theory of focal points and the classic New York City question. Imagine you have to meet a stranger in NYC tomorrow. All channels of communication are blocked, and you and the stranger cannot coordinate on when or where to meet. Where would you go, and when would you go there? A surprising number of respondents say that they would wait at Grand Central Station at noon. Meeting at Grand Central Station at noon is “focal”: any other [*place*, *time*] pair would be exactly as good a solution in terms of QUALITY and STABILITY, but [Grand Central, noon] somehow strikes participants as an obvious answer to the task.[6] In this context, the choice of [Grand Central, noon] is “natural” precisely because it is AVAILABLE: it has a higher probability of emergence over comparable equilibria.

Finally, OC. OC builds on the criteria for naturalness proposed by S&Z, and develops an information-theoretic framework for measuring how AVAILABLE a convention is. She defines the arbitrariness of a convention as the degree to which it “could have been otherwise”,[7] and

[6] Quick reminder on the terminology. In standard game-theoretic parlance, an outcome is “focal” when it is salient *and* optimal, and therefore when its salience does not decrease the chances of selecting a better regularity. Outcomes can be salient without being focal, in which case they may attract player preferences even if they do not fit Nash predictions (e.g., Leland & Schneider 2018; Chowdhury et al. 2021).

[7] I have adopted the same stipulation. See above, fn. 2, and below, fn. 15.

associates conventional naturalness with low arbitrariness, and thus high AVAILABILITY. To illustrate, suppose a population α is playing a game of coordination that can give rise to exactly two conventions of behavior: K and K*. Suppose, further, that the probability of K and K* is determined solely by the rewards they would distribute if they were adopted by α (as was the case with the idealized games of Section 1), and that K would reward α better than K*. In this context, K will be more likely to emerge, hence less arbitrary, hence more "natural" than K*.

One goal of OC's framework is to improve explanation in cultural evolution and add nuance to standard accounts of the dichotomy between functional and conventional traits. Evolutionary work on cultural traits (e.g., patterns of social inequity; see Cochran & O'Connor 2019) often identifies conventionality with the absence of adaptive benefits. If a cultural regularity R confers a boost in fitness, then R's emergence is not arbitrary and R cannot be conventional. If R does not yield any boost in fitness, then R's emergence is arbitrary and R must be conventional. However, R may be highly functional but compete for selection with a large pool of similarly adaptive traits, and therefore represent a conventional outcome among the options in that pool. Or R may be non-conventionally selected despite a low adaptive power because of the absence of (better) alternatives. Thinking that conventions come in so defined measures of "naturalness" can help cultural theory make sense of this complexity.

3. An uneven terrain

Now for some critical remarks on the theoretical situation. To begin with, there are some macroscopic discrepancies between the batteries of criteria we have surveyed. For CG&K, a convention is natural if it aligns with relevant predispositions (cultural, psychological, and so forth) in a population of agents. Call this, for brevity, ACCORD. CG&K add that arbitrary patterns that satisfy ACCORD have high chances of becoming entrenched as conventions. S&Z propose three criteria: likelihood of emergence (AVAILABILITY), the resistance to perturbations

(STABILITY), and the payoff optimality of the pattern (QUALITY). S&Z do consider dispositional alignment among the factors that can make a convention AVAILABLE,[8] but since their focus is on the way we rationally (not dispositionally) settle competitions among alternative regularities, they do not commit to any version of ACCORD. For them, "'natural' does not […] necessarily mean explainable in terms of cognitive or cultural predispositions" (S&Z: 9). Finally, OC models her account of conventional naturalness on S&Z's AVAILABILITY, viewing it as inversely proportional to the ease with which a convention "could have been otherwise". OC acknowledges QUALITY and STABILITY as legitimate criteria for naturalness, but leaves them in the background for the purposes of her analysis.

There is a common thread: CG&K, S&Z and OC converge on the idea that conventional naturalness is linked with probability of emergence. However, note that in CG&K high probability of entrenchment is merely a probable correlate of naturalness. Assume, as seems plausible, that for CG&K patterns like the POV convention are natural because they fit the psychology of film viewers. Assume therefore that in CG&K's framework naturalness is settled by the satisfaction of ACCORD. A framework where conventional naturalness is settled by the satisfaction of ACCORD logically allows for unnatural conventions with high chances of emergence (e.g., conventions that do not align with the associative propensities of their users but nevertheless have a high likelihood of entrenchment), whereas in a framework where AVAILABILITY is a criterion for naturalness, a convention cannot have high chances of emergence while failing to be natural. So while something reasonably close to AVAILABILITY does feature in CG&K's theory, its role in the account is different from the one it plays in S&Z and, by extension, in OC. In CG&K, high probability of entrenchment is a *ceteris paribus* consequence of the naturalness of a convention. In S&Z, it *is* conventional naturalness.

[8] Focal outcomes are a case in point. In the NYC question, [Grand Central, noon] is AVAILABLE because it is psychologically salient, and psychological salience is plausibly a form of ACCORD.

Then, there are some more subtle differences among the criteria themselves and among the ways they are formulated by CG&K, S&Z and OC. Below I discuss three consecutive points: i) in S&Z, AVAILABILITY is a future-oriented criterion that applies to candidate conventions, whereas OC's statement of the measure is counterfactual and ranges on actual conventions; ii) QUALITY, STABILITY, and ACCORD track properties of regular behaviors, whereas AVAILABILITY tracks a property of events of conventionalization; iii) the QUALITY, the STABILITY and the ACCORD of a pattern of behavior can be causes for the AVAILABILITY of its conventional entrenchment, but not the other way around.

First point. S&Z's formulation of AVAILABILITY is officially about the likelihood that a regularity of behavior will become a convention. By contrast, OC's formulation of the criterion (under the conceptual umbrella of arbitrariness) is officially about the ease with which a convention could have been otherwise. The two properties are not mutually entailing: the counterfactual property entails the future-oriented one, but the future-oriented property does not entail the counterfactual one. Specifically, if a regularity R is a conventionally entrenched pattern that could hardly have been otherwise, it follows that in its pre-entrenchment career, R had a high probability of becoming a convention. Conversely, if a regularity R has a high probability of becoming a convention, it does not follow that R is a convention that could hardly have been otherwise, for the simple reason that the future is open and, despite its high probability of entrenchment (which, recall, must be lower than 1 for R to possibly give rise to a convention), R may fail to become a convention. In other words, the property tracked by the counterfactual reading of the measure can only be instantiated by conventionalized patterns of behavior (i.e., *real* conventions). By contrast, S&Z's formulation is about likelihood of conventionalization, and therefore tracks a property that can only be instantiated in the pre-entrenchment career of a pattern of behavior (i.e., *possible* conventions).

This connects to the second point. The capacity to distribute optimal rewards (QUALITY), the resistance to deviations (STABILITY), and the concordance with the predispositions of the population (ACCORD), are all possible properties of patterns of behavior. By contrast, AVAILABILITY designates a property that can be attributed non-derivatively only to events of conventional entrenchment. The thought is the following. Assume a face-value reading of AVAILABILITY whereby x is AVAILABLE *iff* $P(x)$ is high and P is vanilla probability. So $P(x)$ is the probability that x happens. What should we understand the variable x to range over? The variable cannot range simply over regularities of behavior, since the mere probability that a regularity emerges in a population (*sic*) is agnostic about the chances that the regularity in play becomes a convention. Thus, construing AVAILABILITY by having x stand for mere regularities of behavior would render the criterion moot for a theory of natural conventions. To avoid that, the variable must range over events of conventional entrenchment of regularities of behavior.

To illustrate, suppose a population α has to solve a game-theoretic predicament G at a time *t*. Suppose, further, that there are two regularities of behavior R and R*, both potential solutions to G, and that R is vastly more salient than R*. So α will almost certainly try to solve G by converging on R at *t*. Suppose, finally, that R is salient (high ACCORD) and resilient (high STABILITY) but vastly sub-optimal (low QUALITY), and that since α is rational, R will almost certainly be retracted at a time *t′* close to *t*. Next, α will almost certainly turn to R*, which by contrast offers a good solution to G, and thus will almost certainly become a convention at α. What should we regard as AVAILABLE in this context? An orderly use of the vocabulary we have introduced compels us, I believe, to say that the *conventionalization of R** is AVAILABLE, despite the fact that R is more salient than R* and therefore that R *simpliciter* has a higher probability of emerging than R* at the time of α's first encounter with G.

This connects in turn to the third point: the QUALITY, the STABILITY and the ACCORD of a regular behavior can be causes for the AVAILABILITY of its conventional entrenchment, but not the other way around. Suppose an arbitrary regularity R is resilient (STABILITY), distributes good rewards (QUALITY), and fits the dispositions of a population α (ACCORD). All these properties can – and, absent defeating factors, will – cause R to have a high probability of conventional entrenchment, and may be invoked as part of an explanation for why the conventionalization of R is likely to occur at α. By contrast, suppose the conventionalization of R is AVAILABLE at the population α. That the conventionalization of R is highly probable cannot cause R to be resilient, to distribute good rewards, and to align with the predisposition of α.

Here is a real-world example. In children the "arms up" gesture is initially instrumental and sub-intentional, and starts off as a reflex-like adjustment to the bodily actions performed by caregivers in picking children up. It then undergoes a process of ontogenetic conventionalization that turns it into a stable communicative signal children use to convey the request to be picked up (Burling 2005: 105-111). It is perfectly possible to say that the consolidation of the "arms up" gesture into a conventional signal is AVAILABLE because its content is a behavior that scores high on ACCORD (i.e., it piggybacks on the "natural" anatomical affordances of the human body). Conversely, it is clearly not possible to say that raising one's arms to ask to be picked up scores high on ACCORD because its consolidation into a conventional signal is AVAILABLE.

4. Inherent and probabilistic naturalness

If the above is correct, there are a few inconsistencies and loose ends in the theoretical landscape, and the situation would benefit from some streamlining. To stabilize things, I suggest that an account of natural conventions should distribute measures of conventional naturalness along two basic dimensions: a *probabilistic* dimension and an *inherent* one.

Probabilistic naturalness is a statistical dimension tracking likelihood of conventional entrenchment. An actual convention K whose content is a regularity of behavior R is probabilistically natural at a population α if the conventionalization of R was highly likely to occur at α. By contrast, inherent naturalness is a structural dimension tracking relevant ways in which the regularity of behavior constituting a convention can be regarded as natural. So, an actual convention K whose content is a regularity of behavior R is inherently natural at a population α if, to borrow from QUALITY, STABILITY and ACCORD, R distributes (near-)optimal rewards, and/or R is resistant to perturbations, and/or R accords with the predispositions of α.[9]

Probabilistic naturalness is a singular, closed dimension. An actual convention K whose content is a regularity R ranks high on the scale of probabilistic naturalness if the conventionalization of R had, either absolutely or comparatively, significant chances of occurring. This is all probabilistic naturalness consists of (hence the dimension is "singular"), and likelihood of conventional entrenchment exhausts the factors that can be taken into account in assessing the probabilistic naturalness of a convention (hence the dimension is "closed").

By contrast, inherent naturalness is plural and open. Assume, building on CG&K, S&Z and OC, that the concordance with the predispositions of the population (ACCORD), the distribution of optimal rewards (QUALITY), and the resilience of the pattern (STABILITY) are all valid sub-measures of inherent naturalness. The "pluralism" of the dimension indicates that inherent naturalness is a complex feature determined by how much a convention satisfies each relevant sub-measure for inherent naturalness. These sub-measures may be applied individually or in conjunction depending on one's descriptive or explanatory needs. The "openness" of the dimension indicates that QUALITY, STABILITY and ACCORD are not guaranteed to exhaust the factors that can be taken into account in assessing the inherent naturalness of a convention, and that other criteria may be added to the dimension.

[9] I use the adjective "inherent" simply to highlight the contrast between the naturalness of events of conventional entrenchment and the naturalness of the *quid* raised to the rank of a convention by events of conventional entrenchment. As should be clear, inherent naturalness does not have to concern categorical or intrinsic properties, and can concern dispositional or relational properties (as ACCORD does).

Think of cases where chances of conventionalization are boosted or undercut by the absence or presence of a geological boundary or of a natural resource.[10] For instance, suppose a population α about to established an organized religion lives in an environment where clay is abundant and accessible. Population α will be more likely to develop rituals involving objects made of clay than a population β occupying a territory where clay is scarce, even if other materials would have afforded matching levels of conventional resilience (equal STABILITY), the social rewards distributed by the clay-based rituals could have been produced with other materials (equal QUALITY), and neither α or β were particularly predisposed towards using clay in the first place (equal ACCORD). In this scenario, one could reasonably stipulate that clay-based rituals are highly inherently natural at α in light of the resources available in α's environment, and therefore add to the measures of inherent naturalness a criterion of ECOLOGY tracking a convention's reliance on environmental opportunities.

Teasing inherent and probabilistic naturalness apart yields an account of natural conventions that incorporates the insights of CG&K, S&Z and OC into a unitary framework while addressing the discrepancies highlighted in Section 3. The first point was that in S&Z AVAILABILITY is a future-oriented criterion applying to candidate conventions, whereas OC's statement of the measure is counterfactual, it applies to actual conventions, and has entailment privileges over its future-oriented variant. The proposed formulation of probabilistic naturalness removes the ambiguity. By saying that an actual convention K with content R is probabilistically natural at a population α if the conventionalization of R was highly likely to occur at α, we center the analysis on "real" conventions,[11] we bring the counterfactual property (with its entailment

[10] See, e.g., Claidière, Scott-Phillips & Sperber (2014) on the role of ecological variables in cultural evolution. Please note that my objective here is merely to illustrate the open-ended nature of the rubric of inherent naturalness. It is not to provide an argument that we should commit to the criterion I am about to sketch.

[11] As one should, at least assuming that a theory of conventional naturalness should first and foremost provide descriptive purchase on the naturalness of actual conventions.

privileges) to the forefront, and we leave its future-oriented counterpart for possible conventions. The second point was that QUALITY, STABILITY, and ACCORD track properties of patterns of behavior, whereas AVAILABILITY tracks a statistical feature of events of conventionalization. The two dimensions compartmentalize the criteria on the basis of their proper target: we have probabilistic naturalness for events of conventional entrenchment and inherent naturalness for the *quid* raised to the rank of convention by events of conventional entrenchment. The third point was that high QUALITY, high STABILITY, and high ACCORD can be causes for high AVAILABILITY, but not the other way around. We can now formulate the point in a generalized fashion by saying that high scores on one or multiple criteria for inherent naturalness can cause a convention to be probabilistically natural. Conversely, a high level of probabilistic naturalness cannot cause a convention to be inherently natural.

A two-dimensional account coupling a closed criterion of probability of conventionalization with an open-ended battery of measures for inherent naturalness is not only, as I have tried to argue, independently plausible; it also parallels extant practice in work on cultural evolution. For instance, in Cultural Attractor Theory (CAT), cultural attractors are theoretical posits that capture the way in which certain ideational variants (e.g., tool traditions or symbolic conventions) are more likely to be the outcome of cultural transformations than others (Buskell 2017). Cultural attractors come in different kinds, and CAT-style epidemiology distinguishes between cases where the spread of a regularity is likely to occur because it fits the existing cultural inclinations of a population, cases where a regularity is likely to become a convention because it provides an effective solution to a problem, and so forth. The analogy should be clear enough.[12] Probabilistic naturalness corresponds to the statistical feature of cultural magnetism defeasibly generated by all cultural attractors: the tendency towards the dissemination and the

[12] The comparison should not be interpreted as an implicit endorsement of CAT. The observation I am making is merely that the two-dimensional account parallels the descriptive approach of a popular framework in the theory of cultural evolution. The point would remain even if CAT turned out to be irremediably flawed.

entrenchment of an ideational variant. In turn, the measures for inherent naturalness operate on par with the open-ended system of "motives of attraction" which, in frameworks like CAT, defeasibly account for why an ideational variant is likely to become dominant in a population.

5. Optimized messages and opaque words

So far for the principled arguments. To see how the proposal behaves when applied to a concrete case study, enter linguistic conventions. Language has always been a key concern for theories of conventions, and is unsurprisingly central to the contributions discussed in Sections 2 and 3. CG&K's account is inspired by work in semiotics. S&Z focus on indirect request but suggest that the recognition of natural conventions paves the way for new observations about linguistic meaning in general. OC applies her measure of arbitrariness to indirect requests, and adds to the mix basic sentence structure and the conventionality of color terms. Here I will explore two new applications: the tendency towards word-length optimality and the prevalence of shape opacity in spoken vocabularies. The choice of these case studies rests on two main reasons. First, I want to focus on a basic aspect of the organization of languages and of the functional hierarchy of the grammar: how vocabularies are built. Second, I want to examine two cases where probabilistic naturalness and key sub-measures of inherent naturalness exhibit inverse patterns of correlation. As we will see, there is an argument that probabilistically natural patterns of word length score high on inherent naturalness, whereas a spoken lexicon with substantial regions of shape opacity is a probabilistically natural occurrence in spite of its inherent unnaturalness.[13]

[13] Please note that while the distinction between probabilistic and inherent naturalness is neutral about the extent to which actual linguistic conventions are probabilistically or inherently natural, the verdict may be influenced by your preferred grammatical framework. For instance, if you adhere to a generalized version of Optimality Theory (Prince & Smolensky 2004) on which all aspects of linguistic grammars (beyond just phonology) result from optimal constraint satisfaction, you might be inclined to believe that actual linguistic conventions are more inherently natural (assuming optimal constraint satisfaction entails high QUALITY) and more probabilistically natural than linguists with different grammatical allegiances may be willing to grant. In the discussion below I

Let us start with word lengths. Word lengths are conventional. Short words could have been longer, polysyllabic words could have had half their syllables, and languages differ in the mean phonological size of their vocabulary items (Wichmann, Rama & Holman 2011). However, there are multiple constraints on how word lengths are distributed in a language. For instance, from a standpoint of economic efficiency, it would be inconvenient for a community of speakers to use sesquipedalian word forms for common objects, and concise word forms for ecologically infrequent referents. Languages tend to be Zipfian (1935): more frequent words tend to be short, whereas less frequent words tend to be long. Also, frequent words tend to be conceptually simple, whereas less frequent words tend to express more complex concepts, according to various well-defined measures of conceptual complexity. Hence, words for conceptually simple meanings like 'cat' tend to be short, whereas words for conceptually more complex meanings like 'polyphiloprogenitive' tend to be long (Piantadosi 2014; Dahl & Walter 2020).[14]

The use of the verb "tend" is not casual, as we are actually dealing with a propensity towards an optimum achieved in different degrees across languages. While attested languages do tend to map frequent messages to short words, the magnitude of this effect exhibits important cross-linguistic variation and has outliers. For example, in English 'wen' is short but infrequent, whereas 'happiness' is long but frequent (Pimentel et al. 2021). Natural languages can be distant from compression optimality, and ideal code efficiency has proven challenging to reproduce even in artificial agents (Rita, Chaabouni & Dupoux 2020). Yet, however imperfectly, vocabularies do tend to optimize transmission by compressing the length of the more frequent and simple messages, while relaxing the production costs for the more infrequent and complex ones. Crucially, this tendency seems to be tacitly known and exploited by learners in formulating

will remain as non-partisan as possible about the nuts and bolts of the grammar.

[14] I will only discuss word lengths here, but bear in mind that principles of economy shape languages well beyond the realm of word lengths. E.g., see Rooij (2004) on the relationship between Zipfian optimality, Grice's (1989) Maxim of Quantity and Horn's Principle (Horn 1984).

hypotheses about the meaning of unfamiliar words. Learners tend to associate long fictional words with complex objects in comprehension and production tasks, and judgments of conceptual complexity for real words correlate with their length across several languages (Lewis & Frank 2016). Zipfian distributions also appear to facilitate word segmentation for learners (Kurumada, Meylan & Frank 2013). This suggests that word length biases learners towards specific hypotheses about semantic complexity, and that in cases where the mapping between forms and meanings is in fact Zipf-optimal, lexical acquisition is facilitated (Hendrickson & Perfors 2019). This, in turn, suggests that lexical learning is aided by a hardwired disposition to inversely correlate word length with statistical frequency and semantic complexity.

We can use probabilistic and inherent naturalness to describe this. Languages where the conventions of word length tilt towards the Zipfian optimum are probabilistically natural. Though ideal compression is hard to achieve, sufficiently compressed systems are more likely to become conventionally entrenched than massively inefficient vocabularies. Furthermore, the probabilistic naturalness of optimized vocabularies can be attributed to the fact that Zipfian lengths meet key criteria for inherent naturalness. In particular, systems of form-meaning mappings that satisfy the efficiency predictions perform well on the measures of QUALITY, STABILITY, and ACCORD. They minimize production and perception costs for frequent words (QUALITY), they are less likely to undergo further processes of optimization (STABILITY), and they tap into some hardwired psychological propensities of human learners (ACCORD).

However, high values of inherent naturalness only defeasibly cause optimized associations between forms and meanings to have high chances of conventional entrenchment. The link can be severed, e.g., in constrained word introduction games. Suppose a population α of speakers of a language L has to coordinate on a novel word form for a novel meaning M afforded by, say, the rapid spread of a new technology. M is conceptually simple and frequent, but L is lexically overcrowded. Suppose, further, that among the word forms permitted by L's phonology, all the short forms are already in use for other meanings, and that α has strong incentives to minimize ambiguity and avoid assigning M a form homonymous with another word of L. Under these constraints, it is probabilistically natural that M will be conventionally assigned a long word, even if a long word is an inherently unnatural choice for a referent as frequent as M.

I now turn to the prevalence of shape opacity in spoken language vocabularies.[15] Most words in languages like English bear no correspondence between form and meaning. The nouns 'table', 'hat', and 'green' are conventionally tasked with denoting tables, hats and the color green, but their shapes do not offer cues to their meaning. There are, however, translucent words,[16] i.e., words whose shape stands in a relation of perceptual or inferentially exploitable correspondence to the things they denote. Iconic words like onomatopoeias ('buzz', 'ticktock', 'boom') are the paradigm example.

At least since de Saussure (1916), the received wisdom in the language sciences has been that spoken vocabularies are overwhelmingly opaque, with vanishingly rare instances of shape translucency.[17] This consensus is shifting. Translucent correspondences between word forms and meanings are more widespread than previously assumed (Perniss & Vigliocco 2014; Monaghan et al. 2014). For instance, words denoting round objects have a statistically observable tendency to contain phones articulated with round lips (like [o] or [ʊ]); in English, words ending with -*ash* have a sound-symbolic propensity to be associated with abrupt contact ('smash', 'crash',

[15] What I am about to discuss is usually referred to as the principle of the "arbitrary sign". However, I am going to steer clear of arbitrariness-talk to avoid confusion with the interpretation of the label adopted thus far. For more on the interplay of arbitrariness *qua* availability of alternative outcomes or options, and arbitrariness *qua* opaque mapping between form and meaning, see Planer & Kalkman (2021) and Gasparri et al. (2023).

[16] The adjective is borrowed from Pateman's (1986) distinction between "transparent" and "translucent" signs. I adopt the label to signal that even in textbook examples of iconic words elements of transparency are invariably intertwined with residues of opacity.

[17] The comment does not extend to sign languages, where iconicity and other forms of translucency have always been known to be more prevalent. See, e.g., Liddell (2003).

‘mash’); across languages, lexical items pertaining to the same conceptual domain have been found to converge towards form similarity (Dingemanse et al. 2015; Blasi et al. 2016; Dautriche et al. 2017; Winter & Perlman 2021).

In addition to the interest of providing a more accurate estimate of the relative prevalence of shape opacity in spoken vocabularies, research in this area is motivated by the increasingly well-attested advantages that shape translucency provides in word learning. Because their shape contains cues about meaning, translucent words are easier to master for language learners and tend to be acquired earlier than their opaque counterparts in language development (Imai & Kita 2014; Lockwood, Dingemanse & Hagoort 2016; Nielsen & Dingemanse 2021; Sidhu et al. 2021). Iconic mappings between word forms and meanings have been found to emerge spontaneously in human experiments simulating the emergence of novel signaling systems (Kempe et al. 2021), and evidence indicates that their reliance on perceptual and associative mechanisms makes them more resistant to semantic change (Monaghan & Roberts 2021).

So, spoken vocabularies seem to gravitate towards intelligent ratios of shape opacity and translucency. This tendency can be seen as an adaptive (hence, at least partly rationally explicable) result of the compromise between two attractors. The first attractor are the learning and processing benefits afforded by shape translucency. Iconic words are easier to acquire and process because they tap into our perceptual propensity to exploit sign-signified resemblance relations in formulating hypotheses about meaning. Likewise, words exhibiting semantically motivated patterns of shape resemblance are easier to memorize because they tap into our propensity to associate similar forms with similar meanings. The second attractor is expressive power. In spite of its benefits, spoken vocabularies can only incorporate shape translucency to a modest extent. For instance, iconicity is constrained by the limited amount of sounds that can be produced by the human vocal tract, and a vocabulary where identity in sound reliably corresponds to identity in meaning would struggle to express the diversity of meanings encoded by natural lexicons. Looked at through the lens of these constraints, the emergence of sizable amounts of decoupling between word forms and meanings appear something of an inevitability in language evolution. Opaque words entail learning costs. But they provide a rational solution to the problem of constructing a system where the limited inventory of speech sounds available to humans can manage to convey such an astounding variety of meanings.

Distinguishing inherent and probabilistic naturalness can help us tease apart the rival senses in which a large spoken lexicon with considerable amounts of shape opacity can be regarded at once as a highly "natural" and as a highly "unnatural" technology of communication for creatures like us. Shape opacity is inherently unnatural. It appears to be associated with low STABILITY (opaque words are prone to semantic change), low QUALITY (opaque words require a higher learning investment and their manipulation is cognitively more costly), and low ACCORD (opaque words cannot be interpreted on the basis of our perceptual or associative propensities). By extension, a large lexicon where shape opacity is prevalent is an inherently unnatural technology of communication for creatures like us: it is unstable, cognitively costly, and hard to learn. Yet, given the pressure to differentiate the units of the system under the constraints of our articulators, the conventionalization of large lexicons with considerable amounts of shape opacity had to be likely to surface in language evolution. Thus, the prevalence of shape opacity is a probabilistically natural feature of the conventions of present-day spoken vocabularies.

The reverse goes for shape translucency. Shape translucency appears to be associated with higher STABILITY (translucent words are better conserved than their opaque counterparts), higher QUALITY (translucent words have lower learning and processing costs), and higher ACCORD (learners can guess the meaning of translucent words by exploiting their perceptual and associative propensities). By extension, a large lexicon where shape translucency is ubiquitous would be an inherently natural technology of communication for creatures like us: it would be stable, cognitively cheap, and easy to learn. Yet, given the pressure to differentiate the units of the system under the constraints of our articulators, the conventionalization of large translucent lexicons had to be unlikely to surface in language evolution. Our species had powerful incentives to develop systems of lexical conventions that are unnaturally difficult to use, remember, and pass from generation to generation. In sum, there is an argument that an inherently unnatural lexicon with large swaths of opaque words is nonetheless a probabilistically natural convention of acoustic communication for creatures like us.

6. Conclusion

I have argued that an account of natural conventions should distribute measures of conventional naturalness along two basic dimensions. A probabilistic dimension, tracking the likelihood that a regularity of behavior becomes a convention at a population of agents; and an inherent dimension, encompassing multiple sub-measures of naturalness for the regularity of behavior in play. I have argued that this approach to natural conventions generates conceptual benefits. For instance, it incorporates the criteria suggested by CG&K, S&Z and OC into a unitary framework and organizes them as a function of their proper target (behavioral regularities vs. events of conventional entrenchment). Finally, I have showcased the descriptive applicability of the proposal with two case studies in language: the tendency towards Zipfian optimality and the prevalence of shape opacity in spoken vocabularies.

I conclude with a teaser. The applications of Section 5 concerned linguistic behaviors, but the proposal is about natural conventions in general. Thus, it can be applied to the conventions of other (whether symbolic or non-symbolic) domains outside language.[18] And even within

[18] For instance, the conventions of musical notation. In the Western staff notation, notes are represented on a discrete, monotonic grid where higher spatial location corresponds to higher pitch, leveraging our propensity to group musical tones in discrete equivalence classes and represent relationships of relative height in spatial terms. On the staff, linear order corresponds to temporal order, and coincidence defeasibly corresponds to simultaneity, exploiting our tendency to cognitively encode succession in time as succession in space. These

language it could, modulo the due adjustments, assist the description of *non*-conventional and *pre*-conventional linguistic behaviors.

On-the-fly lexical innovations (Armstrong 2016; Gasparri 2022) are a case in point. Imagine you want to tell a colleague of yours, Sue, that John stole your mug from the coffee room. Suppose that you have chosen not to do that by using conventional vocabulary (the reason is irrelevant: you may just want to sound witty). Suppose, further, that another coworker of yours, Mary, is also an accomplished illusionist, and that Sue knows that. You connect the dots and tell Sue: "John pulled a Mary on my mug". In the described context, the expression 'pull a Mary' is an inherently, and thus probabilistically natural solution to the problem of telling Sue that John stole your mug under the constraint of non-conventionality. It is inherently natural because it taps into Sue's propensity to associate Mary with the act of making objects disappear. By being inherently natural, the innovation is also probabilistically natural: in the described situation, it is far more likely to arise than alternative nonce phrases that Sue would have no way of interpreting correctly. E.g., "pull an *x*" with *x* being some obscure magician known only to you.

In summary, the framework holds potential for interesting applications to semantic change, including long-range semantic drifts, the bottom-up emergence of patterns of sociolinguistic variation in the use of a word, and top-down linguistic interventions (i.e., deliberate proposals to reform the conventional meaning of a term). In conceptual engineering and conceptual ethics, there is debate about the conditions that linguistic interventions have to meet to have a chance at generating genuine meaning change (see, e.g., Burgess, Cappelen & Plunkett 2020). One could reflect on a principle of Maximize Inherent Naturalness as a non-exclusive, defeasible requirement for linguistic interventions to have good chances of conventional dissemination.

and similar factors may warrant an argument that the staff notation is a conventional system with unexpected levels of inherent and probabilistic naturalness. For more on musical notation in the West, see Grier (2021).

References

Armstrong, J. (2016). The problem of lexical innovation. *Linguistics and Philosophy*, 39, 87–118.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *PNAS*, *113*, 10818–10823.

Bordwell, D. (2007). *Poetics of cinema*. New York: Routledge.

Burgess, Alexis, Cappelen, H., & Plunkett, D. (eds.) 2020. *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Burling, R. (2005). *The talking ape: How language evolved*. Oxford: Oxford University Press.

Buskell, A. (2017). What are cultural attractors? *Biology & Philosophy*, 32, 377–394.

Chowdhury, S. M., Kovenock, D., Rojo Arjona, D., & Wilcox, N. T. (2021). Focality and asymmetry in multi-battle contests. *The Economic Journal*, *131*, 1593–1619.

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

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Cochran, C., & O'Connor, C. (2019). Inequality and inequity in the emergence of conventions. *Politics, Philosophy & Economics*, *18*, 264–281.

Cumming, S., Greenberg, G., Kaiser, E., & Kelly, R. (2021). Showing seeing in film. *Ergo*, *7*, 730–756.

Cumming, S., Greenberg, G., & Kelly, R. (2017). Conventions of viewpoint coherence in film. *Philosophers' Imprint*, *17*, 1–28.

Dahl, A., & Waltzer, T. (2020). Constraints on conventions: Resolving two puzzles of conventionality. *Cognition*, *196*, 104152.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*, 2149–2169.

de Saussure, F. (1916). *Cours de Linguistique Générale*. Lausanne-Paris: Payot.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*, 603–615.

Eco, U. (1976). *A Theory of Semiotics*. Bloomington: Indiana University Press.

Gasparri, L. (2022). Lexical innovation and the periphery of language. *Linguistics and Philosophy*, *45*, 39–63.

Gasparri, L., Filippi, P., Wild, M., & Glock, H.-J. (2023). Notions of arbitrariness. *Mind & Language*, *38*, 1120-1137.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Grier, J. (2021). *Musical Notation in the West*. Cambridge: Cambridge University Press.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22.

Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context* (pp. 11–42). Washington: Georgetown University Press.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130298.

Kempe, V., Gauvrit, N., Panayotov, N., Cunningham, S., & Tamariz, M. (2021). Amount of learning and signal stability modulate emergence of structure and iconicity in novel signaling systems. *Cognitive Science*, 45. https://doi.org/10.1111/cogs.13057

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*, 439–453.

Leland, J. W., & Schneider, M. (2018). A theory of focal points in 2 × 2 games. *Journal of Economic Psychology*, 65, 75–89.

Lewis, D. K. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Liddell, S. K. (2003). *Grammar, gesture, and meaning in American Sign Language*. Cambridge: Cambridge University Press.

Lockwood, G., Dingemanse, M., & Hagoort, P. (2016). Sound-symbolism boosts novel word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1274–1281.

Metz, C. (1974). *Film language: A semiotics of the cinema*. Chicago, IL: University of Chicago Press.

Millikan, R. G. (2005). *Language: A Biological Model*. Oxford: Oxford University Press.

Monaghan, P., & Roberts, S. G. (2021). Iconicity and diachronic language change. *Cognitive Science*, *45*. https://doi.org/10.1111/cogs.12968

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130299.

Morgan, J. L. (1978). Two types of convention in indirect speech acts. In P. Cole (Ed.), *Pragmatics* (pp. 261–280). Leiden: Brill.

Nielsen, A. K., & Dingemanse, M. (2021). Iconicity in Word Learning and Beyond: A Critical Review. *Language and Speech*, *64*, 52–72.

O'Connor, C. (2021). Measuring conventionality. *Australasian Journal of Philosophy*, *99*, 579–596.

Pateman, T. (1986). Transparent and translucent icons. *The British Journal of Aesthetics*, *26*, 380–382.

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130300.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*, 1112–1130.

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021). How (Non-)Optimal is the Lexicon? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4426–4438.

Planer, R. J., & Kalkman, D. (2021). Arbitrary signals and cognitive complexity. *The British Journal for the Philosophy of Science*, 72, 563–586.

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA: Blackwell.

Rita, M., Chaabouni, R., & Dupoux, E. (2020). "LazImpa": Lazy and Impatient neural agents learn to communicate efficiently. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 335–343. Association for Computational Linguistics.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sidhu, D. M., Westbury, C., Hollis, G., & Pexman, P. M. (2021). Sound symbolism shapes the English language: The *maluma/takete* effect in English nouns. *Psychonomic Bulletin & Review*, *28*, 1390–1398.

Simons, M., & Zollman, K. J. S. (2019). Natural conventions and indirect speech acts. *Philosophers' Imprint*, *19*, 1–26.

van Rooij, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy*, 27, 493–527.

Wichmann, S., Rama, T., & Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*. https://doi.org/10.1515/lity.2011.013

Winter, B., & Perlman, M. (2021). Size sound symbolism in the English lexicon. *Glossa: A Journal of General Linguistics*, 6, 79.

Zipf, G. K. (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Boston, MA: Houghton Mifflin.