Optimal Languages*

Francesc Dilmé†

Summer 2022

Abstract

This paper studies how languages are shaped by the cognitive costs that using them involves. We introduce a new continuous approach to characterize the optimal resolution of the tradeoff between the precision of a language and the complexity of the structures it uses, and its dependence on the information the language is used to describe. Notably, when the cost of communication is endogenized using information theory, all words in an optimal language are equally precise and their precision is independent of the distribution of states.

Keywords: Optimal language, communication, information theory

JEL Codes: D83, Z13

*I thank Andreas Blume, Wooyoung Lim, Mark Machina, Lars Metzger, Frank Riedel, Joel Sobel, Deszo Szalay, and the audiences at the seminars at UC San Diego, Hong Kong University of Science and Technology, National University of Singapore, Bar Ilan University, University of Bonn, and University of Bielefeld for their useful comments.

†University of Bonn. fdilme@uni-bonn.de

1 Introduction

Human communication is based on the use of natural, constructed, and formal languages. The main use of these languages is to transmit and store information, and they evolve (and are developed) over time, in part to accomplish these tasks more efficiently.[1] As a result, languages are shaped by the type of information they describe and the underlying costs of using them, which typically derive from the complexity of their substructures and the cognitive costs that learning and using them involves.

This paper studies optimal languages under the presumption that communication is costly due to its lack of precision, the complexity and length of its coding, and the costs associated with the physical or cognitive limitations that communicators face. We use insights from the economics and information theory literatures to develop a new general framework to study costly communication, and we uncover several features of optimal languages, and a number of comparative statics results. We use them to rationalize some empirical features observed in natural languages.

We model a language as a map from a set of (differently likely) states of the world to a set of heterogeneous words. The set of words is initially taken as given and later endogenized using information theory. In our base model, the communication loss that using a language generates comes from two sources. The first is the *imprecision loss*, which is due to the mistakes or misunderstandings that the lack of precision of the language generates. Lowering the imprecision loss favors increasing the size of the vocabulary (i.e., the set of used words) to lower the amount of states that each word signifies. The second is the (word-dependent) *speaking loss* of using each word, which may derive from its complexity or length.[2] Minimizing the speaking loss favors using a few simple words for communication. Our goal is to characterize how, depending on the likelihood of states, a(n optimal) language gives meaning to words to minimize the total loss that communication generates.

We begin our analysis by characterizing the optimal language for a fixed given set of words. In this setting, increasing the precision of a word lowers the imprecision loss of the

[1] Initial uses of the cost–benefit analysis in linguistics came from Zipf (1949) and Marschak (1965). See Rubinstein (2000) for an evolutionary model of optimal languages, and John (2016) for a recent survey on the theory and evidence of how economic forces influence language dynamics.

[2] Our concept of "word" corresponds to a "label" that can be assigned to describe some piece of information. In a common language, for example, it may correspond to one or more words (or sentences) that may be long and complex, increasing the frictions that communication involves. Section 4 pushes this interpretation further by assuming, in line with standard information theory, that such labels are constructed using combinations of the letters of some alphabet. In this case, using more labels makes them more precise, but implies using longer codes that increase the speaking loss.

states it describes and, because it lowers the likelihood of it being used, the expected speaking loss it generates also decreases. Still, making some words more precise implies lowering the precision and increasing the usage frequency of other words. Conversely, describing some states more precisely implies that other states are communicated more coarsely. We show that in an optimal language, similar words (in terms of their complexity) tend to be similarly precise and also to refer to similarly likely states. Furthermore, simpler words are used more frequently than complex words and also refer to more likely states of the world. We show that the precision of a word depends not only on its complexity but also on the degree of heterogeneity of both states of the world and words. If, for example, all words are similarly complex, an optimal language focusses on minimizing the imprecision loss, and more likely states are described precisely as a result. Alternatively, if the states of the world are similarly likely, more likely states are described less precisely. In this case, an optimal language aims at minimizing the speaking loss, and therefore complex words are used to precisely describe unlikely states.

Conditional on being realized, likely states are shown to generate a lower communication loss than unlikely states, even when they are communicated more coarsely. We use this to prove that the efficiency of optimal communication decreases when the homogeneity of the states of the world increases through a garbling of the state space. Indeed, if information is more concentrated, an optimal language can be focussed on communicating likely states very precisely. Similarly, the optimal communication loss also increases when the set of words becomes more homogenous.

In the second part of the paper, we fix an alphabet and endogenize the complexity structure of the word set using information theory. To do this, we separate the problem of finding optimal languages into two parts. First, for each fixed distribution of words' frequencies, we determine the speaking loss that communicating them with an optimal coding generates. Using the Source Coding Theorem, this loss is shown to be a linear function of the differential entropy of the distribution of word likelihoods. Next, we characterize the optimal language for a given distribution of states, that is, the assignation of states to words that minimizes the sum of the imprecision and the speaking losses. We find that in an optimal language, all words are equally precise and that their precision is independent of the distribution of states and increasing with the size of the alphabet. Consequently, their rank distribution of words coincides with that of the states of the world.

We extend our results to include two additional types of cognitive costs. First, we include a *capacity loss* related to the size of the vocabulary; the more words a vocabulary includes, the higher the cost of learning the language or recalling a word when needed. The second is a state-dependent *importance factor* indicating the loss from miscommunicating each given

state. These additions have little effect on our results for languages with an exogenous word set. When the word set is endogenous, words referring to likely states are more precise and shorter in the first case, while words referring to important states are more precise (but not necessarily shorter) in the second.

Relationship to the Literature. The economics literature on communication has mostly focussed on the study of (strategic) information transmission between (one or multiple) senders and receivers with misaligned incentives. Three notable exceptions, and the papers most similar to ours, are Cŕemer, Garicano and Prat (2007), Jäger, Metzger and Riedel (2011), and Sobel (2015), henceforth CGP, JMR, and S, respectively. The first paper studies optimal codes within firms, assuming that there is a finite set of words available for communication and that the receiver incurs a cost that depends on the number of states each word signifies. The second paper analyzes a multi-dimensional cheap-talk model and uses it to characterize the sets of states that words signify and to provide an algorithm to numerically calculate optimal languages for large numbers of words. The third paper considers a version of the Crawford and Sobel (1982) (henceforth CS) model without conflict of interest between the sender and the receiver and with a limited number words (or messages) available for communication. The key contribution of our paper to this literature and its main focus is the analysis of how cognitive and physical costs shape optimal communication.[3] To this end, we generalize some of the previously obtained results, while providing new insights on the tradeoffs that designing efficient communication imposes when words differ in their complexity. Importantly, we illustrate how information theoretical insights can be used to endogenize the word structure in economics models of communication, obtain new predictions, and open the door to further research on the intersection of the economics and information theory literatures, which have been regarded as disconnected from each other in the past.[4]

---

[3]A common assumption in the cheap-talk literature is that the cost of using each message is the same (normalized to zero). Exceptions include the papers on costly lies by Kartik et al. (2007) and Austen-Smith and Banks (2000), which study communication between agents with conflicting interests when there is a cost of misrepresenting information, and Austen-Smith and Banks (2000), which allows for money burning in addition to the cheap-talk messages. Another exception is Hertel and Smith (2013), which adds a cost of sending messages in the CS model and obtains that when the bias is zero, only messages with costs below a given threshold are used in the equilibrium, thus satisfying the no-incentive-to-separate (NITS) condition.

[4]Our model abstracts from the optimal "grammar", "syntax", or "compositionality" (that is, how the meaning of a complex expression can be deduced from the meaning of its parts see Rubinstein (1996) and Blume (2000, 2004)) and from intentional "vagueness" of the language (that is, the fact that the meaning of some words is not clearly defined see Lipman (2009) and Lim and Wu (2017)). It instead focusses on how to optimally give meaning to a set of heterogeneous words, which can be understood themselves as structures

Since Zipf (1949) first introduced the tradeoff between the preference of the speaker for a non-complex language and the preference of the hearer for precision, part of the linguistics literature has focussed on theoretically recovering the so-called Zipf's law for the rank distribution of words.[5] We model the imprecision loss using (limits of) standard economics models and focus our analysis on the structure of the language. In particular, we characterize how the properties of the information transmitted and the word complexity shape the language's usage and precision. The general setting introduced in the paper (that accommodates both exogenous and endogenous word complexity), as well as our findings, allows combining previous insights from both the economics and the linguistics literatures to shed light on the structure of optimal languages.

Technical Contribution. This paper proposes a continuous model of communication. In our model, rather unconventionally, a language uses a continuum of words to communicate a set with a continuum of states of the world. This modeling choice permits using measure-theoretic analysis to study languages in the presence of both heterogeneous information and word complexity. In particular, we show how measure-preserving transformations of the state space and the word space can be used to transform the problem of finding an optimal language into finding the monotone one-dimensional function that minimizes the communication loss functional, which can be solved using calculus of variations. The optimal assignation of heterogeneous states of the world to heterogeneous words is then characterized by a second-order differential equation that provides us with explicit solutions for some primitives of the model.

Our continuous approach differs from the previous literature where, in the absence of strategic considerations, the set of words is typically assumed to be finite. As argued in Appendix B, our model can be obtained as the limit of some of the previous discrete models when the number of available words increases (or their usage cost decreases). We show this explicitly for the S model, and we argue that this is also true for CGP and JMR. This makes our model suitable for studying natural or constructed languages that use many different words, and therefore their efficient usage is likely to be affected by the word complexity or length and other cognitive costs. We show that the assumption that a large number of words used for communication (infinity in the limit) gives tractability, permits a complete characterization of optimal languages, and consequently sheds light on how they are shaped

(like sentences) constructed using grammar rules.

[5]See Ferrer-i-Cancho (2017) for a recent review of the literature. For example, Ferrer-i-Cancho and Solé (2003) numerically obtains that the Zipf's law is achieved when the objective function adequately balances the two preferences, and states are uniformly distributed.

by their usage and cost structure.[6]

Structure of the Paper. Section 2 sets the basic model, and Section 3 characterizes optimal languages for a fixed set of words. In Section 4, we endogenize the set of words (and its cost structure) using information theory. Section 5 relaxes some simplifying assumptions used in Section 3. Section 6 concludes. The proofs of the results in the main text are in Appendix A. In Appendix B, we show that our model is a limit of a standard communication model without conflict of interest as the cost of the words decreases.

2 A Continuous Model of Languages

In this section we introduce a continuous model of languages. To make the presentation clear, we present a simplified version of it, and we generalize the setting in Section 5. In particular, we here assume (and latter relax) that the state space is one-dimensional, and states are ordered in terms of their likelihood. Appendix B shows that particular versions of our model can be obtained as the limit of standard communication models with a finite vocabulary (and with general payoff functions) as the size of the vocabulary increases.

There is a *set of states (of the world)* $T \equiv [0,1]$. The likelihood with which states are realized is modeled using an absolutely continuous, full-support probability measure F in T. We use f to denote its density function so, for each state $t \in T$, $f(t)$ is interpreted as the frequency (or likelihood) with which t occurs. In this section, for convenience and without loss of optimality (see Remark 3.3), we assume that states are ordered in decreasing likelihood, that is, f is decreasing, and that f is differentiable. Letting μ be Lebesgue measure in $\mathbb{R}$, $\mu(T')$ denotes the amount of states in a set $T' \subset T$, and $F(T')$ their total likelihood.

There is a *set of words* $W \equiv \mathbb{R}_+$ to be used for communication. The *usage cost* of each word is captured with the function $c : W \to \mathbb{R}$, which associates to each word w the cost of using it (coming from, for example, its complexity or length). We assume that c is differentiable, increasing and such that $\lim_{w\to\infty} c(w) = \infty$.

A *language* (or *semantics*) is a measurable function $\ell : T \to W$ satisfying that the pushforward measure of μ, denoted μ_ℓ, is absolutely continuous.[7] Notice that, the density

[6]See Dilmé (2018) for an analysis of the CS model with a small bias between the sender and the receiver. As in our model, a small bias implies that the number of words used (in efficient equilibria) is also large, adding tractability to the analysis.

[7]Throughout the paper we use κ_f to denote the pushforward measure of a measure κ in $\mathbb{R}^n$ under a function $h : \mathbb{R}^n \to \mathbb{R}$, which satisfies $\kappa_h(B) = \kappa(f^{-1}(B))$ for all measurable sets B in $\mathbb{R}$. If κ and κ_h are absolutely

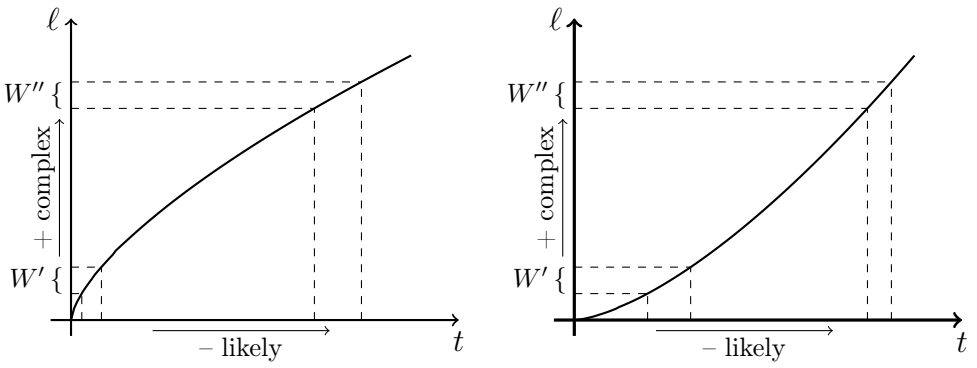

Figure 1: Two examples of a language. In both cases, simple words refer to likely states. In (a) simple words are more precise than complex words, and the contrary is true in (b). See the relative sizes anti-images of the sets W' and W'', which are of the same size.[9] Note that, when ℓ is differentiable, the precision with which a state is communicated is $1_\ell(\ell(t))^{-1} = \ell'(t)$.

of μ_ℓ at a given word $w \in W$, denoted $1_\ell(w)$, can be interpreted as the local ratio of states of the world per word signifies at w or, put differently, $1_\ell(w)^{-1}$ can be interpreted as the *precision* of w (see Figure 1). Therefore, our definition of a language ensures that there is no word which is infinitely imprecise.[8] Intuitively, for a given language ℓ and a set of words $W' \subset W$, the amount of states communicated using words in W', $\mu(\ell^{-1}(W'))$, is equal to $\int_{W'} 1_\ell(w)\,\mathrm{d}w$, so no positive mass of states is assigned to a single word.

In this section, the *communication loss* that a language ℓ generates is given by

$$\mathcal{L}(\ell) \equiv \underbrace{\int_T g(1_\ell(\ell(t)))\, f(t)\,\mathrm{d}t}_{\equiv \mathcal{L}^i(\ell)} + \underbrace{\int_T c(\ell(t))\, f(t)\,\mathrm{d}t}_{\equiv \mathcal{L}^s(\ell)}\,, \tag{2.1}$$

where $g : \mathbb{R}_+ \to \mathbb{R}$ is a twice-differentiable, increasing and convex function such that $\lim_{x\to\infty} g(x) = \infty$. The communication loss is divided between the *imprecision loss* $\mathcal{L}^i$ and

continuous, and k denotes the density of κ, we use k_h to denote the density of κ_h. In particular, since the density function of μ is 1, 1_ℓ denotes the density of μ_ℓ.

[8]The assumption that μ_ℓ is absolutely continuous is convenient and without loss of optimality given our loss function (2.1). Indeed, it is easy to see that for any sequence of languages $(\ell_i)_i$ such that $(\mu_{\ell_i})_i$ converges (in measure) to some measure which is not absolutely continuous, the limit of the corresponding sequence of communication losses given by equation (2.1) tends to $+\infty$. Appendix B shows that optimal languages in discrete models converge, as the communication los shrinks, to languages satisfying our restriction.

[9]Note that in both subfigures, the words in W' are simpler than the words in W''. In panel (a), words in W' jointly signify a set of smaller size than the states signified by the words in W'', so the words in W' are more precise. The reverse is true in (b).

the *speaking loss* $\mathcal{L}^s$. The imprecision loss is owed to the lack of precision of the language. It is the sum across the state space of the product between the likelihood of each state and loss derived from the coarseness with which such a state is communicated (which corresponds to the ratio of states per word, 1_ℓ, see Remark 2.1 below), derived from the potential mistakes that imprecise information generates. The speaking loss averages the usage cost of all used words (related to their complexity), which is the sum across the state space of product of the likelihood of each state and the cost of using its associated word. Section 5.2 analyzes the effect of adding a capacity loss to the right hand side of equation (2.1), that is, a loss term that depends on the amount of words used for communication, $\mu(\ell(T))$, arising from the cost of learning the language. Given that all results in Section 3 apply also in the presence of a capacity loss, but its inclusion makes some of the arguments and expressions unnecessarily tedious, we assume in our base model that there is no capacity loss.

The specification of the communication loss in (2.1) captures a fundamental tradeoff that designing an optimal language involves. On the one hand, lowering the communication's inaccuracy (lowering the imprecision loss) favors using many words in order to increase their precision. On the other hand, making the language simpler (lowering the speaking loss) favors using only (few) "simple" (and cheap) words.

It is sometimes convenient to write the communication loss in the space of words in the following way:

$$\mathcal{L}(\ell) = \int_W \big(g(1_\ell(w)) + c(w)\big)\, f_\ell(w)\, \mathrm{d}w \tag{2.2}$$

where, for each word $w \in W$, the pushforward measure of F under ℓ, $f_\ell(w)$ is the frequency with which w is used.[10] The communication loss associated to a word $w \in W$ is given by the likelihood of it being used (given by $f_\ell(w)$) multiplied by its cost conditional on being used, which equals its imprecision loss $g(1_\ell(w))$ plus its usage cost $c(w)$.

Remark 2.1. Appendix B shows that equation (2.1) is obtained as the limit of standard communication models with a finite word set and a (possibly infinite) bigger set of states. To gather some additional intuition, consider a version of the CGP model with a continuum of states. In this model, the imprecision loss that a word generates when it is used depends on the amount of states it signifies, which is interpreted as the cost that the receiver incurs to distinguish the actual realized state. In order to increase the number of words while keeping their distribution it is convenient to use, for each $\Delta > 0$, the set $\hat{W}^\Delta \equiv \big\{[(k\text{–}1)\,\Delta, k\Delta) \,\big|\, k \in \mathbb{N}\big\}$

[10] It is important to realize that, even if ℓ is injective, $f_\ell(\ell(t))$ is different, in general, from $f(t)$. Indeed, if ℓ is injective, $f_\ell(w) = f(\ell^{-1}(w))\, 1_\ell(w)$, that is, the frequency with which a (set of) word(s) is used depends both on the likelihood of the states it refers to and the corresponding ratio of states of the world per word.

to denote the set of words. A language ℓ (as defined in our model) is interpreted as assigning to each word $\hat{w} \in \hat{W}^{\Delta}$ the set of states of the world $\ell^{-1}(\hat{w})$. For each language ℓ, it is then the case that

$$\lim_{\Delta \searrow 0} \sum_{\hat{w} \in \hat{W}^{\Delta}} g\left(\frac{\mu(\ell^{-1}(\hat{w}))}{\Delta}\right) F(\ell^{-1}(\hat{w})) = \int_T g(1_\ell(w))\, f_\ell(w)\, \mathrm{d}w \ .$$

Consequently, even when in our limit specification the language is injective (as it will be the case for optimal languages), the local density of states per word shall be interpreted as (the limit of) the "amount of states" a word signifies.

Similarly, in the S model, the receiver takes an action (denoted $\alpha(\hat{w})$) after the sender communicates word $\hat{w}$. Hence, in an optimal language (see the details in Appendix B.1), each word signifies an interval $[\inf(\ell^{-1}(\hat{w})), \sup(\ell^{-1}(\hat{w}))] \subset T$, and the action $\alpha(\hat{w})$ minimizes the average square-distance of the states in the set. In this case, for a quadratic payoff loss and a fixed language ℓ, the imprecision loss that the language generates is

$$\lim_{\Delta \searrow 0} \sum_{\hat{w} \in \hat{W}^{\Delta}} \frac{1}{\Delta^2} \int_{\ell^{-1}(\hat{w})} (\alpha(\hat{w}) - t)^2\, f(t)\, \mathrm{d}t = \frac{1}{12} \int_0^1 1_\ell(\ell(t))^2\, f(t)\, \mathrm{d}t \ .$$

An alternative way of interpreting equation (2.1) is considering a multi-dimensional state space. Indeed, even though in our base model we assume (for simplicity) that the state space T is one dimensional, Section 5 shows that the same analysis and results also apply when the state space is multi-dimensional. Hence, if for example T is $[0,1]^2$, each word w used in an optimal language can be interpreted as meaning a one-dimensional set (such as a segment of a curve), so in the normalized (one-dimensional) model $1_\ell(w)$ can be interpreted as the length of this segment.

Remark 2.2. Similarly to the CGP model, our specification of the payoff does not use any structure of the state space other than its likelihood distribution. This contrasts with some communication models, like JMR and S, where close states are typically signified by the same word, since closeness of the meaning of a word plays an important role on determining the imprecision loss it generates. Appendix B argues that, in these models, as the number of words available for communication increases, the precision with which a state is communicated in an optimal language does not depend on its "spatial" allocation within the state space, but only on its local characteristics (such as its likelihood or, in Section 5.3, its relative importance). As a result, the implied payoff loss approaches the right hand side of equation (2.1). Intuitively, as words become more precise, their meanings partition the state space in very small pieces. The possibility of re-assigning words across the state space implies that the properties of a word in an optimal language (such as its precision, its usage frequency, etc) depend only on the local properties of the states it describes.

3 Optimal Languages

This section is devoted to characterizing optimal languages. Their existence and main characterization is established in Proposition 3.1. In order to save notation, in Lemmas 3.1 and 3.2, we use ℓ to denote an optimal language, and these lemmas shall be interpreted as necessary conditions that an optimal language must satisfy.

3.1 Preliminary Results

We say that $x : W \to W$ is a *state-precision-preserving re-assignation* of W if it preserves its Lebesgue measure μ.[11] The composition of a state-precision-preserving re-assignation x with a language ℓ, $x \circ \ell$, shall be interpreted as a "re-labeling" of the language ℓ, that is, a new language obtained by assigning that the meaning of each word to another word. As a result, composing a(ny) language ℓ with a state-precision-preserving re-assignation x (which generates a new language $x \circ \ell$) preserves the precision with which each state is communicated, but not the precision of each word. More formally, we have

$$1_{x\circ\ell}(x(\ell(t))) = 1_\ell(\ell(t)) \quad \text{but} \quad 1_{x\circ\ell}(w) \lesseqgtr 1_\ell(w) \quad \forall t \in T, \forall w \in W,$$

where the symbol $\lesseqgtr$ means that the inequality (or equality) may hold in any direction. Hence, after applying a state-precision-preserving re-assignation x to a language ℓ, the expression for its communication loss (2.2) becomes

$$\mathcal{L}(x \circ \ell) = \int_T \big(g(1_\ell(\ell(t))) + c(x(\ell(t)))\big)\, f(t)\, \mathrm{d}t \ .$$

Notice that applying a state-precision-preserving re-assignation to a language leaves its imprecision loss the same (since the precision with which each state is communicated remains the same), while potentially changes the speaking loss (since it changes the frequency with which each word is used).

Our first lemma claims that the frequency with which words are used is decreasing in their complexity, that is, simpler words are used more frequently.[12] The intuition is clear: if a (set of) word(s) is simpler (and cheaper) than another one (with the same amount of words) and also used less frequently, one can switch their assigned states while keeping the

[11] In general, a function $\phi : \mathbb{R} \to \mathbb{R}$ (or any other measurable space) is a *measure-preserving* transformation (preserving μ) if $\mu(\phi^{-1}(B)) = \mu(B)$ for any measurable set B.

[12] Our result is consistent with the assertion in Zipf (1935), empirically verified in posterior studies, that "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences".

precision with which each state is communicated the same (so keeping the imprecision loss), and therefore lowering the implied communication loss. Even though its proof is illustrative of the use measure-theoretical techniques to analyze languages, we leave it (and the proofs of the other results) to Appendix A.

Lemma 3.1. *$c(w) < c(w') \Rightarrow f_\ell(w) \geq f_\ell(w')$ for almost all $w, w' \in W$.*

We can use similar logic to characterize which states are allocated to which words. Now, for a fixed language ℓ, a *word-precision-preserving re-assignation* is a function $x_\ell : W \to W$ which preserves μ_ℓ (and therefore it preserves the precision of each word):

$$1_{x_\ell \circ \ell}(x_\ell(\ell(t))) \lneqq 1_\ell(\ell(t)) \quad \text{but} \quad 1_{x_\ell \circ \ell}(w) = 1_\ell(w) \quad \forall t \in T, w \in W. \tag{3.1}$$

Using the expression of the payoff in terms of the states (2.2), we see that re-allocating (using a word-precision-preserving re-assignation) highly likely states to words with a low usage cost (weakly) lowers the communication loss.

Lemma 3.2. *$c(\ell(t)) < c(\ell(t')) \Rightarrow f(t) \geq f(t')$ a.e.*

The logic behind Lemma 3.2, as the example below shows, comes from the convexity of our cost function g, and the fact that by Lemma 3.1 simple words are used more frequently. It relies on the fact that re-assignations of states that keep the frequency of each word the same but assign likely states to simple words make words more similarly precise, therefore reducing the imprecision loss without changing the speaking loss.

To gain some intuition on Lemma 3.2, consider a finite version of our model with only 6 states, $\{t_i\}_{i=1}^6$, with likelihoods $f(t_1) = f(t_2) = \frac{1}{4}$ and $f(t_3) = ... = f(t_6) = \frac{1}{8}$. Assume also there are 3 words available, $\{w_i\}_{i=1}^3$, $c(w_1) < c(w_2) = c(w_3)$. Consider the following two languages. The language ℓ_1 assigns $\{t_1, t_2\}$ to w_1, $\{t_3, t_4\}$ to w_2 and $\{t_5, t_6\}$ to w_3. The language ℓ_2, instead, assigns t_1 to w_2, t_2 to w_3 and $\{t_i\}_{i=3}^6$ to w_1. Since the frequency of each word is the same under both languages, these languages generate the same (discrete-analogous) speaking loss, but due to the convexity of g, they differ in their imprecision loss. Indeed, we have that, since g is convex and strictly increasingly,

$$\underbrace{\tfrac{1}{2}\, g(2) + \tfrac{1}{4}\, 2\, g(2)}_{\mathcal{L}^i(\ell_1)} = g(2) < g(\tfrac{5}{2}) \leq \underbrace{\tfrac{1}{2}\, g(4) + \tfrac{1}{4}\, 2\, g(1)}_{\mathcal{L}^i(\ell_2)} .$$

Our continuous approach allows us to generalize this intuition to any distribution of states and structure of costs of the words' set.

Remark 3.1. Lemmas 3.1 and 3.2 resemble Propositions 1 and 2 in CGP, which state that in their model "broader words describe less frequent events" and "are used less frequently",

while in ours these properties apply to more complex words. Interestingly, the logic for the analogous results is switched. For example, the reasoning for their result on the frequency of usage of broader words is analogous to the logic of our result on the likelihood of states signified by complex words: keeping the likelihood of words, assigning likely states to likely words smoothes the distribution of words' precision. Alternatively, the logic behind their result about the likelihood of states being described by broader words comes from the possibility of switching states between words, while we argued that complex words are less frequent by switching the meaning of words. As we will see (Proposition 3.1 below), their results do not apply to our model: depending on the distribution of state likelihoods and word complexity, broader words may be more or less frequent than precise words, and refer to more or less likely states.

3.2 Characterization of Optimal Languages

Lemma 3.2 ensures that, in an optimal language, ℓ can be chosen to be continuous and strictly increasing, and therefore differentiable almost everywhere. Then, a language ℓ gives the following communication loss:

$$\mathcal{L}(\ell) = \int_0^1 \left(g(\ell'(t)^{-1}) + c(\ell(t)) \right) f(t) \, \mathrm{d}t \ . \tag{3.2}$$

Notice that $\ell'(t)$ is equal to $1_\ell(\ell(t))^{-1}$ almost everywhere, since the precision with which state t is communicated is equal to the inverse of its local ratio of states per word. Hence, equation (3.2) illustrates of the main tradeoff of our model: increasing the precision of the language ℓ' lowers the imprecision loss, but requires using more complex words, as it increases $\ell(1)$. The problem of finding an optimal language can then be written as a problem of calculus of variations: it consists on finding the (increasing) map from the state space into the word space which minimizes the functional (3.2). This allows us to find the following characterization of any optimal language:

Proposition 3.1. *An optimal language ℓ exists, is twice-differentiable, and satisfies*

$$h''(\ell'(t)) \, \ell''(t) = -\frac{f'(t) \, h'(\ell'(t))}{f(t)} + c'(\ell(t)) \quad \textit{for all } t \in T \ , \tag{3.3}$$

where $h(x) \equiv g(1/x)$ for all $x \in \mathbb{R}_{++}$, which is decreasing and convex.

Equation (3.3) shows how the characteristics of the state space and the set of words affect optimal communication. The first term on its right hand side is negative and related to the imprecision loss. If the states of the world are heterogeneous (i.e., the slope of the

likelihood function f' is large in absolute value), the need for a precise language decreases fast with the state. If this term dominates the second one, likely states are communicated precisely, while unlikely states are communicated in a coarse manner. Intuitively, likely states tend to contribute more to the communication loss, as they are communicated more frequently. Such contribution is mitigated, in an optimal language, through communicating them more precisely, which implies $\ell''(\cdot) < 0$. Figure 1(a) depicts a language where this happens at all states. The second term on the right hand side of equation (3.3), instead, is positive and related to the speaking loss. If the words' usage cost c increases fast (so the second term dominates), an optimal language communicates less likely states more precisely. Equivalently, as we can see in Example 3.1 below, complex words are optimally made more precise in order to avoid using them frequently. This is corresponds to Figure 1(b).

Remark 3.2. Recent developments in linguistics have indicated that a word length is not only negatively related with its frequency (see footnote 12), but also with its "information content" (see, for example, Piantadosi et al. (2011)). Even though defining and measuring a word's information content is not without difficulty (and some degree of ambiguity), its natural analogous in our model is the "broadness" (i.e. inverse of the precision) of a word, that is, the amount of states it describes. Thus, such a result seems to indicate that the second term on the right hand side of equation (3.3) dominates the first, that is, word heterogeneity plays an important role on determining the structure of a language.

3.3 Comparative Statics: Heterogeneity of States and Words

We devote this section to shedding light on how changes in the heterogeneity of the states of the world and the complexity structure of the set of words affect the communication loss. Our first result generalizes the findings in CGP and S that more "complex" environments (i.e., environments with a less concentrated density function) generate a higher speaking costs. We generalize their results to allowing for word heterogeneity and endogenizing the size of the vocabulary. We provide afterwards an additional result claiming that a similar result applies to words: if they become more homogenous, the communication loss increases.

Before stating our comparative statics results regarding the heterogeneity of the set of state of the worlds and the set of words, we establish the following corollary of Proposition 3.1:

Corollary 3.1. *If $f(t) > f(t')$ then $g(1_\ell(\ell(t))) + c(\ell(t)) \leq g(1_\ell(\ell(t'))) + c(\ell(t'))$, for almost all $t, t' \in T$.*

Corollary 3.1 shows that more likely states generate, conditional on being realized, a lower communication loss. Thus, an optimal language alleviates the communication loss that likely

states generate by assigning them to simple words and, sometimes, communicating them precisely. The intuition is similar to that of Lemma 3.1: if a state of the world is more likely and generates a higher communication loss than another state, they can be "switched" (using the appropriate measure-preserving transformation of the state space) so that the total communication loss decreases.

The next result claims that, for a fixed language, increasing the homogeneity of the state space increases the communication loss.

Proposition 3.2. *Let $\ell(\cdot; f)$ denote an optimal language for a distribution f, and let $\mathcal{L}(\ell(\cdot; f); f)$ be the corresponding communication loss. Let $x : T \to T$ be a transformation preserving μ such that f_x is decreasing. Then, $\mathcal{L}(\ell(\cdot; f_x); f_x) \geq \mathcal{L}(\ell(\cdot; f); f)$, with strict inequality when $f \neq f_x$.*

The interpretation of Proposition 3.2 is the following. A measure-preserving transformation of the state space can be viewed as garbling states of the world into new "mixed" states of the world. As a result, after such a transformation, the amount of states of the world is the same, but they are more homogeneous than the (old) states of the world before the transformation. Given that, by Corollary 3.1, the payoff loss that a state generates (conditional on being realized) is increasing and independent on the likelihood of the state, standard arguments for measure-preserving re-assignations apply, so[13]

$$\mathcal{L}(\ell(\cdot; f_x); f_x) \geq \mathcal{L}(\ell(\cdot; f_x); f) \geq \mathcal{L}(\ell(\cdot; f); f) \ .$$

Hence, Proposition 3.2 establishes that a higher concentration of the states' likelihood distribution makes the communication loss decrease. This result is intuitive: when the likelihood of the states of the world becomes more concentrated, an optimal language describes (very) likely states more precisely (so simple words are precise), while complex words describe less precisely (very) unlikely states, lowering the total communication loss.

A similar intuition applies to increasing the heterogeneity of the set of words: if words become more heterogeneous, simple words can be used to efficiently communicate likely states, while complex words are left to communicate unlikely states.

Proposition 3.3. *Let $\ell(\cdot; c)$ denote an optimal language for a cost function c, and let $\mathcal{L}(\ell(\cdot; c); c)$ be the corresponding communication loss. Let $x : W \to W$ be a transformation preserving μ such that $c \circ x$ is finite and increasing. Then, $\mathcal{L}(\ell(\cdot; c \circ x); c \circ x) \geq \mathcal{L}(\ell(\cdot; c); c)$, with strict inequality when $c \neq c \circ x$ in $\ell(T; c)$.*

[13] As the proof of Lemma 3.1 illustrates, the Hardy-Littlewood inequality for the integral of two monotone functions ensures any measure-preserving re-assignation of f increases the payoff loss.

Remark 3.3. The payoff loss that a language ℓ generates (see equation (2.1)) is well defined independently on whether f is decreasing or not. Furthermore, by Ryff (1970), we have that for any fixed (not necessarily decreasing) f, there is some strictly decreasing f^* and measure-preserving $x : T \to T$ such that $f = f^* \circ x$. It is immediate to prove that, if ℓ is an optimal language for f^*, then $\mathcal{L}(\ell^*; f^*) = \mathcal{L}(\ell^* \circ x; f) \leq \mathcal{L}(\ell; f)$ for any language ℓ. Consequently, the assumption that f is decreasing is, in fact, innocuous. Proposition 3.2 reinforces the result in Lemma 3.2: in optimal languages, the states that a given word signify have similar likelihoods. Furthermore, Remark B.2 illustrates why re-arrangements of the state space do not change the properties of optimal languages with large vocabularies in settings such as those in JMR and S.

3.4 Examples

Example 3.1 (*constant* f). As a first illustrative example, consider the case where the states of the world are homogeneous, so $f(t) = 1$ for all $t \in T$, and g is the identity. Let ℓ be an optimal language. In this case, the solution to equation (3.1) can be written as $1/\ell'(t) = (c(\ell(1)) - c(\ell(t)))/2$ (where we used Proposition 5.2 below to pin down $\ell(1)$). Since the precision of a word is $\ell'(t)$, we see that, as we argued before, complex words are also precise. So, the density of states of the world associated with a particular word is linearly decreasing in its cost. Thus, we can write:

$$\int_0^{\ell(1)} \tfrac{1}{2} \big(c(\ell(1)) - c(w)\big)\, \mathrm{d}w = 1 \ .$$

The corresponding communication loss is $\mathcal{L}(\ell) = \frac{1}{4} \int_0^{\ell(1)} (c(\ell(1))^2 - c(w)^2)\, \mathrm{d}w$. Hence, the communication loss that a word generates is decreasing in its complexity: the imprecision loss of simple words is high because the amount of states associated to them is large, and their low cost does not compensate such a large loss. Conversely, the contribution to the communication loss of a state t is $\frac{1}{2}\,(c(\ell(1)) + c(\ell(t)))$, which is increasing, as established in Corollary 3.1.

Example 3.2 (*power costs*). Consider now the case where $c(w) = \kappa\, w^\beta$ for some $\kappa, \beta > 0$ and all $w \in W$, while now allowing a general form for f, and again assuming that g is the identity. Assume that ℓ is an optimal language and consider another language $\ell_\gamma(t) \equiv \gamma\, \ell(t)$ for all $t \in [0, 1]$, for some $\gamma > 0$. Note that, for this language, the mass of used words is $\gamma\, \ell(1)$. From equation (3.2) it is clear that $\mathcal{L}^i(\ell_\gamma) = \gamma^{-1}\, \mathcal{L}^i(\ell)$ (since $\ell'_\gamma(t) = \gamma\, \ell'(t)$) and that, given the homogeneity of the cost function, $\mathcal{L}^s(\ell_\gamma) = \gamma^\beta\, \mathcal{L}^s(\ell)$. Minimizing $\mathcal{L}(\ell)$ with

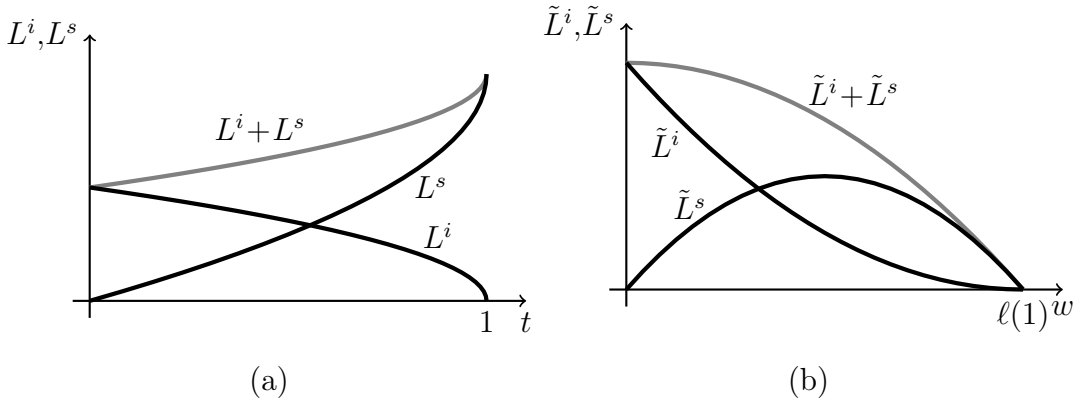

Figure 2: In (a) we depict the contribution of each state of the word t to the imprecision loss $L^i(t) \equiv g(\ell'(t)^{-1})\, f(t)$ and the speaking loss $L^s(t) \equiv c(\ell(t))\, f(t)$ when $g(x) = x$, $c(w) = w$ and $f \equiv 1$. In (b) we depict the contribution of each word $w \in W$ to the imprecision loss $\tilde{L}^i(w) \equiv L^i(\tau(w))\, f_\ell(w)$ and the speaking loss $\tilde{L}^s(w) \equiv L^s(\tau(w))\, f_\ell(w)$.

respect to γ we obtain that

$$\frac{\mathrm{d}\mathcal{L}(\ell_\gamma)}{\mathrm{d}\gamma} = -\gamma^{-2}\,\mathcal{L}^i(\ell) + \beta\,\gamma^{\beta-1}\,\mathcal{L}^s(\ell) \quad \Rightarrow \quad \gamma^{-\beta-1} = \frac{\beta\,\mathcal{L}^s(\ell)}{\mathcal{L}^i(\ell)}\ .$$

Since ℓ is optimal and ℓ_γ is feasible, we have that $\gamma = 1$ has to minimize $\mathcal{L}(\ell_\gamma)$. This implies that $\mathcal{L}^i(\ell) = \beta\,\mathcal{L}^s(\ell)$. As we see, even though a higher value of β makes complex words comparatively more costly, it also flattens the cost of simple words. Then, in an optimal language, the imprecision loss is higher than the speaking loss when $\beta > 1$. Conversely, if β is close to 0, the cost of each word $w \in W$ converges to κ. In this case, the amount of states per word is very small, and therefore the imprecision loss becomes very small too, while the speaking loss converges to κ.

Figure 2(a) depicts the imprecision and speaking losses as a function of the state, for linear costs (i.e., $\beta = 1$). As we just showed, in this case, in an optimal language, the imprecision loss is equal to the speaking loss, so the area below the curves L^i and L^s is the same. Thus, low states (which are communicated with simple words) do not contribute much to the speaking loss, but they contribute significantly to the imprecision loss; while the converse is true for high states. Overall, states associated with complex words have a higher contribution to the communication loss than the states associated with simple words (as established in Corollary 3.1). Conversely, Figure 2(b) shows the contribution of each word to the communication loss, which is obtained dividing the communication loss of the state(s) it means by its precision. Again, the areas below $\tilde{L}^i$ and $\tilde{L}^s$ coincide. Now, the contribution to the speaking loss is low for complex words: even though their cost is high, they are very precise, and therefore they are rarely used. As shown in Example 3.1, the total

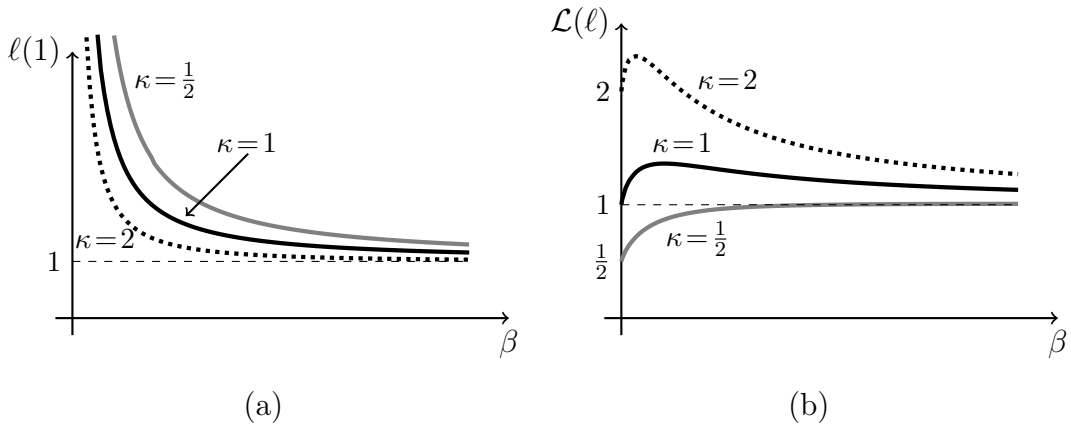

Figure 3: In (a) we depict the optimal vocabulary size $\ell(1)$ as a function of β when $c(w) = \kappa\, w^{\beta}$, $g(x) = x$ and $f \equiv 1$, for different values of κ; and in (b) we depict the corresponding communication losses. When κ is small, the cost of using each word is small too, so the size of the vocabulary is large and the communication loss is low. Similarly, when $\beta \to 0$, we have that $c(w) \to \kappa$ for all $w \in \mathbb{R}_+$, so $\ell(1) \to \infty$ (the precision of communication increases and the imprecision loss vanishes) and $\mathcal{L}(\ell) \to \kappa$. When, instead, $\beta \to \infty$, we have that $c(w) \to 0$ when $w \in [0, 1)$ and $c(w) \to \infty$ when $w > 1$. As a result, in this case, $\ell(1) \to 1$ (and all words have the same precision) and $\mathcal{L}(\ell) \to 1$ (the speaking loss vanishes).

contribution of a word to the communication loss is decreasing in its complexity.

Figure 3 depicts the effect that changes in κ have on the size of vocabulary and the communication loss depending on the value of β. To get some intuition, consider an increase of κ to $\kappa' > \kappa$. It is easy to verify, using equation (3.3), that the new optimal language is $(\kappa'/\kappa)^{-1/(\beta+1)}\, \ell(\cdot)$. So, an increase in the cost function in a factor κ'/κ lowers the mass of words used in a factor $(\kappa'/\kappa)^{-1/(\beta+1)} < 1$. This multiplies the total communication loss by a factor $(\kappa'/\kappa)^{1/(\beta+1)} < \kappa'/\kappa$. Hence, when κ increases, the size of the vocabulary decreases, and therefore the effect of such an increase on the communication loss is reduced.

4 Information Theory

In this section we endogenize the complexity structure of the set of words. To do this, we interpret W as a set of messages to which a language gives meaning, and we separate the problem of obtaining an optimal language as follows. In a first step, a language maps the set of states of the world into W, so as in our model $\ell : T \to W$. Now, additionally, for a fixed alphabet, a language specifies a codification of the set of messages into sequences of letters (or codes). We use Information Theory to model this second step, as it provides us with results on how to optimally codify a set of messages to minimize their speaking loss.

We solve the problem "backwards", so we begin obtaining the loss derived from the optimal coding of a set of messages for given distribution of words F_ℓ, and we then determine the optimal choice of F_ℓ.

4.1 Optimal Word Coding

Since the seminal paper of Shannon (1948), the Information Theory has been developed, in part, to study how to optimally encode information in order to efficiently communicate and store it. In this section we derive the limit cost of an optimal coding of a set of words with a given likelihood distribution as the number of words increases.

We first analyze the following (standard) coding problem in information theory, for a fixed increasing language ℓ such that $\bar{w} \equiv \ell(1) < +\infty$. We begin by discretizing the set of messages to be coded. So, for each Δ, the set $\hat{W}^\Delta$ (similarly to Remark 2.1) denotes the finite set of messages (some times called "symbols") $\left\{ [(k-1)\,\Delta, k\,\Delta) \,\middle|\, k=1, ..., \lfloor \bar{w}/\Delta \rfloor \right\}$, where we interpret that the language ℓ assigns to the message $\hat{w} \in \hat{W}^\Delta$ the states in $\ell^{-1}(\hat{w})$. Fix also an alphabet with n letters, $A \equiv \{1, ..., n\}$. A *coding* is a map from $\hat{W}^\Delta$ to the set of finite sequences of letters, $\hat{\gamma}^\Delta : \hat{W}^\Delta \to \cup_{m=1}^{\infty} A^m$ which is "uniquely decodable" (i.e., such that any sequence of letters is obtained by stacking at most one sequence of codes in $\hat{\gamma}^\Delta(\hat{W}^\Delta)$). For a fixed coding, let $m(\hat{w})$ be the number of letters used to codify message $\hat{w}$. Then, in the standard information theory model, the (speaking) loss of using a coding is the expected length of the code, $\sum_{\hat{w} \in \hat{W}^\Delta} f_\ell^\Delta(\hat{w})\, m(\hat{w})$, where $f_\ell^\Delta(\hat{w}) \equiv F(\ell^{-1}(\hat{w}))$ is the probability with which the message $\hat{w}$ is communicated.

The Source Coding Theorem establishes that the average number of letters of the messages used to code finitely-valued random variable is at least its entropy, and there are codes which/ approximate this limit. Furthermore, in an optimal code, the number of characters associated to each message is approximately equal to the negative of the logarithm of its frequency. As the number of messages increases, the entropy of the optimal coding diverges, but it can be normalized by an additive constant (equal to $\log_n(\Delta)$, so it is independent of the distribution of states) so that the limit of the normalized speaking loss of an optimal coding can be approximated by its differential entropy:[14]

$$\hat{H}(f_\ell) \equiv -\int_0^{\bar{w}} \log_n(f_\ell(w))\, f_\ell(w)\, \mathrm{d}w \ . \tag{4.1}$$

Thus, the previous formula gives us the speaking loss that an optimal coding that a language generates in our model with endogenous vocabulary. Using that $f_\ell(\ell(t)) = f(t)\, 1_\ell(\ell(t))$, it

[14]The approximation holds up to integer constraints that the finiteness of the alphabet imposes.

can written as follows

$$\begin{aligned}\hat{H}(f_\ell) &= -\int_T \left(\log_n(f(t)) + \log_n(1_\ell(\ell(t)))\right) f(t)\,\mathrm{d}t \\ &= \hat{H}(f) + \int_T -\log_n(1_\ell(\ell(t)))\, f(t)\,\mathrm{d}t\ . \end{aligned} \tag{4.2}$$

The first equality of the previous indicates that the length with which an optimal code communicates a state depends, additively, on a term that depends on its likelihood, and another that depends on how imprecisely it is communicated. Given that its likelihood is independent of the language, a language changes the length of an optimal coding only through affecting the precision.

The second equality of equation (4.2) shows that the speaking loss of optimally coding a language generates can divided in two terms. The first term is the differential entropy of the distribution of the states of the world. This term implies that complex (or more entropic) environments are more difficult to communicate. The second term is related to the precision of the language. This term implies that using a more precise language (that is, decreasing the value 1_ℓ) is costly, as it increases the number of messages and, as a result, the expected length of their optimal coding.

To gather more intuition from the previous expression consider the case of a binary code, that is, $n = 2$, and fix some language $\ell : T \to W$ and a small $\Delta > 0$. Now consider doubling the precision by using language $\ell^\dagger \equiv 2\,\ell$ instead of ℓ. In this case, notice that each message $\hat{w} \equiv [w - \Delta, w)$ signifies under ℓ the same states that the messages $\hat{w}_1^\dagger \equiv [2w - 2\Delta, 2w - \Delta)$ and $\hat{w}_2^\dagger \equiv [2w - \Delta, 2w)$ together signify under $\ell^\dagger$. Furthermore, if Δ is small enough, the likelihoods of both $\hat{w}_1^\dagger$ and $\hat{w}_2^\dagger$ under $\ell^\dagger$ are very close to the likelihood of $\hat{w}$ under ℓ divided by two. It is then easy to show that a near-optimal coding of $\ell^\dagger$ assigns to the messages $\hat{w}_1^\dagger$ and $\hat{w}_1^\dagger$ the same code that $\hat{w}$ is assigned in an optimal coding of ℓ with an additional final letter to differentiate them. As a result, doubling the precision of the language (notice that $1_{\ell^\dagger}(\ell^\dagger(\cdot)) = 1_\ell(\ell(\cdot))/2$) requires increasing the length of each of the codings of each message by one, and this is independent of the underlying state distribution (so the speaking loss in expression (4.2) increases by $-\log_2(2^{-1}) = 1$).

4.2 Optimal Languages

The previous subsection uses information theory to endogenize the speaking loss that a language generates. The speaking loss in equation (4.1) is such that, for each state $t \in T$, the speaking loss associated to the word $\ell(t)$ is $-\log_n(f(\ell(t)))$. So, differently from our model in Section 2, the usage cost of using the word endogenously depends on the frequency with which it is used.

Using the expression (4.1), we can write the problem of finding an optimal language as

$$\min_{\ell} \int_T \left(g(1_\ell(\ell(t))) - \log_n(1_\ell(\ell(t))) \right) f(t)\, \mathrm{d}t + \hat{H}(f) \ . \tag{4.3}$$

The following result characterizes optimal languages:

Proposition 4.1. *Any optimal language solving* (4.3) *satisfies* $1_\ell(\ell(t)) = \bar{w}^{-1}$ *for all* $t \in T$, *where* $\bar{w} \equiv \ell(1)$ *is the total mass of words and satisfies* $g'(\bar{w}^{-1}) = \bar{w}/\log(n)$.

A remarkable consequence of endogenizing the complexity structure of the word set is that, in an optimal language, all words are equally precise. This result, follows from the fact that, in an optimal coding, the length of the code (or usage cost) associated to a used word $\ell(t)$ can be approximated (up to a constant independent of the word) by the logarithm of its frequency $f_\ell(\ell(t))$ (see equation (4.1)). In our setting, the frequency of a word is $\ell'(t)^{-1} f(t)$, so its usage cost can be additively separated into a term that depends on the likelihood of the state it refers to, $\log(f(t))$, and a term that depends on its (im)precision $\log(\ell'(t)^{-1})$. Consequently, the minimization problem (4.3) can be solved by minimizing the interior of the integral "state-by-state". For example, if $n = 2$, doubling the value of ℓ' in a small set of states $[t, t+\varepsilon)$ (but keeping it in the rest of the states) only changes the communication loss that these states generate, and the additional communication loss is given by

$$\begin{aligned} &\left(g(\ell'(t)^{-1}/2) + \log_2(2\,\ell'(t)) - g(\ell'(t)^{-1}) - \log_2(\ell'(t)) \right) f(t)\,\varepsilon + O(\varepsilon^2) \\ &\quad = \left(g(\ell'(t)^{-1}/2) - g(\ell'(t)^{-1}) + 1 \right) f(t)\,\varepsilon + O(\varepsilon^2) \ . \end{aligned}$$

The intuition is similar to the one stated above for equation (4.2): the communication of a small set of states can be made more precise by increasing the amount of messages that describe them through adding letters to the existing messages. Since, in a prefix coding, one can expand the set of messages through adding additional characters to some of the existing codes (if $n = 2$, one can double the precision by adding one digit), this can be made without altering the rest of the coding. In our limit coding with an infinite number of messages, we can do this state by state.

The optimal size of the vocabulary given in Proposition 4.1 (which since g' is increasing exists and it is unique) is independent of the distribution of states. This is an immediate consequence on the fact that not only all words are equally precise, but also their precision does not depend on the distribution of states. Notice also that the size of the vocabulary is increasing in the size of the alphabet n. This is intuitive: the more letters are available, the less costly is to communicate a given distribution of words, so more emphasis can be put on minimizing the imprecision loss in an optimal language.

Remark 4.1. A somehow controversial matter in linguistics is the sometimes called "great Eskimo vocabulary hoax." It arose from the assertion in Boas (1911) that Eskimo languages have an unusually large number of words for snow. This claim was latter disputed by, among others, Martin (1986), so even though it remains an open topic, the current more spread view among experts is that the number of words for snow is similar across languages. We shed light on the debate by establishing a feature of efficient information transmission: in our model, the precision and/or the number of words of an optimal language are independent of the distribution of states of the world. Hence, even though snow-related events may be communicated more frequently in Eskimo societies, the optimal precision or amount of words used to efficiently communicate them need not to be different from other languages.

Remark 4.2. A visual inspection of equations (2.1) and (4.3) permits assigning the following usage cost function to the optimal coding of an optimal language:

$$c(w) = -\log_n \left(\frac{f(w/\ell(1))}{\ell(1)^{-1}} \right) \quad \forall w \in [0, \ell(1)] \ .$$

As it can be expected from the result in Proposition 4.1, this usage cost function exactly balances the distribution two forces which drive the precision of the words used in the language, given by the two terms on the right hand side of equation (3.3) (see the discussion after Proposition 3.1). It is easy to verify that it is also the unique that achieves it (up to an additive constant).

Zipf's Law

One of the best established empirical facts in the study of common (and some constructed) languages is that, when words are ranked by the frequency of their usage in a large text, their frequency is approximately proportional to the inverse of their rank, which is commonly known as the Zipf's law (Zipf (1935, 1949)). The two main approaches on understanding this regularity are the following. The first approach is the so-called "least effort" (see Zipf (1949)) where the language is assumed to minimize a convex combination of the communication losses incurred by the sender and the receiver. The other approach relies on the result in Li (1992), who shows that if symbols (including the "blank space") are generated independently with equal probabilities (also called "monkey languages"), one obtains approximations of the Zipf's law for the frequencies of words in a long text.[15]

In our model, words are equally precise in an optimal language, so the rank-distribution of words is the same as the rank distribution of states. Consequently, our model presents

[15]The fact that, in an optimal coding, all sequences of bits are approximately equally likely, reinforces this approach.

another force pushing the distribution of word frequencies towards a Zipf law: using the limit of standard communication models (in economics) without conflict of interest, the Zipf's law for words appears in an optimal language when the rank-frequency of described events follows a Zipf's distribution, which is often the case in natural and social environments.[16] (Notice that our distribution is a truncated Zipf's law since the size of the vocabulary is finite.)

5 Generalization of the Results

In this section we show that some of the simplifying assumptions in the model presented in Section 2 can be relaxed so the results in Section 3 still hold. To do this, we now consider a set of states (of the world) $T \subset \mathbb{R}^n$ for some $n \in \mathbb{N}$, assumed to be compact and with positive (Lebesgue) measure normalized to be 1. The likelihood of the states of the world is given by an absolutely continuous measure F with continuous density f. For convenience we keep $W = \mathbb{R}_+$ (otherwise it can be normalized similarly to the state space).

Notice that the definition of a language can be immediately generalized to a multi-dimensional state space, as well as the communication loss function in equation (2.1) (replacing "dt" by "$\mu(\mathrm{d}t)$") and the analogous expression (2.2), where with some abuse of notation μ now refers to the Lebesge measure both in $\mathbb{R}$ and $\mathbb{R}^n$. Lemmas 3.1 and 3.2 are still valid (their proofs do not rely on the fact that the state space is finite-dimensional). So, to prove that rest of the results (including Propositions 3.1-3.3 and Corollary 3.1) we show that, without loss of generality, we can normalize the state space to $\tilde{T} = [0, 1]$, and we can then derive the results in the normalized state space.

5.1 State Space Normalization

In this subsection, we show that we can normalize the general setting above by mapping the state space T into a normalized state space $\tilde{T} \equiv [0, 1]$. To do this, we first choose some mapping $\tilde{\tau} : T \to \tilde{T}$ such that the push-forward measure of the Lebesgue measure in $\mathbb{R}^n$ is the Lebesgue measure in $[0, 1]$, and such that if $f(t) > f(t')$ then $\tilde{\tau}(t) \leq \tilde{\tau}(t')$ for almost all

[16] A large range of phenomena in social and natural sciences which are rank-distributed according to the Zipf's law. These phenomena range from demography (Auerbach (1913)), biology (Willis (1922)), physics (Nicolis and Tsuda (1989)), etc. In economics, some attention have been put on the Zipf's distribution of cities' and firms' size (Gabaix (1999), Stanley et al. (1995), Axtell (2001)), webpage visits (Cunha et al. (1995), Huberman et al. (1998)) and academic citations (Silagadze (1997)). See Newman (2005) for a review.

$t, t' \in T$, so more likely states are placed at the lower part of $[0,1]$.[17]

We call a function $\tilde{\ell} : \tilde{T} \to W$ such that $\mu_{\tilde{\ell}}$ is absolutely continuous a *normalized language* (i.e., a language in the normalized state space $\tilde{T}$). Note that, for each normalized language $\tilde{\ell}$, we can construct a language by composing it with the normalizing function $\tilde{\tau}$, that is, $\mathcal{L}(\tilde{\ell}) \equiv \mathcal{L}(\tilde{\ell} \circ \tilde{\tau})$. We can then assign to the normalized language $\tilde{\ell}$ following the communication loss (which is analogous to equation (2.1)):

$$\mathcal{L}(\tilde{\ell}) = \int_0^1 \left(g(1_{\tilde{\ell}}(\tilde{\ell}(\tilde{t}))) + c(\tilde{\ell}(\tilde{t})) \right) \tilde{f}(\tilde{t}) \, \mathrm{d}\tilde{t} \, , \tag{5.1}$$

where $\tilde{f} \equiv f_{\tilde{\tau}}$ is the likelihood density of the normalized states of the world, which is independent of the particular choice of $\tilde{\tau}$ (so $\mathcal{L}(\tilde{\ell})$ is also independent of $\tilde{\tau}$). We can then use the analysis of Section 3 and obtain that a optimal normalized language exists with the same properties as in our base model. The following result ensures that focussing the attention on optimal normalized is without loss of generality.

Proposition 5.1. *A language is optimal if and only if it is the composition of an optimal normalized language (characterized in Proposition 3.1) with some function $\tilde{\tau}$ with the above properties.*

5.2 Optimal Languages with Capacity Costs

In our previous analysis, increasing the amount of words used in a language is costly because it involves using more complex words. Still, in practice, when the number of words used in a language is large, the cognitive/time/monetary cost learning languages or recalling words when used may be significant. Such a capacity loss is different from our speaking loss as it depends, in a first approximation, on the number of used words, and not the frequency with which they are used.

We now examine the effect of a capacity loss in our previous result. To this end, we consider the same setting as in Section 4.2, now adding a *capacity loss* term $C(\mu(\ell(T)))$ to the objective function of expression (2.1), where $C : \mathbb{R}_+ \to \mathbb{R}$ is a differentiable function increasing towards infinity. It is important to notice that, even though the speaking and capacity losses typically increase when additional words are included in the vocabulary, the additional capacity loss is independent of the usage of the extra word, while the extra speaking cost depends on its assigned precision and likelihood.

[17] Notice that this implies that $f_{\tilde{\tau}}$ is monotone and continuous, so it is differentiable almost everywhere. Furthermore, since , we have that for almost all $t \in T$, $f(t) = f_{\tilde{\tau}}(\tilde{\tau}(t))$.

All results in Section 3 remain the same in the presence of a capacity loss. The reason is that the arguments used to prove them involve (measure-preserving) reallocations of states of the world or words, which alter the imprecision and speaking losses without changing the total vocabulary size. The presence of a capacity loss allows the following characterization of the optimal vocabulary size:

Proposition 5.2. *The following result is true for an optimal language ℓ:*

$$\lim_{t\to 1}\left(g'(1_\ell(\ell(t))\,1_\ell(\ell(t))\,f_\ell(\ell(t))+C'(\ell(t))\right)=0\ . \tag{5.2}$$

To shed light on Proposition 5.2, assume first that the capacity loss is flat at the optimum (so $C'(\ell(1))=0$). In this case, either $f_\ell(\ell(1))=0$ or $\lim_{t\to 1}1_\ell(\ell(t))=0$, which also implies that $f_\ell(\ell(1))=0$. Hence, extreme words are very rare and, if $f(1)>0$, also very precise. The reason is that, if complex words were not very precise, increasing the precision of the least likely states by adding words in the vocabulary would not affect much the speaking loss (since these states are already communicated with similarly complex words), but would significantly lower their imprecision loss. Consequently, when there is no capacity loss, the expected speaking cost from using a word tends to be hump-shaped: simple words are not very costly, while complex words are used less often (see $\tilde{L}^s$ in Figure 2(b)). When, instead, $C'(w)>0$ for all $w\in W$ and $f(1)=0$, we have that complex words are very imprecise, since necessarily $\lim_{t\to 1}1_\ell(\ell(t))=\infty$, but they remain very unlikely. In this case, an optimal language communicates unlikely states very coarsely because introducing additional words into the language is costly, while its benefit is small.

A similar exercise can be performed by introducing a capacity loss in the payoff loss function for an endogenous word set (equation (4.3)), which leads to the following generalization of Proposition 4.1:

Proposition 5.3. *In the presence of a capacity loss, when the speaking loss is endogenized using information theory, for any optimal language ℓ there exists a constant $K>0$ such that, for almost all words $w\in\ell(T)$, we have $g'(1_\ell(w))\,1_\ell(w)-1/\log(n)=K\,f_\ell(w)^{-1}$.*

Proposition 5.3 establishes that, in an optimal language with capacity costs, simple words are more precise than complex words. Indeed, given that g is convex, $g'(1_\ell(w))\,1_\ell(w)$ is decreasing in the precision of a word, $1_\ell(w)^{-1}$. Similarly, $K\,f_\ell(w)^{-1}$ is decreasing in the word likelihood, so increasing in the word complexity. To see why notice that, from equation (4.3), we have that the effect on the total communication loss of changing the imprecision of a given (used) word $\ell(t)$ is proportional to $f(t)$. Thus, merging the meaning of two similar words (and therefore "saving" one word), and therefore duplicating the imprecision of the

resulting word, is proportional to $f(t)$.[18] Since by Lemma 3.2 complex words refer to less likely states than simple words, it is optimal to decrease their precision more than for the simple words.

5.3 State-Dependent Imprecision Loss

The specification of the communication loss in our base model (equation (2.1)) assumes that the communication loss of each state (conditional on being realized) depends only on the precision and complexity of the word used to communicate it. This is a common assumption in the literature on non-strategic communication, such as Zipf (1949) and Mandelbrot (1953), the Information Theory literature. It is also assumed or the CGP model (who consider a finite state space) and the JMR model.

Still, it is plausible that the importance of communicating some information precisely depends on the state of the world it communicates. Indeed, misunderstandings on information on health or safety information may generate higher payoff losses than misunderstandings on, for example, communication about entertainment-related issues. This is incorporated in the S model by specifying a payoff function analogous to the one used in the CS model (without bias between the sender and the receiver), that is, allowing the payoff loss that the imprecision of the language generates to be state dependent.

Adding to our model an extra layer of heterogeneity of the states of the world is straightforward. Appendix B shows that the limit of the S model (with a general loss function) as the number of words increases corresponds to our specification with an additional term multiplying the imprecision loss that each state generates. So, the communication loss (2.1) becomes

$$\mathcal{L}(\ell) \equiv \int_T \big(g\big(\nu(t)\, 1_\ell(\ell(t))\big) + c(\ell(t))\big)\, f(t)\, \mathrm{d}t \ , \tag{5.3}$$

where the *importance factor* $\nu(t)$ is strictly positive for all $t \in T$, and corresponds to the curvature of the payoff function at t in the CS model (see Appendix B). Intuitively, if the curvature of the payoff function is high in a given region of the state space, it is relatively "important" that they it is communicated precisely (compared with other regions), because when the states in this region are communicated coarsely the implied mistakes owed to the imprecision generate a higher communication loss.

[18] Given that, in an optimal language without a capacity loss, the precision of all words is the same ($\bar{w}$ in Proposition 4.1), the payoff loss of merging the meaning of two words $f(t_1)$ and $f(t_2)$ increases from $(g(\bar{w}^{-1}) + \log_n(\bar{w}))\,(f(t_1) + f(t_2))$ to $(g(2\,\bar{w}^{-1}) + \log_n(\bar{w}/2))\,(f(t_1) + f(t_2))$. Hence, reducing the number of used words is done more efficiently by merging the meanings of words referring to unlikely states.

Proposition 5.4. *Define the following renormalization $r : T \to T$ of the state space:*

$$r(t) = \int_T \nu(t')^{-1} \, \mathrm{d}t' \Big/ \int_T \nu(t')^{-1} \, \mathrm{d}t' \ .$$

Then, for each language ℓ, defining $\tilde{\ell} \equiv \ell \circ r^{-1}$ and $\tilde{f}(t) \equiv f(t)/\nu(t)$, we have that equation (5.3) can be rewritten as follows

$$\mathcal{L}(\ell) = \mathcal{L}(\tilde{\ell} \circ r) = \int_T \left(g\big(1_{\tilde{\ell}}(\tilde{\ell}(t))\big) + c(\tilde{\ell}(t)) \right) \tilde{f}(t) \, \mathrm{d}t \ .$$

Proposition 5.4 establishes that the results in Section 3 hold in the presence of a state-dependent importance factor when the state space is conveniently normalized. Heuristically, in a discrete setting, if a word w describes n states with likelihood f and importance factor ν, this word generates the same communication loss when, instead, it describes $n\,\nu$ states with a importance factor 1 and with likelihood f/ν.

The heterogeneity on the importance of the states of the world can also be incorporated our model with an endogenous word set in Section 4. In this case, as the following result establishes, Proposition 4.1 is generalizes to state that words are used to describe more important states of the world are more precise.

Proposition 5.5. *In the presence of a state-dependent imprecision loss, when the speaking loss is endogenized using information theory, an optimal language ℓ satisfies, for almost all $t \in T$, $\nu(t)\, 1_\ell(\ell(t))\, g'\big(\nu(t)\, 1_\ell(\ell(t))\big) = 1/\log(n)$.*

6 Conclusions

This paper develops a new framework to study how efficient information transmission resolves the tradeoff between the precision of the language and the costs related to its usage and size. We characterize the relationship between the precision of a word, its usage cost, and the kind of information it refers to, shedding light on how optimal languages are shaped by the cognitive costs and the characteristics of the information they are used to communicate or store. The framework introduced in this paper and its results can be used to study the optimality of spoken or artificial (computer/music/formal) languages and to determine their underlying (usage/cognitive) cost structure. In particular, it may help to analyze changes in languages due to the frequency of communication, changes in the information they communicate, or the need for precise information transmission.

A Proofs of the results

A.1 Proofs of the results in Section 3

Proof of Lemma 3.1

Proof. Fix a language ℓ and define a decreasing rearrangement f_ℓ^* of f_ℓ as

$$f_\ell^*(w) = \int_0^\infty \mathbb{I}_{[0,|\{w':f_\ell(w')>y\}|]}(w)\,\mathrm{d}y \quad \text{for all } w \in W\ .$$

Then, by Ryff (1970), there is a measure preserving map $x : W \to W$ such that $f_\ell = f_\ell^* \circ x$, or equivalently, $f_{x\circ\ell} = f_\ell^*$.[19] The function x can be interpreted as redefining (or switching) the meaning of words: it keeps the precision with which each state is communicated, but it changes the states that each word signifies and, as a result, the likelihood of it being used. Furthermore, given that f_ℓ^* is decreasing, x redefines words in such a way that it increases the frequency with which simpler words are used. As a result, using the Hardy-Littlewood inequality for the integral of two monotone functions, we obtain that

$$\int_W c(w)\, f_{x\circ\ell}(w)\,\mathrm{d}w = \int_W c(w)\, f_\ell^*(w)\,\mathrm{d}w \leq \int_W c(w)\, f_\ell(w)\,\mathrm{d}w\ .$$

So, a rearrangement of the word space which makes the frequency with which words are used decreasing, weakly lowers the communication loss, while keeping the imprecision loss the same. When f_ℓ^* and f_ℓ do not coincide almost everywhere it is easy to prove, using the rearrangement inequality, that the inequality is strict. □

Proof of Lemma 3.2

Proof. Consider the following change of variables. Let $h : T \to T$ be such that $h(\hat{t}) \equiv \int_0^{\hat{t}} f(t)\,\mathrm{d}t$. Given that h is bijective, ℓ can be decomposed as $\ell = \hat{\ell} \circ h$, for some $\hat{\ell} : T \to W$. We can then write the imprecision loss as

$$\int_0^1 \Big(g\big(f(h^{-1}(\hat{t}))\,1_{\hat{\ell}}(\hat{\ell}(\hat{t}))\big) + c(\hat{\ell}(\hat{t}))\Big)\,\mathrm{d}\hat{t}\ .$$

Now, notice that we can generate additional languages using rearrangement x of $1_{\hat{\ell}}(\hat{\ell}(\cdot))$. Such a rearrangement, does not change the speaking loss, preserving $1_{\hat{\ell}}$ (as integrals of function do not change after rearrangements) but, by a similar argument as in the proof of Lemma 3.1, it decreases the imprecision loss.[20] □

[19] Indeed, for each $T' \subset T$, we have $\int_{x\circ\ell(T')} f_{x\circ\ell}\,\mathrm{d}t = \int_{T'} f\,\mathrm{d}t = \int_{\ell(T')} f_\ell\,\mathrm{d}w = \int_{\ell(T')} f_\ell^* \circ x\,\mathrm{d}w = \int_{x\circ\ell(T')} f_\ell^*\,\mathrm{d}w$.

[20] Now we have a convex function of the product of two functions (instead of a linear function). It is easy to see that the Hardy-Littlewood inequality (see, for example, Hardy et al. (1952) (Theorem 378)) extends to

Proof of Proposition 3.1

Proof. Fix ℓ with a finite mass of used words $\bar{w} \in \mathbb{R}_{++}$ (so we require $\ell(1) = \bar{w}$). It is useful to rewrite the equation (3.2) in terms of the inverse of ℓ, denoted $\hat{\tau} : [0, \bar{w}] \to [0, 1]$ (which is continuous, increasing and concave). This gives us the following expression:

$$\mathcal{L}(\hat{\tau}) = \int_0^{\bar{w}} \underbrace{\big(g(\hat{\tau}'(w)) + c(w)\big)\, \hat{\tau}'(w)\, f(\hat{\tau}(w))}_{\equiv L(t, \hat{\tau}(w), \hat{\tau}'(w))} \, \mathrm{d}w \ . \tag{A.1}$$

The function L in (A.1) is often called the *Lagrangian* of the problem. Notice that it is convex with respect to the third component, since $\frac{\partial^2}{\partial v^2} L(w, t, v) = 2\, g'(v) + v\, g''(v) > 0$.

The problem of maximizing $\mathcal{L}(\hat{\tau})$ imposing $\hat{\tau}(0) = 0$ and $\tau(\bar{w}) = 1$ is a version of the standard "fundamental problem" in the calculus of variations. The corresponding Euler-Lagrange condition for optimality is

$$\tau''(w) = -\frac{c'(w) + \frac{f'(\tau(w))}{f(\tau(w))}\, g'(\tau'(w))}{2\, g'(\tau'(w)) + \tau'(w)\, g''(\tau'(w))} \ .$$

Given that the previous equation is well behaved, Theorem 3 in Rockafellar (2001) establishes that an optimal solution of the optimization exists and satisfies the equation almost everywhere. Equation (3.3) is obtained by rewriting the previous equation in terms of h and ℓ. Finally, using the fact that the payoff loss of an optimal language depends continuously on $\bar{w}$, there exists some $\bar{w}^* \in \overline{\mathbb{R}}_{++}$ which maximizes it. □

Proof of Corollary 3.1

Proof. Fix an optimal language ℓ. Let $\tilde{\tau} : T \to T$ be a mapping which preserves the measure μ and which makes $g(1_\ell(\ell(\tilde{\tau}(t)))) + c(\ell(\tilde{\tau}(t)))$ an increasing function of t. Since $\tilde{\tau}$ preserves μ, we have that $1_\ell = 1_{\ell \circ \tilde{\tau}}$, and therefore we can use the same argument as in Lemma 3.1 to prove the result. □

this case. Indeed, notice that if $0 < a < a'$, $0 < b < b'$ and g is convex, then $g(a\,b) < \min\{g(a\,b'), g(a'\,b)\} \leq \max\{g(a\,b'), g(a'\,b)\} < g(a'\,b')$, then, since both $g(a\,b')$ and $g(a'\,b)$ are convex combinations of $g(a\,b)$ and $g(a'\,b')$, we have that the convexity of$g$ implies $g(a\,b) + g(a'\,b') < g(a\,b') + g(a'\,b)$.

Proof of Proposition 3.2

Proof. We define $L_\ell(t) \equiv g(1_\ell(\ell(t))) + c(\ell(t))$. Then, we have that

$$\begin{aligned}\mathcal{L}(\ell(\cdot; f_x); f_x) &= \int_0^1 L_{\ell(\cdot;f_x)}(t)\, f_x(t)\, \mathrm{d}t = \int_0^1 L_{\ell(\cdot;f_x)}(x(t))\, f(t)\, \mathrm{d}t \\ &\geq \int_0^1 L_{\ell(\cdot;f_x)}(t)\, f(t)\, \mathrm{d}t \geq \int_0^1 L_{\ell(\cdot;f)}(t)\, f(t)\, \mathrm{d}t = \mathcal{L}(\ell(\cdot; f); f)\ ,\end{aligned}$$

where the second equality holds since x is a measure-preserving transformation, the first inequality (in the second line) is the Hardy-Littlewood inequality once one realizes that, by Corollary 3.1, $L_{\ell(\cdot;f_x)}$ is a decreasing rearrangement of $L_{\ell(\cdot;f_x)} \circ x$; and the second inequality holds by the optimality of $\ell(\cdot; f)$. The second inequality is strict when $f_x \neq f$ since, in this case, the languages $\ell(\cdot; f)$ and $\ell(\cdot; f)$ satisfy different optimality conditions (3.3). □

Proof of Proposition 3.3

Proof. The proof is analogous to the proof of Proposition 3.2, now using that measure-preserving transformation of the word space leaves the imprecision loss unchanged (recall equation (3.1)). In this proof, the measure-preserving transformation $x : W \to W$ has to be chosen such that $c \circ x$ is increasing. □

A.2 Proofs of the results in Sections 4 and 5

Proof of Proposition 4.1

Proof. Proposition 4.1 is derived using standard calculus of variations. □

Proof of Proposition 5.1

Proof. We first prove the "only if" direction. Assume that $\ell : T \to W$ is an optimal language. We make the proof for the case where c is strictly increasing and $\tilde{f}$ is strictly decreasing (the other cases are analogous).[21] This guarantees that, by Lemmas 3.1 and 3.2, we have that if $\tilde{f}(\tilde{t}) > \tilde{f}(\tilde{t}')$ then $\ell(\tilde{t}) \leq \ell(\tilde{t}')$ for almost all $\tilde{t}, \tilde{t}' \in \tilde{T}$. In this case, we can define $\tilde{\tau} = h \circ \ell$, where $h(w) \equiv \int_0^w 1_\ell(w')\, \mathrm{d}w'$ for all $w \in [0, \ell(1)]$. Notice that $\tilde{\tau}$ satisfies the equations in the main text: $\mu_{\tilde{\tau}}(\tilde{\tau}(T')) = \mu(T')$ for all $T' \subset T$, and if $f(t) > f(t')$ then $\tilde{\tau}(t) \leq \tilde{\tau}(t')$ for almost

[21] Relaxing such assumptions implies that there are sets of states with the same likelihood with positive measure, and/or positive-measure sets of words with the same complexity. This makes the proof more tedious, but the steps are in it the same.

all $t, t' \in T$. Let us define the normalized language $\tilde{\ell}(\tilde{t}) = h^{-1}(\tilde{t})$ for all $\tilde{t} \in [0,1]$. Clearly, $\tilde{\ell} \circ \tilde{\tau} = \ell$, and therefore we have that $\mathcal{L}(\tilde{\ell}) = \mathcal{L}(\ell)$, so $\tilde{\ell}$ is an optimal language. (Notice that, trivially, $\mathcal{L}(\tilde{\ell}) \leq \mathcal{L}(\ell)$ (for any normalized language $\tilde{\ell}$) since from any normalized language we can generate a language on the original state space using the function $\tilde{\tau}$ described above.)

To prove the "if" part, assume now that $\tilde{\ell}$ is an optimal normalized language. Assume that there is some choice of $\tilde{\tau}$ with the properties in the main text such that $\tilde{\ell} \circ \tilde{\tau}$ is not optimal, that is, there is some language ℓ such that $\mathcal{L}(\tilde{\ell}) < \mathcal{L}(\ell)$. In this case, we can use the arguments in the proofs of Lemmas 3.1 and 3.2 to construct a language ℓ^* such that $\mathcal{L}(\ell^*) \geq \mathcal{L}(\ell)$ and such that satisfies both lemmas (note that the proofs are constructive). Then, we can use the first part of the proof to show that $\mathcal{L}(\ell^*) = \mathcal{L}(\tilde{\ell}^*)$ for some normalized language ℓ^*. Finally, we have

$$\mathcal{L}(\tilde{\ell}) < \mathcal{L}(\ell) \leq \mathcal{L}(\ell^*) = \mathcal{L}(\tilde{\ell}^*) \leq \mathcal{L}(\tilde{\ell}) \ ,$$

which is a contradiction. Then, $\tilde{\ell} \circ \tilde{\tau}$ is an optimal language. □

Proof of Proposition 5.2

Proof. Fix an optimal language ℓ. Given that C increases towards infinity, it is clear that $\ell(1) < \infty$ in an optimal language. Let $K \in \mathbb{R}_+ \cup \{+\infty\}$ be defined as $\lim_{t \nearrow 1} \ell'(t) = K$. Assume $K < +\infty$ (the case $K = +\infty$ is analogous). Fix $\varepsilon \in (0,1)$ and $\delta > -1$. Define the language $\tilde{\ell}$ such that

$$\tilde{\ell}(t) = \begin{cases} \ell(t) & \text{if } t \leq 1-\varepsilon \text{ and} \\ \ell(1-\varepsilon) + (1+\delta)\left(\ell(t) - \ell(1-\varepsilon)\right) & \text{if } t > 1-\varepsilon. \end{cases}$$

Notice that $\tilde{\ell}$ is continuous at $1-\varepsilon$, strictly increasing and coincides with ℓ whenever $\delta = 0$. Since $\tilde{\ell}$ is a feasible language, the optimality of ℓ implies that

$$0 \leq \int_{1-\varepsilon}^{1} \left(h(\ell'(t)) + c(\ell(t)) - h(\tilde{\ell}'(t)) - c(\tilde{\ell}(t))\right) f(t)\, \mathrm{d}t + C(\ell(1)) - C(\tilde{\ell}(1)) \ ,$$

where h is defined in the statement of Proposition 3.1. As $\delta \to 0$, the previous inequality can be written as follows:

$$0 \leq \delta \int_{1-\varepsilon}^{1} \left(\left(\ell(t) - \ell(1-\varepsilon)\right) c'(\ell(t)) + h'(\ell'(t))\, \ell'(t)\right) f(t)\, \mathrm{d}t + C'(\ell(1))\, \delta\, \varepsilon + O(\delta^2) \ .$$

Given that the previous inequality holds for small δ (both positive and negative) and small $\varepsilon > 0$ then the equation (5.2) holds. □

Proof of Propositions 5.3-5.5

Propositions 5.3 and 5.5 are derived using standard calculus of variations. Proposition 5.4 follows from standard change-of-variables arguments.

B Limits of Discrete Languages

In this section we show that our model (Section 2) can be obtained as the limit of some communication models in the economics literature. We will focus our proofs on showing the convergence of a version of the S model (Sobel, 2015) where we allow words to be heterogeneously complex, and we will provide some intuitions on how our model also can be obtained as the limit of the CGP model (see Remark B.1 below).

Consider a set of states of the world $T \equiv [0,1]$. We define a set of words $W \equiv [0,\bar{w})$ for some $\bar{w} > 0$. Using a notation similar to the one used in Remark 2.1, for a fixed $\Delta > 0$ and $k \in \{0, ..., \lfloor \bar{w}/\Delta \rfloor - 1\}$, the interval $w \equiv [k\,\Delta, (k+1)\,\Delta)$ is interpreted as a single word, also denoted (with an obvious abuse of notation) $w \equiv k\,\Delta$. Note that for each value $\Delta > 0$ there are $\lceil \bar{w}/\Delta \rceil$ available words. The set of available words is denoted $W_\Delta \equiv \{0, \Delta, ..., \lfloor \bar{w}/\Delta \rfloor\}$.

The usage cost that using a word $w \in W_\Delta$ involves is $c(w)$, where $c : W \to \mathbb{R}_+$ is an increasing function. As in our base model, the states are distributed according to an absolutely-continuous full-support F, which density f, but now we do not require it to be decreasing. Note that decreasing the value of Δ implies increasing the number of available words (equal to $\lfloor \bar{w}/\Delta \rfloor$) but approximately keeping their complexity distribution. Note that the results can be generalized to the case where there is an infinite number of words (and $\lim_{w\to\infty} c(w) = +\infty$, by taking the limit $\bar{w} \to \infty$.

As our model in Section 2, a language is a function $\ell : T \to W$. Additionally, now without loss of generality, we again require that the corresponding pushforward measure of μ, denoted μ_ℓ, is absolutely continuous.[22] We use the following form of the loss function for the CS model with complexity costs

$$\mathcal{L}_\Delta^{\text{CS}}(\ell) \equiv \sum_{w \in W_\Delta} \int_{\ell^{-1}(w)} \left(\Delta^{-2}\, L(t, \alpha(\ell^{-1}(w))) + c(w)\right) f(t)\, \mathrm{d}t \tag{B.1}$$

where, with some abuse of notation, $\ell^{-1}(w) \equiv \ell^{-1}([w, w+\Delta))$, and where $\alpha(\ell^{-1}(w))$ is an

[22] It is convenient to define a language as a map from T to W instead of to W_Δ, so the definition is independent of Δ. Furthermore, notice that for a given Δ, the requirement that "the corresponding pushforward measure of μ, denoted μ_ℓ, is absolutely continuous" is innocuous in terms of the payoff in (B.1) and the observation below that an optimal language must be an interval language.

"optimal" action taken by the receiver in the set of states $\ell^{-1}(w)$, that is,

$$\alpha(\ell^{-1}(w)) \in \arg\min_{a\in\mathbb{R}} \int_{\ell^{-1}(w)} L(t,a)\, f(t)\, \mathrm{d}t \ ,$$

and where L is a function from $T \times \mathbb{R}$ to $\mathbb{R}$. Sometimes, with some abuse of notation, when the language used is clear, we will use $L(t,w)$ to denote $L(t,\alpha(\ell^{-1}(w)))$. The term Δ^{-2} in the loss function (B.1) is irrelevant for a fixed $\Delta > 0$, but it will be useful to keep the imprecision and speaking costs balanced in optimal languages when $\Delta \to 0$.

As it is usual in the cheap talk literature, L is assumed to be twice differentiable almost everywhere and $-L(t,\cdot)$ to be single peaked around t, and we normalize $L(t,\cdot) = 0$. . Furthermore, we define

$$L_2(t) \equiv \frac{1}{2}\frac{\partial^2 L(t,a)}{\partial a^2}\bigg|_{a=t} \ ,$$

and we assume that $L_2(t) > 0$ for all $t \in T$. We finally assume that $\frac{\partial^2 L(t,a)}{\partial a \partial t} > 0$. Using these assumptions and an argument similar to the one in Crawford and Sobel (1982), it is easy to show that if a language ℓ is optimal then it is essentially equivalent (i.e., generates the same joint distribution of states and actions) to a language where states are pooled in intervals, that is, for all $w \in W_\Delta$ satisfies

$$F\big(\ell^{-1}([w, w+\Delta))\big) > 0 \quad \Rightarrow \quad \mathrm{supp}(F|_{\ell^{-1}([w,w+\Delta))}) \text{ is an interval.}$$

It is important to notice that, for a fixed $t \in T$ and $\Delta > 0$, we have that, as $\varepsilon \to 0$, the following holds:

$$\begin{aligned}\int_t^{t+\varepsilon} \Delta^{-2}\, L(t', \alpha(t, t+\varepsilon))\, f(t')\, \mathrm{d}t' &= \varepsilon^2\, \Delta^{-2} \int_t^{t+\varepsilon} \frac{t'^2}{4}\, L_2(t')\, f(t')\, \mathrm{d}t' \\ &= \frac{\varepsilon^2 \Delta^{-2}}{12} \left(L_2(t)\, f(t)\, \varepsilon + O(\varepsilon^2)\right) . \qquad \text{(B.2)}\end{aligned}$$

This is true because when $\varepsilon > 0$ is small, $L(t', \alpha(t, t+\varepsilon))$ approximates a quadratic function $L_2(t)\,(t' - \alpha(t, t+\varepsilon))^2$ for $t' \in [t, t+\varepsilon]$, because $\alpha(t, t+\varepsilon)$ is close to $t + \frac{\varepsilon}{2}$. and the distribution of states in $[t, t+\varepsilon]$ approximates a locally-uniform distribution with density $f(t)$.

Remark B.1. In the CGP model, the imprecision loss that a word w generates (conditional on being used) depends mass of states it means, that is, it is a function g of $\mu(\ell^{-1}(w))/\Delta$ (see Remark 2.1). In the S model, this loss in an optimal language is approximated by $\frac{\Delta^{-2}}{12}\, L_2(t)\, \mu(\ell^{-1}(w))^3$ (see equation (B.2)), where t is any state in $\ell^{-1}(w)$. Then, the length of equilibrium intervals is small (which we will see corresponds to Δ small) the CGP model allows for a more general function of the precision of the words (in the S model it is locally

quadratic), but the S model allows for an imprecision loss that depends on the state through a state-dependent factor $L_2(t)$. Our model in Section 5.3 can then be obtained as a generalized model combining the features of both models.

Remark B.2. The JMR model considers a multidimensional state space, and assumes (like the CGP and S models) that there is a finite number of words to be used in communication. To make the comparison to our model easier, assume that the state space is $[0,1]^2$. In their model, the (imprecision) payoff loss that a word generates (conditional on being used) is equal to a function of the average distance between the states signified by the word. They show that, for each fixed number n of available words, an optimal language partitions the state space into "ceils" (instead of the one-dimensional intervals), which form a "Voronoi tessellation" of the state space. Their construction suggests that, as the number of words n increases, a close-to-optimal language can be (heuristically) obtained as follows:

Divide the state space in $\lfloor\sqrt{n}\rfloor$ identical squares. Consider the most efficient language with n words *restricted* so that each ceil is fully contained in one of the $\lfloor\sqrt{n}\rfloor$ squares. This language is, in general, not optimal: an optimal language is likely to have some ceils which are not fully included in one of the squares. Nevertheless, it is not difficult to see that, as n increases, the payoff loss of the constrained-optimal language is going to become increasingly similar (in relative terms) to the unconstrained-optimal language. The reason is that the fraction of ceils in the optimal language that intersect more than one square decreases as $(\sqrt{n})^{-1}$. It is also not difficult to see that, given that the states in each of the $\lfloor\sqrt{n}\rfloor$ squares are going to be increasingly similarly likely as n grows (by the continuity of f), a constrained-optimal language is going to assign a similar number of words to squares containing states with a similar likelihood. Therefore, as n grows, the precision of the word describing a given state (compared to the precision of the word describing other states) is going to be determined by the likelihood of such a state. One can finally show that the structure of the language as n grows (i.e., the relationship between complexity of a word and its precision or states it describes) will only depend on the level sets of f, that is, rearrangements of the state space or mappings into lower-dimensional state spaces which preserve the quantile function associated to f do not change the structure of optimal languages.

B.1 Convergence of Sobel (2015)

The main result of this section, Proposition B.1, states that there is a language ℓ such that, for any $(\Delta_n)_n$ converging to 0, the payoff losses that the corresponding optimal languages generate converge to the payoff loss generated by the language ℓ. It further establishes that ℓ minimizes a payoff loss function which can be expressed similarly at in the general-

ized version of our model in Section 5.3 (equation (5.3)), with an additional state-dependent factor multiplying the imprecision loss that each state generates.

Proposition B.1. *Fix a sequence $(\Delta_n)_n$ decreasing towards 0 and a sequence $(\ell_n)_n$ of corresponding optimal languages. Then, there exists language ℓ such that*

$$\lim_{n\to\infty} \mathcal{L}^{\mathrm{CS}}_{\Delta_n}(\ell_n) = \mathcal{L}^{\mathrm{CS}}(\ell) \equiv \int_T \left(\frac{L_2(t)\, 1_\ell(\ell(t))^2}{12} + c(\ell(t)) \right) f(t)\, \mathrm{d}t \ .$$

Furthermore, ℓ minimizes $\mathcal{L}^{\mathrm{CS}}(\tilde{\ell})$ among all languages $\tilde{\ell}$.

Proof. We prove our result for the case that the payoff loss function, the usage cost function, and distribution function are step functions. The result for general functions can be obtained taking limits of step functions to approximate them. Hence, we assume that there is an interval partition of $[0,1]$ with N_f intervals, referred to as f-intervals, where both f and L_2 are constant; and that there is an interval partition of $[0,\bar{w}]$ with N_c intervals, referred to as c-intervals, where the function c is constant. There exists then a set of thresholds $\{t_i\}_{i=1}^{N_f+1}$ such that both f and L_2 are constant in $[t_i, t_{i+1})$ for all $i = 1, ..., N_f$. Analogously, there exists a set of thresholds $\{w_j\}_{j=0}^{N_c}$ such that c is constant in $[w_j, w_{j+1})$ for all $j = 1, ..., N_c$.

We fix, for the rest of the proof, some sequence $(\Delta_n)_n$ strictly decreasing towards 0, and a corresponding sequence of languages $(\ell_n)_n$ where, for each n, ℓ_n is an optimal language when the value of Δ is Δ_n. The first lemma establishes that, as n increases, the mass of states signified by a word in an optimal language asymptotically decreases as Δ_n or faster:

Lemma B.1. *The following inequality holds:*

$$\limsup_{n\to\infty} \max_{w\in W_{\Delta_n}} \frac{\mu(\ell_n^{-1}(w))}{\Delta_n} < \infty \ .$$

Proof. Assume, for the sake of contradiction, that $\lim_{n\to\infty} \max_{w\in W_{\Delta_n}} \frac{\mu(\ell_n^{-1}(w))}{\Delta_n} = \infty$.[23] Let $(w_n)_n$ be a sequence of words such that $w_n \in W_{\Delta_n}$ for all n and such that, defining $\varepsilon_n \equiv \mu(\ell_n^{-1}(w_n))$ for all n, we have $\lim_{n\to\infty} \varepsilon_n/\Delta_n = \infty$. Note that the contribution of the complexity cost of w_n is of order $\mu(\ell_n^{-1}(w_n)) = \varepsilon_n$. Then, using equation (B.2), the payoff loss that the interval $\ell_n^{-1}(w_n)$ generates is, as $n \to \infty$,

$$\frac{\Delta_n^{-2}}{12} L_2(t_n)\, f(t_n)\, \varepsilon_n^3 + O(\varepsilon_n + \varepsilon_n^4\, \Delta_n^{-2}) \ , \tag{B.3}$$

where t_n is the state in $\ell_n^{-1}(w_n)$ with highest value for $L_2(t_n)\, f(t_n)$. Given that, for each $n \in \mathbb{N}$ there are $\lfloor \bar{w}/\Delta_n \rfloor$ available words, there is an f-interval and two words w_n^1 and w_n^2 such that

[23] If $\lim_{n\to\infty} \max_{w\in W_{\Delta_n}} \frac{\mu(\ell_n^{-1}(w))}{\Delta_n}$ does not exist, but $\lim_{n\to\infty} \max_{w\in W_{\Delta_n}} \frac{\mu(\ell_n^{-1}(w))}{\Delta_n} = \infty$, then use the same argument using a sub sequence of Δ_n such that $\lim_{n\to\infty} \max_{w\in W_{\Delta_n}} \frac{\mu(\ell_n^{-1}(w))}{\Delta_n} = \infty$.

$\mu(\ell^{-1}(w_n^1)) \cup \mu(\ell_n^{-1}(w_n^2))$ is an interval and such that $\max_{i\in\{1,2\}} \mu(\ell_n^{-1}(w_n^i)) \leq 2/\lfloor \bar{w}/\Delta_n \rfloor$.[24] We can then define a new language where all words have the same meaning as in ℓ_n except for w_n, w_n^1 and w_n^1, which are now assigned as follows. In the new language, w_n^1 refers to the interval $\ell_n^{-1}(w_n^1) \cup \ell_n^{-1}(w_n^2)$. Now, the meaning of w_n is "split" in two parts of the same length, the first now meant by w_n^2 and the second by w_n. In this case, the payoff loss that the resulting intervals generate is given by[25]

$$2\frac{\Delta_n^{-2}}{12} L_2(t_n)\, f(t_n)\, \frac{\varepsilon_n^3}{8} + O(\varepsilon_n + \varepsilon_n^4\, \Delta_n^{-2}) \tag{B.4}$$

which, if n is large, it is roughly 4 times smaller than its original payoff loss, contradicting the optimality of the original language. □

We proceed in the proof of Proposition B.1 by introducing the concept of a step language. A language ℓ is said to be a *step language* if:

1. $\ell^{-1}([w_j, w_{j+1})) \cap [t_i, t_{i+1})$ is either empty or an interval for all i and j, and
2. ℓ linear and strictly increasing in $\ell^{-1}([w_j, w_{j+1})) \cap [t_i, t_{i+1})$ (whenever is not empty), for all i and j, and we use $\ell'_{i,j}$ to denote the derivative.

Step languages are natural candidates of optimal languages in the limit $n \to \infty$, given the step structure of the payoff, likelihood and word-cost functions. While the first condition is technical and for convenience, the second condition requires that if two words w, w' are equally costly (that is, belong to $[w_j, w_{j+1})$ for some j) and refer to the same type of states (that is, $\ell^{-1}(w), \ell^{-1}(w') \subset [t_i, t_{i+1})$ for some i) then they are equally precise. Notice that a step language is differentiable almost everywhere.

Lemma B.2. *Let ℓ be a step language. Then,*

$$\lim_{n\to\infty} \mathcal{L}^{\mathrm{CS}}_{\Delta_n}(\ell) = \mathcal{L}^{\mathrm{CS}}(\ell)\ .$$

Proof. To prove the result, define first, for each n,

$$W^{i,j}_{\ell,\Delta_n} \equiv \left\{ w \in W_{\Delta_n} \,\middle|\, w \in [w', w'+\Delta_n) \subset [w^j, w^{j+1}) \cap \ell([t^i, t^{i+1})) \text{ for some } w' \in W_{\Delta_n} \right\}.$$

The set $W^{i,j}_{\ell,\Delta_n}$ contains the words w such that are fully included in the interval $[w^j, w^{j+1})$, and their meaning is fully included in the interval $[t^i, t^{i+1})$. We define $W_{\ell,\Delta_n} \equiv \cup_{i,j} W^{i,j}_{\ell,\Delta_n}$.

[24] If two words or more are not used the result is trivial, since these can be taken as w_n^1 and w_n^2. If, instead, all words except maybe one are used, then given that there are $\lfloor \bar{w}/\Delta_n \rfloor$ words, at least half of the words satisfy that the mass of states they mean is no higher than $1/(\lfloor \bar{w}/\Delta_n \rfloor - 1) < 2/\lfloor \bar{w}/\Delta_n \rfloor$, so there is at least one pair of consecutive words with the required conditions.

[25] Note that the additional speaking cost remains $O(\varepsilon_n\, \Delta_n^2)$, which converges to 0 faster than the terms (B.3) and (B.4), since $\lim_{n\to\infty} \varepsilon_n/\Delta_n \to \infty$.

Note then, that for each $w \in W_{\ell,\Delta_n}$ there are two unique indexes $i_{\ell,\Delta_n}(w)$ and $j_{\ell,\Delta_n}(w)$ such that $[w, w+\Delta_n) \subset [w^{j_{\ell,\Delta_n}(w)}, w^{j_{\ell,\Delta_n}(w)+1})$ and $\ell^{-1}(w) \subset [t^{i_{\ell,\Delta_n}(w)}, t^{i_{\ell,\Delta_n}(w)+1})$.

Note that, since there are at most $N^c\, N^f$ words which do not belong to some $W^{i,j}_{\ell,\Delta_n}$, Lemma B.1 implies that the contribution of these words to the payoff loss vanishes as Δ_n gets small. Also, if $w, w' \in W^{i,j}_{\ell,\Delta_n}$ for some i, j, then $\mu(\ell^{-1}(w)) = \mu(\ell^{-1}(w')) = \Delta_n \ell'_{i,j}$.

Using equations (B.1) and (B.2), we have that, since ℓ is piecewise linear, as $n \to \infty$,

$$\begin{aligned}
\mathcal{L}^{\mathrm{CS}}_{\Delta_n}(\ell) &= \sum_{w \in W_{\ell,\Delta_n}} \left(\frac{L_2^{i(w)}}{12\, \ell'^2_{i_{\ell,\Delta_n}(w), j_{\ell,\Delta_n}(w)}} + c^{j(w)} \right) \mu(\ell^{-1}(w))\, f^{i(t)} + o(\Delta_n^0) \\
&= \sum_{i,j} \left(\frac{L_2^i}{12\, \ell'^2_{i,j}} + c^j \right) \mu(\ell^{-1}(W^{i,j}_{\ell,\Delta_n}))\, f^i + o(\Delta_n^0) \\
&= \int_T \left(\frac{L_2^{i(t)}}{12\, \ell'(t)^2} + c(\ell(t)) \right) f(t)\, \mathrm{d}t + o(\Delta_n^0)
\end{aligned}$$

as $n \to \infty$, where we used again Lemma B.1. Since $1_\ell(\ell(t))$ is equal to $\ell'(t)$ almost everywhere, the last expression of the previous equation is equal to $\mathcal{L}^{\mathrm{CS}}(\ell) + o(\Delta_n^0)$ as $n \to \infty$, so it is clear that our result holds. □

We now establish the existence of a step language which is asymptotically optimal, that is, such that, as $n \to \infty$, generates a payoff loss which is close to the one generated by an optimal language.

Lemma B.3. *There exists a step language ℓ such that*

$$\lim_{n\to\infty} \mathcal{L}^{\mathrm{CS}}_{\Delta_n}(\ell_n) = \mathcal{L}^{\mathrm{CS}}(\ell)\ .$$

Proof. Define, for each n, the double-matrix $M_n \equiv \big(\mu(W^{i,j}_{\ell_n,\Delta_n}), \mu(\ell_n^{-1}(W^{i,j}_{\ell_n,\Delta_n}))\big)_{i,j}$, where $W^{i,j}_{\ell_n,\Delta_n}$ is defined in the proof of Lemma B.2. Since $(M_n)_n$ is a bounded finite-dimensional sequence, by the Bolzano-Weierstrass Theorem it has a subsequence $(M_{k_n})_n$ converging to some $(\mu_1^{i,j}, \mu_2^{i,j})_{i,j}$. Also, given the convexity of the loss function, if w, and w' belong to $W^{i,j}_{\ell_n,\Delta_n}$ for some i, j, then the intervals $\ell_n^{-1}(w)$ and $\ell_n^{-1}(w')$ have the same length. Then, using a similar approximation as in the proof of Lemma B.2, we have that

$$\begin{aligned}
\mathcal{L}_{\Delta_n}(\ell_n) &= \sum_{i,j} \left(\frac{\mu\big(\ell_n^{-1}(W^{i,j}_{\ell_n,\Delta_n})\big)^2}{12\, \mu(W^{i,j}_{\ell_n})^2}\, L_i + c_j \right) \mu\big(\ell_n^{-1}(W^{i,j}_{\ell_n,\Delta_n})\big)\, f_i + O(\Delta_n) \\
&\to \underbrace{\sum_{i,j} \left(\frac{(\mu_2^{i,j})^2}{12\, (\mu_1^{i,j})^2}\, L_i + c_j \right) \mu_2^{i,j}\, f_i}_{(*)}\ ,
\end{aligned}$$

with the convention that $\frac{0}{0} = 0$.

Note that, necessarily, $\sum_i \mu_1^{i,j} \leq w^{j+1} - w^j$, and $\sum_j \mu_2^{i,j} = t^{i+1} - t^i$. Then, there exists some step language ℓ satisfying $\mu_1^{i,j} = \mu([w_j, w_{j+1}) \cap \ell([t_i, t_{i+1}))$ and $\mu_2^{i,j} = \mu(\ell^{-1}([w_j, w_{j+1})) \cap [t_i, t_{i+1}))$. Furthermore, $(*)$ coincides with the payoff loss of μ. Thus, limit of the optimal payoff losses $(\mathcal{L}_{\Delta_n}(\ell_n))_n$, which corresponds to $(*)$, is equal to $\mathcal{L}^{\text{CS}}(\ell)$.

Finally note that, necessarily, the limit of the payoffs of any subsequence of $(\ell_{k'_n})_n$ of $(\ell_n)_n$ is $\mathcal{L}^{\text{CS}}(\ell)$. Otherwise, there would be a subsequence of optimal languages with payoffs converging to either $y < \mathcal{L}^{\text{CS}}(\ell)$ or $y > \mathcal{L}^{\text{CS}}(\ell)$. The first case can be easily ruled out since, for n large enough, we would have $\mathcal{L}^{\text{CS}}_{\Delta_n}(\ell) > \mathcal{L}^{\text{CS}}_{\Delta_n}(\ell_{k'_n})$, contradicting the optimality of $\ell_{k'_n}$. In the second case the argument can be inverted, using the fact that there would a step language $\hat{\ell}$ with $\mathcal{L}^{\text{CS}}(\hat{\ell}) = y$. □

It is only left to prove that the language ℓ obtained in Lemma B.3 minimizes $\mathcal{L}^{\text{CS}}(\tilde{\ell})$ among all languages $\tilde{\ell}$. The reasoning to complete the proof standard: Assume, for the sake of contradiction, there is a language $\hat{\ell}$ such that $\mathcal{L}^{\text{CS}}(\hat{\ell}) < \mathcal{L}^{\text{CS}}(\ell)$. In this case, since $\lim_{n\to\infty} \mathcal{L}^{\text{CS}}_{\Delta_n}(\hat{\ell}) = \mathcal{L}^{\text{CS}}(\hat{\ell})$, we have that $\mathcal{L}^{\text{CS}}_{\Delta_n}(\hat{\ell}) < \mathcal{L}^{\text{CS}}_{\Delta_n}(\ell_n)$ for n large enough, contradicting the optimality of ℓ_n. □

References

Auerbach, Felix (1913) "Das Gesetz der Bevölkerungskonzentration," *Petermanns Geographische Mitteilungen*, Vol. 59, pp. 74–76.

Austen-Smith, David and Jeffrey S Banks (2000) "Cheap talk and burned money," *Journal of Economic Theory*, Vol. 91, pp. 1–16.

Axtell, Robert L (2001) "Zipf distribution of US firm sizes," *Science*, Vol. 293, pp. 1818–1820.

Blume, Andreas (2000) "Coordination and learning with a partial language," *Journal of Economic Theory*, Vol. 95, pp. 1–36.

——— (2004) "A learning-efficiency explanation of structure in language," *Theory and decision*, Vol. 57, pp. 265–285.

Boas, Franz (1911) "Introduction to the handbook of North American Indians," *Smithsonian Institution Bulletin*, Vol. 40.

Crawford, Vincent P and Joel Sobel (1982) "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, pp. 1431–1451.

Cŕemer, Jacques, Luis Garicano, and Andrea Prat (2007) "Language and the Theory of the Firm," *The Quarterly Journal of Economics*, pp. 373–407.

Cunha, Carlos, Azer Bestavros, and Mark Crovella (1995) "Characteristics of WWW client-based traces,"Technical report, BU-CS-95-010, Computer Science Department, Boston University.

Dilmé, Francesc (2018) "Strategic Communication with a Small Conflict of Interest."

Ferrer-i-Cancho, Ramon (2017) "Optimization models of natural communication," *Journal of Quantitative Linguistics*, pp. 1–31.

Ferrer-i-Cancho, Ramon and Ricard V Solé (2003) "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences*, Vol. 100, pp. 788–791.

Gabaix, Xavier (1999) "Zipf's law for cities: an explanation," *Quarterly journal of Economics*, pp. 739–767.

Hardy, Godfrey Harold, John Edensor Littlewood, and George Pólya (1952) *Inequalities*: Cambridge university press.

Hertel, Johanna and John Smith (2013) "Not so cheap talk: Costly and discrete communication," *Theory and decision*, Vol. 75, pp. 267–291.

Huberman, Bernardo A, Peter LT Pirolli, James E Pitkow, and Rajan M Lukose (1998) "Strong regularities in world wide web surfing," *Science*, Vol. 280, pp. 95–97.

Jäger, Gerhard, Lars P Metzger, and Frank Riedel (2011) "Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals," *Games and economic behavior*, Vol. 73, pp. 517–537.

John, Andrew (2016) "Dynamic Models of Language Evolution: The Economic Perspective," in *The Palgrave Handbook of Economics and Language*: Springer, pp. 101–120.

Kartik, Navin (2009) "Strategic communication with lying costs," *The Review of Economic Studies*, Vol. 76, pp. 1359–1395.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani (2007) "Credulity, lies, and costly talk," *Journal of Economic theory*, Vol. 134, pp. 93–116.

Li, Wentian (1992) "Random texts exhibit Zipf's-law-like word frequency distribution," *Information Theory, IEEE Transactions on*, Vol. 38, pp. 1842–1845.

Lim, Wooyoung and Qinggong Wu (2017) "Vague Language and Context Dependence."

Lipman, Barton L. (2009) "Why is language vague?".

Mandelbrot, Benoit (1953) "An informational theory of the statistical structure of language," *Communication theory*, Vol. 84, pp. 486–502.

Marschak, Jacob (1965) "Economics of language," *Systems Research and Behavioral Science*, Vol. 10, pp. 135–140.

Martin, Laura (1986) ""Eskimo words for snow": A case study in the genesis and decay of an anthropological example," *American anthropologist*, Vol. 88, pp. 418–423.

Newman, Mark EJ (2005) "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, Vol. 46, pp. 323–351.

Nicolis, John S and Ichiro Tsuda (1989) "On the parallel between Zipf's law and 1/f processes in chaotic systems possessing coexisting attractors a possible mechanism for language formation in the cerebral cortex," *Progress of Theoretical Physics*, Vol. 82, pp. 254–274.

Piantadosi, Steven T, Harry Tily, and Edward Gibson (2011) "Word lengths are optimized for efficient communication," *Proceedings of the National Academy of Sciences*, Vol. 108, pp. 3526–3529.

Rockafellar, RT (2001) "Convex analysis in the calculus of variations," in *Advances in Convex Analysis and Global Optimization*: Springer, pp. 135–151.

Rubinstein, Ariel (1996) "Why are certain properties of binary relations relatively more common in natural language?" *Econometrica: Journal of the Econometric Society*, pp. 343–355.

——— (2000) *Economics and Language: Five Essays. The Churchill Lectures in Economic Theory*: Cambridge, Cambridge University Press.

Ryff, John V (1970) "Measure preserving transformations and rearrangements," *Journal of Mathematical Analysis and Applications*, Vol. 31, pp. 449–458.

Shannon, Claude E (1948) "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27, pp. 379–423.

Silagadze, Z K (1997) "Citations and the Zipf–Mandelbrot Law," *Complex Systems*, Vol. 11, pp. 487–499.

Sobel, Joel (2015) "Broad Terms and Organizational Codes," working paper.

Stanley, Michael HR, Sergey V Buldyrev, Shlomo Havlin, Rosario N Mantegna, Michael A Salinger, and H Eugene Stanley (1995) "Zipf plots and the size distribution of firms," *Economics letters*, Vol. 49, pp. 453–457.

Willis, John Christopher (1922) *Age and area: a study of geographical distribution and origin of species*: Cambridge University Press Cambridge.

Zipf, George Kingsley (1935) *The psycho-biology of language: an introduction to dynamic philology*, The MIT paperback series: Houghton Mifflin.

——— (1949) *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*: Addison-Wesley.