Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Artificial Intelligence 154 (2004) 1–42
Artificial Intelligence
www.elsevier.com/locate/artint

Optimizing the mutual intelligibility of linguistic agents in a shared world

Natalia Komarova [a,b], Partha Niyogi [c,*]

[a] *Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, USA*
[b] *Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT, UK*
[c] *Department of Computer Science, University of Chicago, Chicago, IL 60637, USA*

Received 4 October 2001; received in revised form 11 May 2003

Abstract

We consider the problem of linguistic agents that communicate with each other about a shared world. We develop a formal notion of a language as a set of probabilistic associations between form (lexical or syntactic) and meaning (semantic) that has general applicability. Using this notion, we define a natural measure of the mutual intelligibility, $F(L, L')$, between two agents, one using the language L and the other using L'. We then proceed to investigate three important questions within this framework: (1) Given a language L, what language L' maximizes mutual intelligibility with L? We find surprisingly that L' need not be the same as L and we present algorithms for approximating L' arbitrarily well. (2) How can one learn to optimally communicate with a user of language L when L is unknown at the outset and the learner is allowed a finite number of linguistic interactions with the user of L? We describe possible algorithms and calculate explicit bounds on the number of interactions needed. (3) Consider a population of linguistic agents that learn from each other and evolve over time. Will the community converge to a shared language and what is the nature of such a language? We characterize the evolutionarily stable states of a population of linguistic agents in a game-theoretic setting. Our analysis has significance for a number of areas in natural and artificial communication where one studies the design, learning, and evolution of linguistic communication systems.

Keywords: Linguistic agents; Optimal communication; Language learning; Language evolution; Game theory; Multi-agent systems

* Corresponding author.
E-mail address: niyogi@cs.uchicago.edu (P. Niyogi).

0004-3702/$ – see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.artint.2003.08.005

1. Introduction

Consider two linguistic agents in a shared world. The agents desire to communicate different messages (meanings) to each other. Such a situation arises in a number of different contexts in natural and artificial communication systems and it is important in such cases to be able to quantify the rate of success in information transfer, in other words, the *mutual intelligibility* of the agents. Each agent possesses a communicative device or a language that allows it to relate code (signal) and message, form and meaning, syntax and semantics, depending upon the context in which the communication arises. If they share the same language and this language is expressive enough and unambiguous, then mutual intelligibility will be very high. If on the other hand, they do not share the same language, or the languages are inexpressive or ambiguous, the mutual intelligibility will be much lower. This is often the case in the real world and in this paper, we present an analysis of this situation. We view languages as probabilistic associations between form and meaning and develop a natural measure of intelligibility, $F(L_1, L_2)$, between two languages, L_1 and L_2, which is a generalization of a similar function introduced in [10]. We ask the following question: if there is a biological/cultural/technological advantage for an agent to increase its intelligibility with the rest of the population, what are the ways to do this?

The task of increasing intelligibility reduces ultimately to three related sub-problems:

- Given a language L, what language L' maximizes the mutual intelligibility $F(L, L')$ for two way communication about the shared world?
- What are some acquisition mechanisms/learning algorithms that can serve the task of improving intelligibility?
- What are the consequences of individual language acquisition behavior on the population dynamics and the communicative efficiency of an interacting population of linguistic agents?

In this paper, we create a mathematical framework to address these questions analytically. We find, surprisingly, that the optimal language L' need not be the same as L, and we present an algorithm for approximating L' arbitrarily well (Section 3). The optimal language, L', can be either learned or inherited by each individual from its "parents". In the former case, we find some bounds on the performance of appropriate learning algorithms (Section 4). In the latter case, we study the resulting population dynamics in the context of an evolutionary language game (Section 5).

1.1. Communicability in animal, human and machine communication

The simplest situation where communicability is readily defined corresponds to the case where the "language" may be viewed as an association matrix, A. Such a matrix simply links referents to signals. If there are M referents and N signals, then A is an $N \times M$ matrix. The entries, a_{ij}, define the relative strength of the association between signal i and meaning j. The matrix A thus characterizes the behavior of the linguistic agent in (i) *production mode* where it may produce any of the signals corresponding to a particular meaning in proportion to the strength of the association, and in (ii) *comprehension mode*

where it may interpret a particular signal as any of the meanings in proportion to the association strengths.

The specific settings in which such a scheme is a useful description include animal communication, human languages and artificial languages. For instance, it often makes sense to talk about a *lexical matrix* as a formal description of human mental vocabularies. It is introduced to describe the arbitrary relations between discrete words and discrete concepts of human languages ([10,14,19,26]; also see [36] for a more Bayesian perspective). Each column of the lexical matrix corresponds to a particular word meaning (or concept), each row corresponds to a particular word form (or word image). In the Saussurean terminology of arbitrary sign, the lexical matrix provides the link between signifié and signifiant [28].

An equivalent of a lexical matrix is also at the basis of any animal communication system, where it defines the relation between animal signals and their specific meanings [4, 8,15,31,32]. A classic example of this is alarm calls in primates. There are a finite number of referents that are coded using acoustic signals and decoded appropriately by recipients.

Infinite association matrices can be used as a description of human languages [13,25]. Human grammars mediate a complex mapping between form and meaning. There, the space of possible signals is the set of all strings (sentences) over a finite syntactic alphabet and the set of possible meanings is the set of all strings over some semantic alphabet. Most crucially, the sets of possible sentences and meanings are infinite. This accounts for the infinite expressibility of human grammars.

In artificial intelligence, the problem arises in many different settings. A number of studies have emerged where linguistic agents interact with each other in simulated worlds and one studies whether coherent or coordinated communication ultimately emerges (see, for example, [2,12,21–23,33–35]). Much of this kind of research employs the simulation methodology of Artificial Life. In this paper, we create a mathematical framework for these kinds of problems and derive a number of analytic results. We also study language coordination in a game-theoretic setting and our results have consequences for the Nash equilibria for such problems (for related research on multi-agent systems and game-theoretic foundations, see [1,37] among others).

In the design of natural language understanding systems, the goal is to develop a computer system that is able to communicate with a human. The statistical approach to this problem assumes an underlying probabilistic model for the human source. This probabilistic model is then recovered or learned from data either by randomly drawn samples as in the case of corpus linguistics or statistical language modeling (see [3] and [16] for overviews of this point of view) or via some interactive exchanges and semantic reinforcement [7,11]. The primary implication of this paper is that optimal communication with a language user might require one to learn a language that is *different* from the target source.

1.2. Main results in the context of previous work

Here we outline the three main sets of results presented, respectively, in Sections 3, 4 and 5.

1.2.1. How to maximize the mutual intelligibility?

Let us consider a population of agents and assume that each of them has a language. An evolutionary process can then be described where individuals reproduce and the offspring do not have an innate language, but acquire a language on the basis of interaction with the population. This process was first explicitly modeled in [10] and later in [23] and [21]. In the approach of the latter two works, at each (discrete) moment of time, a randomly chosen individual is replaced by a new one, which then learns the language of the population; in [10], the generations do not overlap. It is clear that the choice of a learning procedure used by the offspring will influence the evolutionary dynamics that ensues and, in particular, whether or not the population will converge to (and maintain) a reasonably coherent language.

Several basic learning mechanisms have been considered. The "imitator" simply learns the averaged language of the population, both in the production and in the comprehension mode. The "calculator" of Hurford (called "obverter" in [23] and "Bayesian learner" in [21]) does not copy the language of the population but rather constructs the "best response" to it: it adopts the production behavior which is best understood by the population, and the comprehension strategy which is the best decoder for the population, thus maximizing its communicative efficiency with the population. The "Saussurean learner" of [10] imitates the production mode of the population and then adopts the comprehension behavior that maximizes its chances of understanding itself.

It turns out that imitators do not do very well and a coordinated communication system seems to be unstable in a population of such learners. Saussurean learners show a better performance, but the obverters are the most efficient (in the setting of [23] and [21]). Starting from a randomly chosen initial condition, a population of obverters quickly develops a highly coordinated communicative system, and reaches a state where signals and meanings are related in the one-to-one fashion (plus perhaps some isolated synonyms or homonyms).

A peculiar feature of both imitators and obverters is that their production and comprehension modes are completely de-correlated.[1] Before the perfect coordination of language is reached, some obverters might find themselves speaking a very strange language. Imagine a case where the language has two sentences, s_1 and s_2, and two meanings, m_1 and m_2. A pathological linguistic agent might use s_1 to communicate the meaning m_1 and s_2 to communicate m_2 in production mode but interpret s_1 to mean m_2 and s_2 to mean m_1 in comprehension mode. Such a linguistic agent is therefore self-contradictory in its associations between form and meaning.

In this paper, we avoid such internal contradictions by requiring that a linguistic agent's production and comprehension modes be linked via a common association matrix. In doing so, we have two motivating considerations in mind. First, from a cognitive standpoint, it seems natural to give symmetric consideration to form and meaning and treat language as a relation between form and meaning rather than two separate functional mappings for production and comprehension. Second, from a computational standpoint, a

[1] In other contexts, the obverter may be defined differently. A more general definition is that an obverter performs to maximize comprehension on the assumption that the hearer's reception behavior is the same as its own.

common association matrix provides a compact representation for a language from which production and comprehension modes may easily be derived. The Saussurean learner satisfies such a criterion; obverters that have reached a co-ordinated language do not violate this constraint (see also [20]). The communicating neural networks also provide a link between the production and comprehension modes, see, e.g., [20,29,30].

In this paper, we execute a comprehensive analysis of this situation. The first question we address is whether the obverter algorithm can still be carried out if the self-consistency constraint is imposed on each of the learners.[2] This requires us to understand what the best response to an arbitrary language is, when it exists, and how to approximate it. We demonstrate the following:

- If the language L is *not* self-consistent, then it is in general not possible to use the obverter procedure for finding the best response. In other words, the comprehension behavior and the production behavior designed to (separately) maximize the communicative efficiency, do *not* obey the self-consistency requirement.
- If the language L is fully co-ordinated (defines a one-to-one correspondence between signals and meanings), then the best response exists and is equal to the language L itself.
- Next, suppose that the language L is self-consistent, but not fully co-ordinated. Then even though it is not in general possible to find the best response, we can approximate it within any given accuracy, ε.
- Finally, suppose that the language L is fully co-ordinated, but the communication is noisy. Then, under some mild restrictions on the magnitude of the noise, we can still find the best response, and, under slightly stronger conditions, it is the language L itself.

Incidentally, the first of the above statements suggests that the obverter mechanism *cannot* be used for learning from a population of individuals. Even though each agent might have a self-consistent language, the average language of the population may not be self-consistent. The obverter procedure can be used by each newly introduced agent to learn from one randomly chosen individual, or from its "parent" (i.e., an individual chosen proportionally to its linguistic performance).

1.2.2. Learning the optimal language

A second set of results relate to the problem of *learning* a self consistent language for the purpose of optimal communication with the chosen teacher or "parent". Since the teacher's language is not known at the outset, the learner must obtain relevant estimates of it over a finite number of interactions with the teacher. This situation arises in a number of artificial intelligence settings where a machine learning approach is taken to acquire a language for communicative purposes. For example, statistical language modeling for spoken language understanding between human and machine, or language learning for robotic communication systems between two robots (machines) are natural applications. In

[2] The precise definition of self-consistency is that there exists a probability measure over the space of signals and meanings, which is common for both the production and the comprehension modes.

most such cases, particularly in statistical language modeling, it has been tacitly assumed that collecting a corpus by sampling the target language and then reconstructing it on the basis of this corpus is a sufficiently good strategy for designing a natural language based human–computer interaction system. In the light of the results presented in this paper, we will see that this assumption is mistaken. Specifically,

- We consider two different frameworks for learning: (i) learning with full information, where the learner has access to both the sentence and its meaning in all interactions, and (ii) learning with partial information, where the learner has direct access only to the sentence. The meaning is not directly accessible but the learner knows whether communication was ultimately successful. We present algorithms to learn how to communicate optimally in both settings.
- We present explicit bounds on the number of examples (interactions) needed for an agent to be able to learn a self consistent language that yields communicability that is arbitrarily close to optimal with high probability. In a partial information setting, the number of examples is seen to be proportional to N^2M^2/γ^2, where N is the number distinct signals and M is the number of distinct meanings. In the full information setting where meanings are directly observable, the number of examples reduces to a quantity proportional to N^2/γ^2. In both cases, γ is a margin parameter that characterizes the learning difficulty of the teacher's language.

It is interesting to compare our approach with the approach taken in the studies of populations of neural networks. Oliphant [21] and Smith [29] used numerical simulations to investigate the dynamics of an iterative learning model. While they did not address the question of convergence to a maximum communicability in a teacher-learner pair, they looked at the convergence of the population of networks to an optimal communication system. By varying the update rules of individual networks, they were able to show that a learning bias toward a one-to-one mapping between meanings and signals led to an emergence of a coordinated communication system. In their setting, each individual did not necessarily optimize its communication ability with the current population, but rather, each individual had a learning strategy which eventually facilitated a high long-term communicability outcome.

1.2.3. Communicability and the evolutionary language game

Finally, we examine the implications of the communicability function in the language game framework. There has been considerable recent activity with work on computational models for the evolution of natural languages and animal communication [8,12,23,25]. In models that are based on selective fitness, the communicability function determines the payoff of different languages. Individuals that communicate well receive a high payoff which translates into biological fitness, or reproductive success: individuals with higher fitness produce more children who learn their language. Alternatively, one can assume that individuals with a high payoff have a high reputation (or standing in the group) and are more influential as language teachers. The assumption is that language performance measured by the function F contributes to the rate with which each language is spread.

The function F defines the equilibria of the language game equation. In [38], such equilibria (called the *Nash equilibria*) were found for a system where the production and comprehension modes are independent. In this paper, we show that these results continue to hold if the requirement of self-consistency is imposed. In other words, all the stable and neutrally stable states of the language system can be attained even if the production and comprehension modes are cognitively related.

We also characterize all the evolutionarily stable strategies (ESS) [18] of the language system. It can be proved that the ("strict") ESS correspond to fully co-ordinated languages. However, there is another kind of an evolutionarily significant state, called *weak ESS*, which is stable modulo some random drift. This state have been observed in many numerical studies of language systems including the above mentioned [10,21,23]. In this paper we analytically prove that:

- If the frequency of occurrence of events (subjects of communication) in the shared world is exactly uniform (i.e., all events occur with exactly the same frequency), then weak ESS can be characterized as perfectly co-ordinated languages which might have some *isolated* synonyms or homonyms (see Section 5.3 for the precise definition).
- In the more general case of non-uniform frequencies of events, only isolated synonyms are possible, and homonyms are unstable.

The latter result means that ambiguous languages are evolutionarily unstable. Indeed, while true homonyms will reduce the communicability potential of a language, (isolated) synonyms will not. On the other hand, it is commonly observed that human languages have numerous homonyms, whereas true synonymy is extremely rare. To resolve this apparent contradiction, we have to remember about the presence of *context*.

Indeed, the relevant communicative accuracy of individuals should not be defined per-word, but rather per utterance. Therefore, the entries of the "lexical matrix" are not words as such, but slightly larger objects, which can roughly defined as "words in a context". As soon as we accept this level of description, then the results of the mathematical model correctly describe the following observation: *in human languages, there are practically no true homonyms that remain homonyms in the presence of contextual clues, and, on the other hand, context-dependent synonyms are rather common*.

To give some examples, let us first consider a lexical homonymy, such as "fall" (autumn) and "fall" (down). This is a complete homonymy on the level of words, but not on the level of larger utterances (clearly, the utterances "in the fall" and "to fall down" can never be confused). Similarly, the words "beautiful" and "fair" cannot be regarded as true synonyms on the level of words, but if we consider the utterances "beautiful lady" and "fair lady", they are interchangeable. This illustrates the point that when context is taken into account, then synonyms are possible, and homonyms are unstable.

The rest of the paper is structured as follows. In Section 2 we develop a general notion of association matrices as probability measures on the cross product of forms and meanings. We then show how a measure of communicative efficiency or mutual intelligibility may be naturally defined. In Section 3 we show how to construct an approximating family of languages that converges to the optimal communicator. We examine an extension of this in Appendix B where we study communication with a perfect language across a noisy

channel. We continue by examining the implications of our results for learning theory: in Section 4 we discuss algorithms for learning to communicate and present bounds on their sample complexity. Finally, in Section 5 implications for evolution are discussed; in particular, we classify all Nash equilibria and characterize the possible evolutionarily stable strategies. Conclusions are found in Section 6.

2. Communicability for linguistic systems

2.1. Basic notions

We regard a linguistic system to be an association between form and meaning. Let $\mathcal{S} \subset \mathbf{N}$ be the set of all possible linguistic forms (sentences or signals) and $\mathcal{M} \subset \mathbf{N}$ be the set of all possible semantic objects (meanings or referents). Note that depending on the context, the elements of $\mathcal{S}$ can be words, codes, expressions, forms, signals or sentences. The elements of $\mathcal{M}$ can be meanings, messages, events or referents. We will use the general term *signals* for elements of $\mathcal{S}$ and *meanings* for elements of $\mathcal{M}$.

The sets $\mathcal{S}$ and $\mathcal{M}$ need not be finite, but it is essential that they are enumerable. The reason the sets $\mathcal{S}$ and $\mathcal{M}$ can be viewed as countable for human languages has to do with the discrete nature of language. In the lexical setting, $\mathcal{S}$ is the set of all words, and therefore is naturally countable, and the countability of $\mathcal{M}$ (the meanings) is assured by categorization. In the case of human grammars, we may let $\mathcal{S} = \Sigma_1^*$ be the set of all possible strings over a syntactic alphabet (Σ_1) and $\mathcal{M} = \Sigma_2^*$ be the set of all possible strings over a semantic alphabet (Σ_2). Note that in this case $\mathcal{S}$ and $\mathcal{M}$ are infinite.

We define a communication system, or a language, as a probability measure μ over $\mathcal{S} \times \mathcal{M}$. Note that in the case of finite languages (human or artificial lexicons and animal communication systems), μ is related to the association matrix, A, by means of a simple rescaling.

Let us enumerate all possible signals, i.e., the elements of set $\mathcal{S}$, as $s_1, s_2, s_3, \ldots$ and all possible meanings (elements of $\mathcal{M}$) as $m_1, m_2, m_3, \ldots$. The coding and decoding schemes of the agent are contained in the measure μ in the following manner. Each user of μ is characterized by an *encoding matrix* P and a *decoding matrix* Q where

$$P_{ij} \equiv \mu(s_i|m_j) = \begin{cases} \mu(s_i, m_j)/\sum_p \mu(s_p, m_j), & \text{if } \sum_p \mu(s_p, m_j) > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$$Q_{ji} \equiv \mu(m_i|s_j) = \begin{cases} \mu(s_j, m_i)/\sum_p \mu(s_j, m_p), & \text{if } \sum_p \mu(s_j, m_p) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Both P and Q matrices are easily interpreted. P_{ij} is simply the probability of producing the signal s_i given that one wishes to convey the meaning m_j. Similarly, Q_{ij} is the probability of interpreting the expression s_i to mean m_j by the same user.

Matrices analogous to P and Q were introduced in [10], however, they were not explicitly related through a common measure, μ. An effective connection between P and Q has been employed for a particular learning mechanism, called the *Saussurean* [10,21].

Remarks.

1. The user of a language is characterized in *production mode* by the matrix P and in *comprehension mode* by the matrix Q. This captures the fact that given a particular meaning, there might be many different ways to express it. Correspondingly given a particular signal, there may be no unique interpretation. Thus ambiguities in sentence interpretation or polysemy in lexical semantics are incorporated.
2. A measure μ uniquely defines the corresponding P and Q matrices. The converse is not generally true: given the P and Q matrices it might be possible to find more than one μ which would have the correct encoding and decoding matrices. An example with 2×2 matrices is $P = Q = I$ and

$$\mu_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix},$$
$$\mu_2 = \begin{pmatrix} 1/3 & 0 \\ 0 & 2/3 \end{pmatrix}.$$

   Clearly, both μ_1 and μ_2 lead to the same P and Q. In order to avoid such ambiguities we introduce the equivalence classes of measures. We will say that two measures μ_1 and μ_2 are equivalent to each other ($\mu_1 \equiv \mu_2$) if and only if the corresponding P and Q matrices are equal, i.e., $P^{(1)} = P^{(2)}$ and $Q^{(1)} = Q^{(2)}$.
3. For a probability measure μ let us introduce

$$\mathcal{S}_\mu = \{s \in \mathcal{S} \mid \exists m \in \mathcal{M} \text{ s.t. } \mu(s,m) > 0\}.$$

   This defines the set of signals that are used in production or comprehension by a linguistic agent. In the sense of formal language theory, this is the set of well formed syntactic expressions. In fact, the set $\mathcal{S}_\mu$ is what is normally called "language". Our definition of language as a measure μ contains this notion of language as the *support* of μ. Similarly, one may define

$$\mathcal{M}_\mu = \{m \in \mathcal{M} \mid \exists s \in \mathcal{S} \text{ s.t. } \mu(s,m) > 0\}.$$

   This defines the set of all meanings that are expressible by the linguistic agent. If $\mathcal{M}_\mu = \mathcal{M}$ then all meanings can be expressed. If $\mathcal{M}_\mu$ is a proper subset of $\mathcal{M}$ then some meanings are left unexpressed.
4. The probability measure μ, the sets $\mathcal{S}_\mu$ and $\mathcal{M}_\mu$, and the matrices P and Q in humans and animals arise out of highly structured systems in the brain. In fact, it is clear that in human languages, these objects may not vary arbitrarily. A significant activity in generative linguistics attempts to characterize the nature of this structure and the variation that exists among natural languages of the world.

2.2. Probability of events and communicability function

The communicating agents are immersed in a world and the need to communicate messages arises as the corresponding events occur in this shared world. Thus one may define a measure σ on the set of possible meanings $\mathcal{M}$ according to which the agents need to communicate each of these meanings to each other. Given two communication

systems, i.e., languages μ_1 and μ_2, the probability that an event occurs whose meaning is successfully communicated from μ_1 to μ_2 is given by

$$P[1 \to 2] = \sum_i \sigma(m_i) \sum_j \mu_1(s_j|m_i)\mu_2(m_i|s_j).$$

Similarly, one may compute the probability with which an event is successfully communicated from μ_2 to μ_1 as

$$P[2 \to 1] = \sum_i \sigma(m_i) \sum_j \mu_2(s_j|m_i)\mu_1(m_i|s_j).$$

We may then define the effective *communicability function* of μ_1 and μ_2 as

$$F(\mu_1, \mu_2) = \tfrac{1}{2}\big(P[1 \to 2] + P[2 \to 1]\big).$$

In matrix notation, this may be written as

$$F(\mu_1, \mu_2) = \tfrac{1}{2}\big[\mathrm{tr}\big(P^{(1)}\Lambda\big(Q^{(2)}\big)^{\mathrm{T}}\big) + \mathrm{tr}\big(P^{(2)}\Lambda\big(Q^{(1)}\big)^{\mathrm{T}}\big)\big], \tag{3}$$

where Λ is a diagonal matrix such that $\Lambda_{ii} = \sigma(m_i)$, and $P^{(i)}$, $Q^{(i)}$ refer to the coding and decoding matrices associated with measure μ_i. Note that $\mathrm{tr}(P^{(1)}\Lambda(Q^{(2)})^{\mathrm{T}})$ is simply the probability that an event occurs and is successfully communicated from user of μ_1 to user of μ_2.

Remarks.

1. The function $F(\mu_1, \mu_2)$ is the average probability with which μ_1 and μ_2 understand each other in two way communication mode. The function $F(\mu_1, \mu_2)$ is symmetrical with respect to its arguments. If μ_1 is a probability measure with support only on the diagonal elements of $\mathcal{S} \times \mathcal{M}$, then the P and Q matrices are identity and the communicative efficiency is 1.
2. $F(\mu_1, \mu_1)$ is the communicability of two identical linguistic agents. We have

   $$0 < F(\mu_1, \mu_1) \leqslant 1.$$

   For two different agents μ_1 and μ_2 we also have

   $$0 \leqslant F(\mu_1, \mu_2) = F(\mu_2, \mu_1) \leqslant 1.$$

3. The marginals $\mu(m)$ and $\sigma(m)$ are not equal to each other. In other words, the language of an agent is simply given by μ and the conditional probabilities associated with it. The probability with which agents communicate different meanings is determined not by the language but by the external world in which the agents are grounded. Therefore, two agents might have high communicative efficiency in some world and low communicative efficiency in another one.
4. A function similar to our communicability function was introduced by Hurford [10]. However, all meanings were treated to have equal probabilities (a uniform measure σ), and thus the function was not suitable for infinite matrices.

3. Reaching the highest communicability

Let us assume that one of the languages is given and call this language μ_0. According to definition (3), for any language μ we have

$$F(\mu_0, \mu) = \frac{1}{2} \sum_{i,j} \sigma_j \big[\mu_0(s_i|m_j)\mu(m_j|s_i) + \mu(s_i|m_j)\mu_0(m_j|s_i) \big]. \tag{4}$$

Let us define the *best response* as a language μ_*, such that

$$F(\mu_0, \mu_*) = \sup_{\mu} F(\mu_0, \mu). \tag{5}$$

In what follows we will present an algorithm of building a best response or a language which in some sense approaches the best response. In particular, we show that the best response need not exist. However, an arbitrarily good response can be constructed. We show how to construct a sequence of languages (μ_ε where $\varepsilon > 0$) such that $F(\mu_0, \mu_\varepsilon)$ can be made arbitrarily close to $\sup_\mu F(\mu_0, \mu)$—the maximum possible mutual intelligibility between a user of μ_0 and a user of any allowable language.

The interesting question of finding the best response in a noisy environment is considered in Appendix B.

3.1. A special case of finite languages

In order to keep the argument as transparent as possible, we will first make three simplifying assumptions. The effect of relaxing these assumptions will be demonstrated in Section 3.2. For now we will assume that the following conditions are satisfied:

(i) The languages are finite, and the matrices have the size $N \times M$.
(ii) The distribution σ is uniform, i.e., $\sigma_i = 1/M \ \forall i$.
(iii) The measure μ_0 satisfies the *property of unique maxima*, i.e., for each i, there exist a unique $p_0(i)$ and a unique $r_0(i)$ such that

$$\mu_0(s_i|m_{p_0(i)}) = \max_p \mu_0(s_i|m_p), \quad \mu_0(m_i|s_{r_0(i)}) = \max_r \mu_0(m_i|s_r). \tag{6}$$

The last condition states that there exists strictly one element of each column of $\mu_0(s|m)$ (row of $\mu_0(m|s)$) such that it is the biggest element in the column (row).

Let us maximize each of the two terms in the right hand side of expression (4) separately. First, we find a matrix Q^* such that

$$\sum_{i,j} \mu_0(s_i|m_j) Q^*_{ij} = \max_Q \sum_{i,j} \mu_0(s_i|m_j) Q_{ij}, \tag{7}$$

where we maximize over all matrices Q whose elements are non-negative and sum up to one within each row. This results in the following definition of Q^*:

$$Q^*_{ij} = \begin{cases} 1, & \mu_0(s_i|m_j) = \max_p \mu_0(s_i|m_p), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

In other words, in order to construct the *best decoder*, Q^*, we need to find the largest elements in each of the rows of $\mu_0(s|m)$ and put "ones" at the correspondent slots of Q^*. The rest of the entries of the matrix Q^* are zero. This is a well defined operation because of the property of unique maxima. Similarly, we can find the matrix P^* such that

$$\sum_{i,j} P^*_{ij}\mu_0(m_j|s_i) = \max_P \sum_{i,j} P_{ij}\mu_0(m_j|s_i),$$

where we maximize over all matrices P whose elements are non-negative and sum up to one within each column. The *best encoder*, P^*, is given by

$$P^*_{ij} = \begin{cases} 1, & \mu_0(m_j|s_i) = \max_p \mu_0(m_j|s_p), \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

i.e., we maximize each column of the matrix $\mu_0(m|s)$. Now, we have the best encoder and the best decoder for the language μ_0. Finding the matrices P^* and Q^* completes the task of the obverter of [23]. However, in our setting, the two matrices cannot be independent, but they need to be related by a common measure. If a measure μ_* existed such that

$$\mu_*(s|m) = P^*, \qquad \mu_*(m|s) = Q^*,$$

then it would satisfy Eq. (5), thus defining the best response. It turns out that in general, μ_* does not exist. However, there always exists a measure which approaches the performance of P^* and Q^* arbitrarily close. It is convenient to use the following short hand notation:

$$P^0_{ij} = \mu_0(s_i|m_j), \qquad Q^0_{ij} = \mu_0(m_j|s_i).$$

We are ready to formulate the following

Theorem 3.1. *For any finite language μ_0 satisfying the property of unique maxima, and a uniform probability distribution σ, we have*

$$\sup_\mu F(\mu_0, \mu) = 1/(2M)\,\mathrm{tr}\big(P^0(Q^*)^{\mathrm{T}} + P^*(Q^0)^{\mathrm{T}}\big).$$

In order to prove this statement, we need to show that

(a) for all μ,

$$F(\mu_0, \mu) \leqslant 1/(2M)\,\mathrm{tr}\big(P^0(Q^*)^{\mathrm{T}} + P^*(Q^0)^{\mathrm{T}}\big),$$

(b) there exists a family of languages, μ_ε, such that

$$\lim_{\varepsilon\to 0}\big|\sup_\mu F(\mu_0, \mu) - F(\mu_0, \mu_\varepsilon)\big| = 0.$$

The proof of (a) immediately follows from the definitions of the best decoder and the best encoder. The rest of this subsection is devoted to developing an algorithmic proof of (b). Given the matrices Q^* and P^*, we will build a family of measures, μ_ε, such that

$$\lim_{\varepsilon\to 0}\mu_\varepsilon(s|m) = P^*, \qquad \lim_{\varepsilon\to 0}\mu_\varepsilon(m|s) = Q^*. \tag{10}$$

This is not a trivial task, which is demonstrated by the following example. Suppose that the P^* and Q^* matrices are given by

$$P^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad Q^* = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It is clear that we cannot find a measure μ_ε which would satisfy conditions (10) for this pair (P^*, Q^*). Fortunately, it turns out that situations like this never arise. In order to prove this we will need to consider some auxiliary matrices.

3.1.1. The auxiliary matrix and the absence of loops

Let us define an auxiliary matrix X in the following way:

$$X_{ij} = \begin{cases} 1, & P^*_{ij} + Q^*_{ij} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This means that the matrix X contains nonzero entries at the slots where either of the matrices, P^* or Q^*, contains a non-zero entry. Now let us draw lines connecting all the "ones" of the X matrix that belong to the same row, and all the "ones" of the X matrix that belong to the same column. We will obtain some (disjoint) graphs. Let us refer to the "ones" of the X matrix as vertices.

Lemma 3.2. *Suppose that a finite measure μ_0 has the property of unique maxima. Graphs constructed as described above do not contain any closed loops.*

Proof. Let us assume that there exists a closed loop. It looks like a polygon with right angles. Let us consider its "turning points", i.e., such points which simultaneously belong to a horizontal and a vertical line. Suppose there are $2K$ such vertices (this can only be an even number). We will refer to these vertices as x_{α_i,β_j}, where the pair of integers, (α_i, β_j), gives the coordinates of the vertex. Clearly, $1 \leqslant i, j \leqslant K$.

Without loss of generality, let x_{α_1,β_1} be connected with x_{α_1,β_2} with a horizontal line. Then x_{α_1,β_2} is connected with x_{α_2,β_2} with a vertical line, …, x_{α_K,β_1} is connected with x_{α_1,β_1} with a vertical line, thus closing the loop (see Fig. 1, where we used $K = 3$). It is possible to show that exactly a half of the vertices corresponds to "ones" of the P^* matrix, and the rest—to "ones" of the Q^* matrix. If a vertex correspond to a "one" of the Q^* matrix then the corresponding slot of the P^* matrix is zero, and vice versa. This is a direct consequence of the property of unique maxima.

Let us now suppose that $Q^*_{\alpha_1,\beta_1} = 1$, $P^*_{\alpha_1,\beta_1} = 0$ (the alternative is that $P^*_{\alpha_1,\beta_1} = 1$, $Q^*_{\alpha_1,\beta_1} = 0$, in which case the proof remains very similar). This means that $Q^*_{\alpha_1,\beta_2} = 0$, because by construction (see (8)), there can be only one nonzero element in the same row of the Q^* matrix. Then the element $P^*_{\alpha_1,\beta_2} = 1$, because the corresponding vertex is present in the X matrix. This leads to $P^*_{\alpha_2,\beta_2} = 0$ (we can only have one positive element in each column of the P^* matrix, Eq. (9)). This argument can be continued around the loop. The Q^* elements along the loop are alternating between 0 and 1, and so are the elements of the P^* matrix, see Fig. 1.

We can conclude that $P^0_{\alpha_1,\beta_1} > P^0_{\alpha_1,\beta_2}$, because by construction, positive elements in the Q^* matrix correspond to the largest elements in the corresponding rows of the P^0 matrix. Similarly, we obtain $2K$ inequalities:

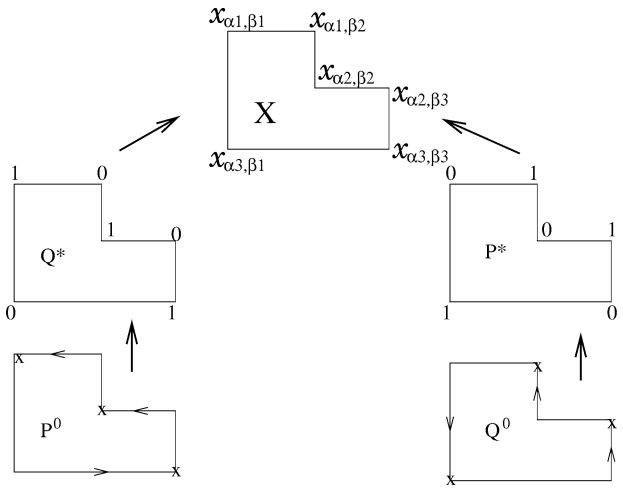


Fig. 1. No loops in graphs.

$$P^0_{\alpha_i,\beta_i} > P^0_{\alpha_i,\beta_{i+1}}, \tag{11}$$

$$Q^0_{\alpha_{i+1},\beta_{i+1}} > Q^0_{\alpha_i,\beta_{i+1}}, \quad 1 \leqslant i \leqslant K \tag{12}$$

(here we set $\alpha_{K+1} \equiv \alpha_1$ and $\beta_{K+1} \equiv \beta_1$). In Fig. 1, the maximum elements of the rows of P^0 and the columns of Q^0 are marked by crosses. The arrows indicate the direction toward the larger elements.

We will now show that system (11)–(12) is incompatible. In order to do this, we write $\mu_0(s_i, m_j) = Q^0_{ij} M_i$, where M_i is the sum of the elements of the ith row of the matrix μ_0: $M_i \equiv \sum_k \mu_0(s_i, m_k)$. Then we can rewrite P^0_{ij} in terms of Q^0 and M:

$$P^0_{ij} = \frac{\mu_0(s_i, m_j)}{\sum_k \mu_0(s_k, m_j)} = \frac{Q^0_{ij} M_i}{\sum_k Q^0_{kj} M_k}.$$

System (11)–(12) can be presented as a closed chain of inequalities for Q^0:

$$Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_1,\beta_2} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_2} M_k}, \quad Q^0_{\alpha_1,\beta_2} > Q^0_{\alpha_2,\beta_2}, \tag{13}$$

$$Q^0_{\alpha_2,\beta_2} > Q^0_{\alpha_2,\beta_3} \frac{\sum_k Q^0_{k\beta_2} M_k}{\sum_k Q^0_{k\beta_3} M_k}, \quad Q^0_{\alpha_2,\beta_3} > Q^0_{\alpha_3,\beta_3},$$

$$\ldots$$

$$Q^0_{\alpha_i,\beta_i} > Q^0_{\alpha_i,\beta_{i+1}} \frac{\sum_k Q^0_{k\beta_i} M_k}{\sum_k Q^0_{k\beta_{i+1}} M_k}, \quad Q^0_{\alpha_i,\beta_{i+1}} > Q^0_{\alpha_{i+1},\beta_{i+1}},$$

$$\ldots$$

$$Q^0_{\alpha_K,\beta_K} > Q^0_{\alpha_K,\beta_1} \frac{\sum_k Q^0_{k\beta_K} M_k}{\sum_k Q^0_{k\beta_1} M_k}, \quad Q^0_{\alpha_K\beta_1} > Q^0_{\alpha_1,\beta_1}. \tag{14}$$

From the first two inequalities we know that

$$Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_2,\beta_2} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_2} M_k},$$

then using the next pair we similarly derive that

$$Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_3,\beta_3} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_3} M_k}.$$

Continuing along the chain, at the Kth step we have

$$Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_K,\beta_K} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_K} M_k}.$$

Using the last two inequalities, we finally obtain: $Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_1,\beta_1}$. This contradiction proves that there can be no closed loops in the matrix X. □

3.1.2. Constructing the matrix μ_ε

Now we can systematically build the matrix μ_ε. From Lemma 3.2 it follows that if we connect all the vertices of the matrix X by horizontal and vertical lines, the resulting (disjoint) graphs will contain no closed loops. Some of the graphs might only consist of one vertex.

For each of these graphs we will perform the following procedure. Take a pair of vertices. If they are connected by a horizontal (vertical) line, refer to the corresponding entries of the Q^* matrix (P^* matrix). One of them will be one and the other—zero. Draw an arrow on the graph from the element corresponding to zero to the element corresponding to one. Repeat this for all pairs of vertices. Next, starting from some vertex, replace the corresponding element in the X matrix by ε, and then, following the arrows, keep replacing the elements of X by entries of the form ε^k, where the integer k increases or decreases from one vertex to the next depending on the direction of the arrow (we can always do this because by Lemma 3.2, there are not closed loops in the graphs of matrix X). We will call the resulting matrix A^ε. The measure μ_ε is obtained by re-normalizing the elements of the matrix A^ε:

$$\mu_\varepsilon(s_i, m_j) = A^\varepsilon_{ij} \Big/ \sum_{k,l} A^\varepsilon_{kl}. \tag{15}$$

Remark 3.3. In the algorithm above we used powers of the small parameter ε, ε^k, to assign to vertices of the matrix X. More generally, one can use *any* functions of ε, $f_k(\varepsilon)$, such that $\lim_{\varepsilon\to 0} f_k(\varepsilon)/f_{k+1}(\varepsilon) = 0$. Thus, the family μ_ε found above is just one of many such families.

3.1.3. Proof of Theorem 3.1

We are now ready to complete the proof of Theorem 3.1, part (b).

Proof. Let us show that Eq. (10) holds. In order to find entries of $\mu_\varepsilon(s|m)$, we need to re-normalize each column of the matrix μ_ε so that its elements sum up to one. Obviously, each

column will contain at most one segment of one of the graphs. By construction, the biggest element of this segment of the graph corresponds to the positive element of Q^*. In the limit $\varepsilon \to 0$, the other elements will be vanishingly small in comparison with the biggest one, and the resulting column of the $\mu_\varepsilon(s|m)$ matrix will be identical to the corresponding column of the P^* matrix. The same argument holds for rows of the $\mu_\varepsilon(m|s)$ matrix which in the limit become the rows of the Q^* matrix. Thus we conclude that the algorithm of Section 3.1.2 leads to constructing a family of measures μ_ε which satisfy the requirements of Theorem 3.1. □

Example 3.4. Consider the following 5×5 matrix:

$$\mu_0 = \frac{1}{1245}\begin{pmatrix} 1 & 64 & 2 & 23 & 90 \\ 92 & 8 & 42 & 81 & 42 \\ 53 & 77 & 60 & 2 & 50 \\ 88 & 15 & 68 & 73 & 59 \\ 39 & 48 & 66 & 65 & 37 \end{pmatrix}. \tag{16}$$

For this language, $\sup_\mu F(\mu_0, \mu) = 394/6225$. In Fig. 2 we show the calculated P^* and Q^* matrices, and then construct the X and the A^ε matrices. The family μ_ε is given by

$$\mu_\varepsilon = \frac{1}{3(1+\varepsilon)+\varepsilon^2}\begin{pmatrix} 0 & \varepsilon & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & \varepsilon^2 & 0 & 0 & 0 \\ \varepsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \varepsilon & 0 \end{pmatrix}.$$

As $\varepsilon \to 0$, $F(\mu_0, \mu_\varepsilon) \to \sup_\mu F(\mu_0, \mu)$.

Remark 3.5. If we let $\mu_* = \mu_\varepsilon|_{\varepsilon=0}$, i.e., μ_ε evaluated at 0, we note that $\mu_* \neq \mu_0$ in general. Further, $F(\mu_0, \mu_*) < \sup_\mu F(\mu_0, \mu)$. Thus, we have that $\lim_{\varepsilon\to 0} \mu_\varepsilon = \mu_*$ yet $F(\mu_0, \mu_*) < \lim_{\varepsilon\to 0} F(\mu_0, \mu_\varepsilon) = \sup_\mu F(\mu_0, \mu)$. This is a consequence of a discontinuity in the definition of the communicability function, $F(L_1, L_2)$. Namely, the conditional probabilities entering definition (4) are discontinuous when all the elements of a column or a row of μ are zero, see Eqs. (1)–(2). Thus the value of $F(\mu_0, \mu_\varepsilon)$ may have a jump at $\varepsilon = 0$.

3.2. General languages

Now we will demonstrate the effect of relaxing assumptions (i) through (iii) of Section 3.1.

3.2.1. Multiple maxima and neutral vertices

If condition (iii) of the previous section is not satisfied, that is the language μ_0 does not possess the property of unique maxima, then definitions (8) and (9) have to be changed. For instance, if $\mu_0(s_k|m_{\alpha_1}) = \cdots = \mu_0(s_k|m_{\alpha_n})$ are all maximal values of the kth row of matrix $\mu_0(s|m)$, then we can take

$$Q^*_{\alpha_1,k} = \gamma_1, \ldots, Q^*_{\alpha_n,k} = \gamma_n,$$

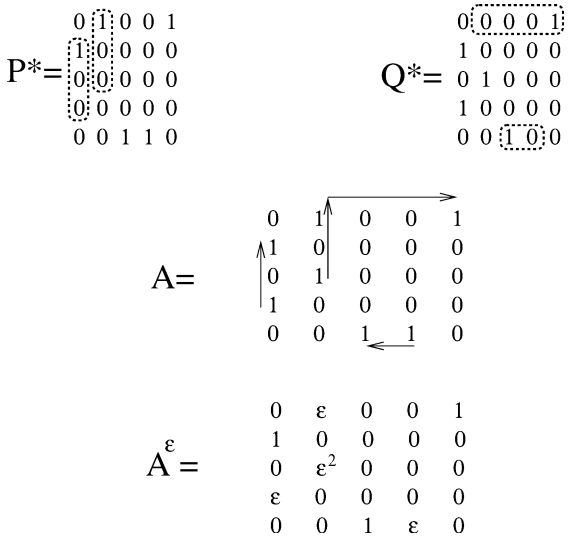


Fig. 2. Construction of A^{ε} for Example 3.4. We first form P^0 and Q^0 matrices by normalizing columns and rows of μ_0 respectively; this step is not shown here. Then we can construct the best encoder, P^*, by identifying the maximal elements in the columns of Q^0, and the best decoder, Q^*, by identifying the maximal elements in the rows of P^0, see the top of the figure. Next, we combine the positive elements (or vertices) of P^* and Q^* to create the auxiliary matrix X. The vertices of X that belong to the same column (row) are connected. In order to define the direction of the arrows, we have to refer to the matrices P^* and Q^*. If two vertices are connected by a vertical line, we find the corresponding elements of the P^* matrix (they are encircled); the direction of the arrow is always toward the "one" of the P^* matrix. Similarly, if two vertices are connected by a horizontal line, we find the corresponding elements of the Q^* matrix (encircled) and direct the arrow toward the "one" of the Q^* matrix. Finally, we build the A^{ε} matrix by replacing the "ones" of the X matrix by powers of ε. The powers of ε must be arranged in such a way that in each of the connected graphs, the arrows point from a smaller entry to a larger entry. Note that in this example P^* and Q^* are not compatible with any single measure.

so that $\gamma_1, \ldots, \gamma_n$ are arbitrary positive numbers with the only restriction that $\sum_{i=1}^{n} \gamma_i = 1$. The result of evaluating the function $\sum_{i,j} \mu_0(s_i|m_j) Q^*_{ij}$ (Eq. (7)) does not depend on the values of the coefficients γ_i. The same argument can be repeated for P^*. Next we note that some closed loops are possible in this case, so that Lemma 3.2 has to be modified. Let us generalize the procedure of assigning direction to the graphs in the case where the language μ_0 does not possess the property of unique maxima. We will not assign a direction to segments of the graph corresponding to rows of P^0 (columns of Q^0) which do not have a unique maximum. We will call the corresponding vertices *neutral vertices*. For vertices which are not neutral, we proceed as before, i.e., if two vertices are connected with a horizontal (vertical) line, then the arrow points toward the larger element of the Q^* (P^*) matrix.

For a closed loop of the auxiliary matrix X, we define the direction of each segment as positive (negative) if it is clockwise (counterclockwise). The direction is zero if there is no arrow. We say that the direction changes sign if it changes from positive to negative or from negative to positive. Instead of Lemma 3.2 we have

Lemma 3.6. *The following is true for loops of the auxiliary matrix X*:

(a) *they contain more than one neutral vertex.*
(b) *the direction of the graph changes sign at least once, or it is identically zero.*

Proof. Statement (a) can be proved by assuming that there are no neutral vertices in a loop and applying Lemma 3.2. To prove statement (b) let us assume that the direction of the arrows in a loop is always positive or neutral. Then we can repeat the argument of Lemma 3.2 and write down a chain of equations/inequalities similar to (13)–(14). The only difference is that some of the inequalities will in fact have an "equals" sign. More precisely, a segment with a positive (zero) direction will correspond to a "$<$" ("$=$"). We immediately get a contradiction unless all signs are "$=$", or the strict inequalities change direction at least once. This proves statement (b). □

The statement of Theorem 3.1 still holds in this case if the perfect encoder and decoder are redefined as indicated above. The algorithm of building the "best response" language stays very similar. We assign powers of ε to all nodes so that the power decreases in the direction of arrows. For adjacent neutral nodes, the power of ε must be the same, and some arbitrary weights can be assigned to the neutral nodes. If a loop is present, it is still possible to assign powers of ε in a consistent way because of statement (b) of Lemma 3.6. It may be necessary to use non-integer powers.

Example 3.7. Consider the following 3×3 matrix:

$$\mu_0 = \frac{1}{43}\begin{pmatrix} 8 & 5 & 2 \\ 2 & 10 & 2 \\ 3 & 9 & 2 \end{pmatrix}. \tag{17}$$

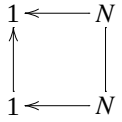
It is easy to calculate the P^0 and the Q^0 matrices:

$$P^0 = \begin{pmatrix} 8/13 & 5/24 & 1/3 \\ 2/13 & 5/12 & 1/3 \\ 3/13 & 3/8 & 1/3 \end{pmatrix}, \qquad Q^0 = \begin{pmatrix} 8/15 & 1/3 & 2/15 \\ 1/7 & 5/7 & 1/7 \\ 3/14 & 9/14 & 1/7 \end{pmatrix}. \tag{18}$$

We can see that this language does not possess the property of unique maxima: the third column of Q^0 contains two maximal elements. We have

$$P^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \gamma \\ 0 & 0 & 1-\gamma \end{pmatrix}, \qquad Q^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}. \tag{19}$$

The directed graph contains one loop with two neutral vertices marked by "N":



The graph changes direction once along the loop. Applying our algorithm we obtain the following A^ε matrix:

$$A^\varepsilon = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \gamma_1\varepsilon \\ 0 & \sqrt{\varepsilon} & \gamma_2\varepsilon \end{pmatrix}.$$

For any positive numbers $\gamma_{1,2}$ this family satisfies the conditions of Theorem 3.1, i.e., approaches the best communicability. Note that if $\gamma_1/\gamma_2 = \gamma$ then in the limit of $\varepsilon \to 0$, the matrices P^* and Q^* are recovered, see Eqs. (19).

Note that in general it is not always possible to find an A^ε matrix which would give rise to given P^* and Q^*; some conditions on the arbitrary neutral coefficients in P^* and Q^* matrices may be imposed (see also Lemma 5.3).

3.2.2. Non-uniform distributions

Before we only considered the uniform distributions $\sigma_i = 1/M$ (condition (ii) above). Now let us assume some general distribution. It turns out that the argument changes very little. Namely, definition (8) becomes

$$Q^*_{ij} = \begin{cases} 1, & \mu_0(s_i|m_j)\sigma_j = \max_p \mu_0(s_i|m_p)\sigma_p, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

(and similarly for the case when we do not have the unique maxima property), and definition (9) stays the same. In the proof of Lemma 3.2, the argument follows the same logics, and the only change comes into inequalities (11): we have

$$P_{\alpha_i,\beta_i}\sigma_i > P_{\alpha_i,\beta_{i+1}}\sigma_{i+1}. \tag{21}$$

However, the multipliers σ_l also get canceled out when we go around the loop, so the statements of Lemmas 3.2 and 3.6 remain true in this case, and thus the algorithm of building the best response is the same.

3.2.3. Infinite matrices

Finally, we will deal with restriction (i) of Section 3.1. First of all let us show that definition (4) makes sense in the case of infinite matrices. We have

$$\begin{aligned} &\tfrac{1}{2}\sum_i \big(\mu_0(s_i|m_j)\mu(m_j|s_i) + \mu(s_i|m_j)\mu_0(m_j|s_i)\big) \\ &\quad\leqslant \tfrac{1}{2}\sum_i \big(\mu_0(s_i|m_j) + \mu(s_i|m_j)\big) = 1. \end{aligned}$$

Since σ is a measure, we have $\sum_j \sigma_j = 1$, which leads to the conclusion $F(L_0, L) \leqslant 1$.

Now, us define the following quantities:

$$A_i = \sup_j P^0_{ij}\sigma_j, \qquad B_i = \sup_j Q^0_{ji}.$$

The generalization of Theorem 3.1 in the case of infinite matrices is given by

Theorem 3.8. *For infinite matrices,* $\sup_\mu F(\mu_0, \mu) = \frac{1}{2}\sum_i (A_i + B_i)$.

It is straightforward to see that for any μ, $F(\mu_0, \mu) \leqslant \frac{1}{2}\sum_i (A_i + B_i)$. In order to conclude the proof of Theorem 3.8, it is necessary to construct a family of languages, μ_ε, such that $\lim_{\varepsilon\to 0} |\sup_\mu F(\mu_0, \mu) - F(\mu_0, \mu_\varepsilon)| = 0$. This is done in Appendix A.

3.2.4. The existence of the best response

To end this section, we will address one more issue. From Theorem 3.1 and extensions it follows that there exists a family of languages, μ_ε, such that $\lim_{\varepsilon\to\infty} F(\mu_0, \mu_\varepsilon) = \sum_\mu F(\mu_0, \mu)$. Can it happen that the supremum is actually reached by some language? From previous considerations it is clear that a language μ_* satisfies Eq. (5) if and only if it also obeys

$$\mu_*(s|m) = P^*, \qquad \mu_*(m|s) = Q^*$$

(where P^* and Q^* may not be unique, like in the case where the property of unique maxima is not satisfied). The question of existence of μ_* is answered by the following

Theorem 3.9. *For a language μ_0, the limiting measure μ_* exists if and only if the auxiliary matrix X satisfies the following property*: *the only vertices of the matrix X that share the same row* (*column*) *are neutral vertices.*

Proof. The "if" part is easy: given the condition of the theorem, we can apply the algorithm of building the family μ_ε (Section 3.1.2 and its extension from Section 3.2.1) and observe that the powers of ε simply get canceled when we normalize the matrix μ_ε. This means that μ_ε does not depend on ε, and since we know that it satisfies the conditions of Theorem 3.1, we conclude that $\mu_\varepsilon = \mu_*$.

To finish the proof we need to show that if the condition of the theorem is not satisfied, then μ_* does not exist. Let us assume that there are two adjacent (non-neutral) vertices, a and b, in the matrix X. Without loss of generality let us assume that they are connected by a horizontal line. This means that Q^* has a 1 at the one of the vertices, say vertex a, and a 0 at vertex b. Therefore, if μ_* exists then it must have a positive entry at a. On the other hand, P^* has a 1 at vertex b and a zero at vertex a. Therefore, the matrix μ_* must have a zero at vertex a, which leads to a contradiction. □

Corollary 3.10. *If $P^* = Q^*$ are extended permutation matrices, i.e., square permutation matrices perhaps with extra rows or columns consisting entirely of zeros, then μ_* is defined as P^*, properly normalized.*

Corollary 3.11. *If the property of unique maxima is not satisfied and μ_* exists, it is* not *unique.*

4. Implications for learning

From the preceding discussion it is now clear that in order to maximize mutual intelligibility with a language user (characterized by the measure μ), it may be necessary

to use a different measure, μ_*, where $\mu_* \neq \mu$. This fact has implications both for learning and evolution of populations of linguistic agents.

Let us first consider the problem of an agent trying to learn a language in order to communicate with some other agent whose language is characterized by the measure μ. Recall that μ_* (the best response) itself may not exist, however, an arbitrarily close approximation μ_ε (for any ε) does exist. Therefore, the learner's task becomes to estimate μ_ε. What degree of accuracy, ε, is useful or necessary will depend upon the particular application in mind. Since the measure μ is unknown to the learner at the outset, there are two natural learning scenarios depending upon how much information is available to the learner on each interaction.

(1) *Full information*: This corresponds to the situation where the learner is able to sample μ directly to get (sentence, meaning) pairs. Thus, when the teacher speaks, both sentence and meaning are directly accessible. The strategy of the learner is to estimate μ as well as it can and derive from it the P^* and Q^* matrices and ultimately μ_ε using the procedure described in the previous sections.
(2) *Partial information*: In most natural settings, however, the meaning may not be directly accessible. In other words, the learner only hears the sentence while the intended meaning is latent. What the learner reasonably may have access to is whether its interpretation of the sentence was successful or not. On the basis of this information, the learner must somehow derive the optimal communication strategy. We refer to this as learning with partial information. Note that we assume that the learner (hearer) receives *weak* reinforcement regarding the communicative exchange. This is similar to the setting of selfish games developed in the work of Steels and pursued also in [40] (always through simulations). There are variants where the learner could get *strong* reinforcement either in the extreme form of being told the true meaning after a failed communicative exchange or some alternative corrective feedback. We do not explore the strong reinforcement setting here.

Thus we see that (1) full information and (2) partial information suggest two different frameworks for learning; in either case, the learner has to estimate P and Q matrices of the teacher.

4.1. Estimating P

An important task for the learner is to estimate Q^* which is derivable from the P matrix of the teacher. Recall that

$$Q^*_{ij} = \begin{cases} 1, & \sigma_j \mu(s_i|m_j) = \max_p \sigma_p \mu(s_i|m_p), \\ 0, & \text{otherwise.} \end{cases}$$

4.1.1. Learning with full information

The learner, in this case, has access to (s, m) (sentence,meaning) pairs every time the teacher produces a sentence. We can define the event

$$A_{ij} = \text{Teacher produces } s_i \text{ to communicate } m_j.$$

The probability of event A_{ij} is simply $\sigma_j \mu(s_i|m_j)$. Therefore, if the teacher produces n (sentence, meaning) pairs (which are random, independent and identically distributed), then the ratio

$$\hat{a}_{ij}(n) = \frac{k_{ij}}{n}$$

is an empirical estimate of the probability of the event A_{ij}. By the law of large numbers, as $n \to \infty$ we have

$$\hat{a}_{ij}(n) \to \sigma_j \mu(s_i|m_j)$$

with probability 1. For the case under consideration, we can even bound the rate at which this convergence occurs. For example, applying Hoeffding's inequality, we have

$$P\left[\left|\hat{a}_{ij}(n) - \sigma_j \mu(s_i|m_j)\right| > \varepsilon\right] \leqslant 2\,\mathrm{e}^{-\varepsilon^2 n/2}.$$

This convergence is guaranteed for fixed (i, j). In general, the learner must estimate a collection of events. The total number of events are given by the total number of possible (sentence, meaning) pairs. As before, let us assume that there are N possible sentences and M possible meanings. Therefore, there are NM different events whose probabilities need to be estimated. The collection of events A_{ij}, $i = 1, \ldots, N$; $j = 1, \ldots, M$, are disjoint. For a finite collection of such events, we will derive a uniform law of large numbers.

Let event E_{ij} be

$$E_{ij} = \left|\hat{a}_{ij}(n) - \sigma_j \mu(s_i|m_j)\right| > \varepsilon.$$

Then, by the union bound, we obtain

$$P\left[\bigcup_{i,j} E_{ij}\right] \leqslant \sum_{i,j} P(E_{ij}) \leqslant NM2\,\mathrm{e}^{-\varepsilon^2 n/2}.$$

Therefore, we have

$$P\left[\overline{\bigcup_{i,j} E_{ij}}\right] = P\left[\forall i, j\ \left|\hat{a}_{ij}(n) - \sigma_j \mu(s_i|m_j)\right| \leqslant \varepsilon\right] > 1 - NM2\,\mathrm{e}^{-\varepsilon^2 n/2}.$$

Thus, with high probability (depending upon the number of examples, n) all empirical estimates $\hat{a}_{ij}(n)$ are close to $\sigma_j \mu(s_i|m_j)$, respectively. Estimating the $\sigma_j \mu(s_i|m_j)$'s is the step to estimating the Q^* matrix that is required for the optimal communication system.

4.1.2. Learning with partial information

Now consider the setup in (2) where the learner has no access to the meaning directly but has to guess a meaning and is told after the event whether the guess was correct or incorrect. Thus the learner has access to asymmetric information: if the guess was correct, the learner knows the true intended meaning; if the guess was incorrect, the learner merely knows what the meaning was not. As it turns out, this does not dramatically change the state of affairs. To see this, let the learner guess a meaning uniformly at random. Thus with probability $1/M$ the learner chooses a meaning m_j. Each time the teacher produces

a sentence, the intended meaning may be successfully communicated or not. Define the event

$$A_{ij} = \text{Teacher produces } s_i;\ \text{Learner guesses } m_j;\ \text{Communication is successful.}$$

The probability of event A_{ij} is simply $\frac{1}{M}\sigma_j\mu(s_i|m_j)$. The event A_{ij} is observable since the learner knows (i) what sentence has been uttered by the teacher, (ii) what meaning it (the learner) assigned to the sentence, and (iii) whether communication was successful. Therefore after n sentences have been produced by the teacher, the learner can count k_{ij}—the number of times event A_{ij} has occurred, and can make an empirical estimate of the probability of A_{ij} as

$$\hat{a}_{ij}(n) = \frac{k_{ij}}{n}.$$

By the same argument as before, $\hat{a}_{ij}(n)$ converges in probability to $\frac{1}{M}\sigma_j\mu(s_i|m_j)$ and the rates are provided by the Hoeffding bounds. Since M is fixed in advance and known, this allows the learner to guess $\sigma_j\mu(s_i|m_j)$ for each i, j arbitrarily well. Let us be a little more precise about the rates of convergence. The learner's estimate of $\sigma_j\mu(s_i|m_j)$ is really $M\hat{a}_{ij}$ where $\hat{a}_{ij}$ is defined above. Therefore we have that

$$P\left[\left|M\hat{a}_{ij} - \sigma_j\mu(s_i|m_j)\right| > \varepsilon\right] = P\left[\left|\hat{a}_{ij} - \frac{1}{M}\sigma_j\mu(s_i|m_j)\right| > \frac{\varepsilon}{M}\right] \leqslant 2\,\mathrm{e}^{-\varepsilon^2 n/(2M^2)}.$$

Thus the confidence in the ε-good estimate of $\sigma_j\mu(s_i|m_j)$ is poorer than before. By the same argument as in case (2), we have a uniform bound as follows:

$$P\left[\forall i, j\ \left|M\hat{a}_{ij} - \sigma_j\mu(s_i|m_j)\right| \leqslant \varepsilon\right] > 1 - NM2\,\mathrm{e}^{-\varepsilon^2 n/(2M^2)}. \tag{22}$$

4.2. Estimating Q

Let us now consider the task of estimating P^* which is derivable from the Q matrix of the teacher. The same arguments of the previous section apply. Recall that

$$P^*_{ij} = \begin{cases} 1, & \mu(m_j|s_i) = \max_p \mu(m_j|s_p), \\ 0, & \text{otherwise.} \end{cases}$$

4.2.1. Learning with full information

Here the learner has direct access to the meaning assigned by the teacher to each sentence. Therefore, the learner need only pick a sentence uniformly at random (with probability $1/N$) and produce it for the teacher to hear. Let us define the event

$$A_{ij} = \text{Learner produces } s_i;\ \text{Teacher interprets as } m_j.$$

The event A_{ij} is observable on each trial. The probability with which it occurs is given by $\frac{1}{N}\mu(m_j|s_i)$. After n trials (where the learner speaks in this manner), the learner simply counts the number k_{ij} of times event A_{ij} occurs and its estimate of $\frac{1}{N}\mu(m_j|s_i)$ is k_{ij}/n. Therefore, we have

$$P\left[\left|\hat{a}_{ij} - \frac{1}{N}\mu(m_j|s_i)\right| > \varepsilon\right] \leqslant 2\,\mathrm{e}^{-\varepsilon^2 n/(2N^2)}.$$

Using the same arguments as before, we have

$$P\left[\forall i, j \left| MN\hat{a}_{ij} - \mu(m_j|s_i)\right| \leqslant \varepsilon\right] > 1 - NM2\,\mathrm{e}^{-\varepsilon^2 n/(2N^2)}.$$

4.2.2. Learning with partial information

The learner simply picks a (sentence, meaning) pair uniformly at random (with probability $1/(NM)$). Define the event

$$A_{ij} = \text{Learner produces } (s_i, m_j);\ \text{Communication is successful.}$$

The event A_{ij} is observable by the learner on each trial. The probability of event A_{ij} is $\frac{1}{NM}\mu(m_j|s_i)$. After n trials (where the learner speaks), the learner counts the number k_{ij} of times event A_{ij} occurs. Therefore, we again have

$$P\left[\left|\hat{a}_{ij} - \frac{1}{NM}\mu(m_j|s_i)\right| > \varepsilon\right] \leqslant 2\,\mathrm{e}^{-\varepsilon^2 n/(2M^2N^2)}.$$

Using the same arguments as before, we have

$$P\left[\forall i, j \left| MN\hat{a}_{ij} - \mu(m_j|s_i)\right| \leqslant \varepsilon\right] > 1 - NM2\,\mathrm{e}^{-\varepsilon^2 n/(2M^2N^2)}. \tag{23}$$

4.3. Sample complexity bounds

Now we can put the pieces together to determine the number of learning events that need to occur so that with high probability, the learner will be able to develop a language with ε-good communicability. Let the teacher's measure be μ. We will assume that μ is such that the P and Q matrices have unique row-wise and column-wise maxima respectively. First let us introduce the *margin* by which the maximum value clears all other values in the row and column respectively. This *margin* will play an important role in determining the number of learning events.

Definition 1. For each i, let

$$j_i^* = \arg\max_j \sigma_j \mu(s_i|m_j)$$

and for each j, let

$$i_j^* = \arg\max_i \mu(m_j|s_i).$$

Then, we define the *margin* γ to be the largest real number such that

$$\sigma_{j_i^*}\mu(s_i|m_{j_i^*}) \geqslant \sigma_j \mu(s_i|m_j) + \gamma \quad \forall j \neq j_i^*$$

and

$$\mu(m_j|s_{i_j^*}) \geqslant \mu(m_j|s_i) + \gamma \quad \forall i \neq i_j^*.$$

4.3.1. Learning with partial information

We have described how to estimate the Q^* and P^* matrices; the following theorem provides a bound on the number of examples needed to ensure correct estimates:

Theorem 4.1. *If the total number, n, of interactions between teacher and learner (with partial information) is greater than $(64M^2N^2/\gamma^2)\log(4MN/\delta)$, then with high probability $> 1-\delta$, the learner can construct a measure that will give arbitrarily good communicability with the teacher.*

Proof. Let there be $n/2$ interactions where the teacher speaks and the learner listens and $n/2$ interactions of the other form. The learner constructs estimates of $\sigma_j\mu(s_i|m_j)$ and $\mu(m_j|s_i)$ in the manner described previously. Let the estimates be denoted by $\hat{p}_{ij}$ and $\hat{q}_{ij}$, respectively. By setting $\varepsilon = \gamma/4$ in Eqs. (22) and (23), we obtain:

$$P\big[\forall i, j \ \big|\hat{p}_{ij} - \sigma_j\mu(s_i|m_j)\big| \leqslant \gamma/4\big] > 1 - 2NM\,\mathrm{e}^{-\gamma^2 n/(64M^2)}$$

and

$$P\big[\forall i, j \ \big|\hat{q}_{ij} - \mu(m_j|s_i)\big| \leqslant \gamma/4\big] > 1 - 2NM\,\mathrm{e}^{-\gamma^2 n/(64M^2N^2)}.$$

Using the fact that $P(A\cap B) \geqslant P(A) + P(B) - 1$, we can see that with probability greater than $1 - 2NM(\mathrm{e}^{-\gamma^2 n/(64M^2N^2)} + \mathrm{e}^{-\gamma^2 n/(64M^2)})$, the estimates $\hat{p}_{ij}$ and $\hat{q}_{ij}$ are both within $\gamma/4$ of the true values. The learner chooses Q^* and P^* using the estimated matrices. Let us first consider the case of Q^*. For each i the learner desires to obtain j_i^* given by

$$j_i^* = \arg\max_j \sigma_j\mu(s_i|m_j).$$

The learner chooses

$$\hat{j}_i = \arg\max_j \hat{p}_{ij},$$

and we claim that $\hat{j}_i = j_i^*$. In order to prove this, assume that this is not the case. Then we get immediately:

$$\sigma_{j_i^*}\mu(s_i|m_{j_i^*}) \geqslant \sigma_{\hat{j}_i}\mu(s_i|m_{\hat{j}_i}) + \gamma.$$

However, we have the following chain of inequalities:

$$\sigma_{\hat{j}_i}\mu(s_i|m_{\hat{j}_i}) \geqslant \hat{p}_{i\hat{j}_i} - \gamma/4 \geqslant \hat{p}_{ij_i^*} - \gamma/4 \geqslant \sigma_{j_i^*}\mu(s_i|m_{ij_i^*}) - \gamma/2,$$

which leads to a contradiction. This argument holds for every i, therefore, since $\hat{j}_i = j_i^*$ for each i, the Q^* matrix is identified exactly. Similarly, one can show that the P^* matrix is also identified exactly.

The only thing that remains is to ensure that n is large enough so that this occurs with high probability. We have

$$2NM\big(\mathrm{e}^{-\gamma^2 n/(64M^2N^2)} + \mathrm{e}^{-\gamma^2 n/(64M^2)}\big) \leqslant 4NM\,\mathrm{e}^{-\gamma^2 n/(64M^2N^2)} \leqslant \delta.$$

This is satisfied for $n > (64M^2N^2/\gamma^2)\log(4MN/\delta)$. Thus, with probability greater than $1-\delta$, both P^* and Q^* are identified exactly. Now the procedure of approximating the measure may be applied. □

Remarks.

1. The number of examples is seen to be a function of M, N and γ. The margin γ that depends upon the teacher's language, μ, determines, in some sense, how easy it is to estimate Q^* and P^* matrices for the learner. It therefore characterizes the learning difficulty of μ in this setting.
2. Finite matrices are applicable to settings such as alarm calls in animal communication systems and lexical learning in human linguistic systems. For example, [27] discusses the problem of learning mappings between signals and meanings using a variety of schemes from associative learning to Bayesian estimation.
3. Infinite matrices are not learnable in general. In fact, infinite dimensional spaces are known to be unlearnable (see [39]) and therefore further constraints will be required on the space of possible measures to which the teacher's language belongs. There are several ways in which one could explore reasonable constraints on linguistic measures. One possibility might be to pursue the approach of Chomsky [5] and restrict the range of variation on possible syntactic forms and thereby on possible measures μ. Another possibility might be to pursue some theory of compositional semantics (e.g., [6]) where the meanings of larger units like phrases and sentences are derived from compositional rules applied to the meanings of smaller units like morphemes and words. Thus the true learning task is really to learn the meanings of words appropriately and then apply these compositional rules for all other syntactic forms. Since words are finite, this reduces the infinite case to the finite one. A third possibility is to make use of context heavily and claim that learning proceeds in a context by context (case based) fashion and in each particular context there are only a finite number of possible forms and their interpretations. A proper development of these issues is the subject of further research and beyond the scope of the current paper.
4. The constants in the bound on sample complexity may be tightened, although the order is essentially correct. For example, we have let the interactions be symmetric, i.e., the numbers of sentences the learner produces and receives are the same. It is easy to check that a more favorable bound is obtained when the learner speaks N^2 times as often as it listens. In this case, it is enough to have $(32M^2(N^2+1)/\gamma^2)\log(4MN/\delta)$ interactions in all.

4.3.2. Learning with full information

For completeness, let us state the number of interactions needed to learn in setting (1). This is given by the following

Theorem 4.2. *If the total number, n, of interactions between teacher and learner* (*with full information*) *is greater than* $(64N^2/\gamma^2)\log(4MN/\delta)$, *then with high probability* $> 1-\delta$, *the learner can construct a measure that will give arbitrarily good communicability with the teacher.*

The proof is very similar to that of the previous theorem and we omit it for this reason. It is noteworthy that learning with full information requires M^2 fewer interactions to learn. This is not surprising since the meanings are accessible, and the larger is the number, M,

of different concepts, the greater is the difference between learning with full and partial information.

5. Implications for evolution

In this section we address the question of evolutionary significance of communicability. This has application in several different contexts.

First, in artificial intelligence, one way to create communicating machines is to start with a population of agents with a sub-optimal communication system and let them evolve and learn from each other. This general approach is pursued in various forms by researchers in evolutionary computation, genetic algorithms, and artificial life. Since the goal is to increase information transfer, the function F can conveniently play a role of the "score" of different communication systems. Based on this, agents with higher intelligibility can be arranged to proceed while agents with lower intelligibility score will be gradually eliminated. The main question is whether such a process will ultimately lead to a coherent communication system. In what follows we develop a formalism that will allow one to characterize possible outcomes of such a dynamical system.

Second, game-theoretic approaches may be relevant to the biological evolution of simple, innate signaling systems in the animal world. In this setting, the function F contributes to the biological fitness of individuals. The framework developed here has some obvious drawbacks, such as the assumption that the signaler and the receiver are both equally interested in the successful transfer of the information, which may not necessarily be the case in many natural settings. However, studying basic properties of evolutionarily stable states of the simplest system may explain certain aspects of evolution, and this approach can later be extended to include more sophisticated scenarios, such as clashes of interest of signalers and receivers.

Finally, in application to human languages, evolutionary theory can also potentially make a contribution. Here, it is not the innate genetic endowment that is considered to be under the selective pressure, but the (learned and culturally transmitted) languages themselves [13,25]. For natural selection to act on language ability, there must be a reward for successful communication, which links language to biological *fitness*. We can therefore assume that successful communication leads to a *payoff* for both the speaker and the hearer. In the spirit of evolutionary game theory, we link payoff to reproductive success [9, 17]. Individuals that communicate more successfully have increased survival probabilities and leave more offspring. An alternative interpretation of this kind of dynamics, is that such individuals will acquire a higher standing or reputation in the group and leave more followers who will learn their language. This simplified model, while leaving out many potentially important aspects of the system, serves as a logical tool of reasoning about the evolution of human language. The inferred properties of the evolutionarily stable states can be compared with the observed properties of human languages. The results of such a comparison may shed light on the role of communicability in the evolution of human language.

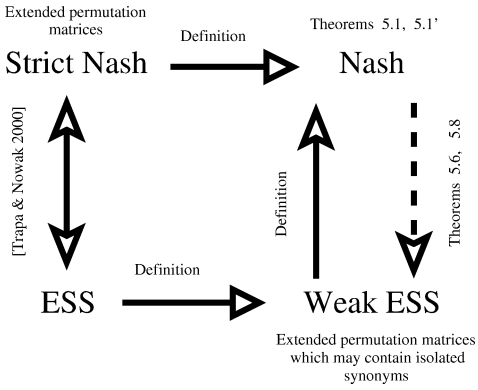


Fig. 3. Nash equilibria and ESS.

5.1. Preliminaries

Let us characterize the attracting states of the evolutionary system by means of the payoff function, $F(\mu, \mu')$. It is useful to recall some important definitions of the classical game theory [9]. Language μ is *strict Nash equilibrium* if we have $F(\mu, \mu) > F(\mu, \mu')$ for all $\mu' \neq \mu$. It is *Nash equilibrium* if $F(\mu, \mu) \geqslant F(\mu, \mu')$ for all $\mu' \neq \mu$.

Language μ is called an *evolutionarily stable state* (*ESS*) [18] if μ is Nash and for every μ' with $F(\mu, \mu) = F(\mu, \mu')$ we have $F(\mu, \mu) > F(\mu', \mu')$. Language μ is a *weak ESS* if the final strict inequality is relaxed to a weak one, $F(\mu, \mu) \geqslant F(\mu', \mu')$.

It can be shown (see [38]) that a language is an ESS if and only if it is a strict Nash equilibrium, see Fig. 3. It is clear that μ' is a strict Nash equilibrium iff there exists a unique best response which is equal to μ'. From the algorithm of finding the best response given in this paper it follows that strict Nash equilibria have to be square matrices and are given by permutation languages, which is in accordance with [38]. In the presence of a non-trivial transition matrix (the noisy environment), we can derive the following result: permutation languages are strict Nash equilibria if the T matrix is diagonally dominant both column-wise and row-wise, see Appendix B. (It is interesting that if these conditions on the T matrix are not satisfied then permutation languages are no longer stable! However, such situations correspond to the kind of noise which changes the signals beyond recognition and can hardly be considered relevant in natural settings.) These results indicate that perfectly coordinated systems with no homonyms or synonyms are evolutionarily stable.

Strict Nash equilibria do not exhaust the set of important rest points of the system. In order to get the full picture we also need to characterize the weak ESS of the system. Once the system has reached one of the weak ESS, random drift is possible without change in the average communicability.

5.2. Nash equilibria

The classification of Nash languages for uniform σ distributions was found in [38]; we will reproduce their result because we will need to use it later:

Theorem 5.1 [38]. *For a uniform probability distribution, σ, a language is Nash if the supports of its P and Q matrices coincide and if each row (column) of its P (Q) matrix contains at most two distinct values, one of which is zero.*[3]

The case of general distributions will be considered in Section 5.4. An example of a Nash language is given by

Example 5.2. Consider the language with

$$P'' = \begin{pmatrix} a_1 & 0 & a_1 & a_1 & 0 & a_1 & 0 & 0 \\ a_2 & 0 & 0 & a_2 & a_2 & 0 & a_2 & 0 \\ 0 & a_3 & 0 & 0 & a_3 & a_3 & a_3 & 0 \\ 0 & 0 & a_4 & 0 & 0 & 0 & 0 & a_4 \\ 0 & a_5 & 0 & 0 & 0 & 0 & 0 & a_5 \end{pmatrix},$$

$$Q'' = \begin{pmatrix} c_1 & 0 & c_3 & c_4 & 0 & c_6 & 0 & 0 \\ c_1 & 0 & 0 & c_4 & c_5 & 0 & c_7 & 0 \\ 0 & c_2 & 0 & 0 & c_5 & c_6 & c_7 & 0 \\ 0 & 0 & c_3 & 0 & 0 & 0 & 0 & c_8 \\ 0 & c_2 & 0 & 0 & 0 & 0 & 0 & c_8 \end{pmatrix}.$$

The conditions $\sum_{i=1}^{N} P''_{ij} = 1$ and $\sum_{j=1}^{M} Q''_{ij} = 1$ lead to

$$a_i = \tfrac{1}{2}, \qquad c_1 = \tfrac{1}{2} - c_4, \qquad c_2 = c_3 = 1 - c_8, \qquad c_5 = \tfrac{1}{2} - c_6,$$
$$c_6 = c_8 - \tfrac{1}{2}. \tag{24}$$

This language satisfies the conditions of Theorem 5.1 and therefore, it is a Nash equilibrium.

We need to check whether the P and Q matrices of Nash equilibria are related through a common measure. We have the following useful

Lemma 5.3. *The P and Q matrices of Nash languages of Theorem* 5.1 *always correspond to a common measure μ.*

Proof. We will simply construct the measure μ. Let us form the diagonal $M \times M$ matrix D^Q such that D^Q_{ii} is the value of the non-zero elements of the ith column of Q. Similarly, for the diagonal $N \times N$ matrix D^P, the element D^P_{ii} is the value of the non-zero elements of the ith row of P. We have $P = XD^P$ and $Q = D^Q X$, where X is the auxiliary matrix of P and Q, i.e., it has ones on the support of Q and P and zeros everywhere else. Let us define

$$A = D^P X D^Q. \tag{25}$$

[3] These conditions can be derived by methods of Section 3.2.1 using the arbitrariness in the algorithm in the presence of neutral vertices.

It is easy to check that the matrix μ obtained by the proper normalization of A corresponds to the matrices P and Q. □

Let us assume that μ is Nash and there exists a language μ' such that $F(\mu,\mu) = F(\mu,\mu')$, and $F(\mu',mu') > F(\mu,\mu)$. There are selective pressures in the system for the language μ to be replaced by the language μ'. A much more "reliable" equilibrium states are given by weak ESS. If a system is at a weak ESS, it can only change its state because of a random drift which takes a very long time in large populations [24]. Therefore, it is important to be able to characterize all weak ESS of the language system.

It has been observed by [38] that Nash equilibria may or may not correspond to weak ESS. Here we derive specific conditions under which Nash equilibria are weak ESS languages.

5.3. Weak ESS for uniform σ-distributions

We will say that a language μ has synonyms (homonyms) if some column (row) of the matrix μ has more than one positive entries. Let us call a synonym (homonym) *isolated* if the corresponding rows (columns) of the matrix μ contain no other positive elements.

Example 5.4. Consider the language

$$\mu = \frac{1}{2+\beta}\begin{pmatrix} 0 & \alpha & 1-\alpha & 0 \\ \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma \\ 0 & 0 & 0 & 1-\gamma \end{pmatrix}. \tag{26}$$

The synonyms μ_{34} and μ_{44} are isolated because the 3rd and the 4th rows do not contain other positive entries. The homonyms μ_{12} and μ_{13} are also isolated because they are the only entries in columns 2 and 3.

We will call *syno-homonyms* such sets of elements that for each two of them it is possible to find a chain of elements connecting the given two elements via *synonym-synonym* or *homonym-homonym* relationships. Example 5.2 contains syno-homonyms.

Suppose that the only synonyms and homonyms of a language, μ, are isolated ones. We have the following observation.

Observation 5.5. For any μ_* which is a best response to μ, the function $F(\mu,\mu_*)$ does not depend on the actual entries in the matrices but is simply equal to $1/M$ times the number of "effective" elements, if we count all the synonyms corresponding to the same meaning as one effective element, and all the homonyms expressed by the same word as one effective element.

To illustrate this we note that in Example 5.4 we have three effective elements (the two synonyms counted as one and the two homonyms counted as one). Thus $F(\mu,\mu) = \frac{3}{4}$.

The following statement holds:

Theorem 5.6. *For the uniform probability distribution σ, the language μ is a weak ESS if and only if the only kind of synonyms* (*homonyms*) *it has are isolated synonyms* (*homonyms*).

Proof. The proof contains two parts. First we assume that the language satisfies the conditions of the theorem and prove that (i) it is Nash and (ii) if for some μ', $F(\mu,\mu') = F(\mu,\mu)$ then $F(\mu',\mu') \leqslant F(\mu,\mu)$. Then, we will show that the languages of Theorem 5.6 are the only weak ESS.

The languages μ of Theorem 5.6 are Nash because they obey the conditions of Theorem 5.1. Next, we can see that by construction, the auxiliary matrix X contains no closed loops and no turns, and thus by Theorem 3.9, the best response exists. From Observation 5.5 it follows that $F(\mu,\mu_*) = F(\mu,\mu) = F(\mu_*,\mu_*)$, i.e., μ is a weak ESS.

To conclude the proof we need to show that no other Nash language is an ESS. Let us suppose that a language μ'' does not satisfy the conditions of Theorem 5.6. Then it contains sets of syno-homonyms. We will show that there exists a language $\tilde{\mu}$ satisfying the conditions of Theorem 5.6 such that $F(\mu'',\mu'') = F(\mu'',\tilde{\mu})$ but

$$F(\tilde{\mu},\tilde{\mu}) > F(\mu'',\mu''). \tag{27}$$

Using the algorithm of finding the best response, we can see that the P^* and Q^* matrices for language μ'' have the same support as the language itself, but have no symmetries. It is easy to check that the choice of P^* and Q^* (and thus the best response, μ_*) is not unique.

Let us identify all the groups of syno-homonyms in μ_*; generally, they can be represented as sub-matrices of the size $k \times l$ for some $1 \leqslant k \leqslant N$ and $1 \leqslant l \leqslant M$, with no rows or columns consisting entirely of zeros. Each of the sub-matrices has to be considered separately to maximize its contribution to the function $F(\mu_*,\mu_*)$. Below we will assume for simplicity that there is only one group of syno-homonyms, as the generalization to multiple groups is straightforward.

Now we will build such $\tilde{\mu}$ that $F(\tilde{\mu},\tilde{\mu}) = \max_{\mu_*} F(\mu_*,\mu_*)$; for clarity it is useful to consult the example considered below. In the support of the matrices P^* and Q^*, X, let us identify the largest extended permutation matrix which belongs to the support. It is not unique, and its size cannot be bigger than $m_{kl} \equiv \min(k,l)$. This is the skeleton of the matrix $\widetilde{X}$, the support of the matrix $\tilde{\mu}$. Next, we need to make sure that the support of $\widetilde{X}$ contains at least one element from each row and column of the sub-matrix—otherwise $F(\mu'',\mu'') \neq F(\mu'',\tilde{\mu})$. This can be done by adding some elements from matrix X to the skeleton permutation matrix, one per each row (column) that are missing from $\widetilde{X}$. It is easy to check that the resulting $k \times l$ matrix can only contain isolated synonyms or homonyms. To build the $\widetilde{P}$ and $\widetilde{Q}$ matrices out of $\widetilde{X}$, we simply make sure that they satisfy the standard normalization conditions. This leaves the entries corresponding to the isolated homonyms in the $\widetilde{Q}$ matrix (synonyms in the $\widetilde{P}$ matrix) undefined.

All is left is to show that for this measure $\tilde{\mu}$, condition (27) holds. Let us consider the function $F(\mu_*,\mu_*)$, where μ_* is any best response to μ''. $F(\mu_*,\mu_*)$ is a linear function of its arguments, the entries of the matrices P^* and Q^*. A linear function can only reach its maximum on the boundary of the domain. For each row of the Q^* matrix, its entries lie on a simplex. The maximum is reached when one of these elements is one and the rest are zero. Because of the restriction that the supports of the matrices P^* and Q^* must coincide, the

non-zero entries corresponding to the vertices of the simplexes must form a permutation matrix. Note that the value of the function $F(\mu_*, \mu_*)$ does not depend on the value of the entries corresponding to isolated synonyms/homonyms. This means that by construction, $\tilde{\mu}$ defined as above corresponds to the maximum of the function $F(\mu_*, \mu_*)$, and therefore inequality (27) holds.

We conclude that for any language μ'' containing syno-homonyms, one can always find a matrix $\tilde{\mu}$ satisfying conditions of Theorem 5.6 such that $F(\tilde{\mu}, \tilde{\mu}) > F(\mu'', \tilde{\mu}) = F(\mu'', \mu'')$. Theorem 5.6 is proven. □

To illustrate Theorem 5.6 and the above algorithm, we consider Example 5.2. It presents a Nash language which does not satisfy the conditions of Theorem 5.6, i.e., contains syno-homonyms. Its best response is defined by the following matrices:

$$P^* = \begin{pmatrix} \alpha_1 & 0 & \beta_1 & \gamma_1 & 0 & \delta_1 & 0 & 0 \\ \alpha_2 & 0 & 0 & \beta_2 & \gamma_2 & 0 & \delta_2 & 0 \\ 0 & \alpha_3 & 0 & 0 & \beta_3 & \gamma_3 & \delta_3 & 0 \\ 0 & 0 & \alpha_4 & 0 & 0 & 0 & 0 & \beta_4 \\ 0 & \alpha_5 & 0 & 0 & 0 & 0 & 0 & \beta_5 \end{pmatrix},$$

$$Q^* = \begin{pmatrix} a_1 & 0 & a_3 & a_4 & 0 & a_6 & 0 & 0 \\ b_1 & 0 & 0 & b_4 & a_5 & 0 & a_7 & 0 \\ 0 & a_2 & 0 & 0 & b_5 & b_6 & b_7 & 0 \\ 0 & 0 & b_3 & 0 & 0 & 0 & 0 & a_8 \\ 0 & b_2 & 0 & 0 & 0 & 0 & 0 & b_8 \end{pmatrix},$$

with the usual normalization restrictions plus the condition that P^* and Q^* must have identical support. The normalization conditions state that the elements of the columns of P^* and rows of Q^* belong to some simplexes, e.g., $a_1 + a_3 + a_4 + a_6 = 1$. Some of the entries of the matrices P^* and Q^* are allowed to be zero, but there can be no rows or columns consisting entirely of zeros. If this holds, we have $(\mu'', \mu_*) = 5/2$ no matter what the entries of μ_* are (this follows from Eqs. (24) and normalization conditions on P^*, Q^*).

The function $F(\mu_*, \mu_*)$, on the other hand, depends on the entries of μ_* and reaches its maximum at the corners of the simplexes. In order to find the maximum, we need to identify the largest permutation matrix in the support of μ_*; in Fig. 4 its elements are encircled by solid lines. Then we make sure that there are no zero rows or columns by adding three more elements to $\widetilde{X}$ (they are encircled by a dotted line). The maximum value of $F(\mu_*, \mu_*)$ is 5, because such is the number of effective entries of $\tilde{\mu}$ (see Observation 5.5); in Fig. 4, the isolated homonyms are underlined. The matrices $\widetilde{P}$ and $\widetilde{Q}$ obtained from $\widetilde{X}$ are:

$$\widetilde{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

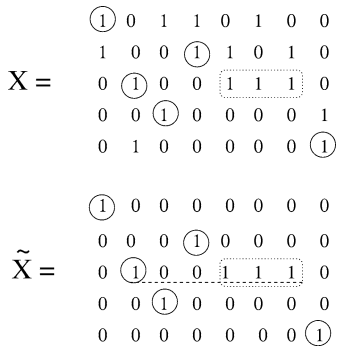


Fig. 4. Building the matrix $\tilde{\mu}$ for Example 5.2.

$$\widetilde{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & x_1 & 0 & 0 & x_2 & x_3 & x_4 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

with $x_1 + x_2 + x_3 + x_4 = 1$. The corresponding measure $\tilde{\mu}$ satisfies (27). We conclude that language of Example 5.2 is not a weak ESS.

5.4. *Weak ESS for general σ-distributions*

First of all we will generalize Theorem 5.1 for the case of non-uniform σ. We have

Theorem 5.1′. *Let Λ be the diagonal matrix with elements $\Lambda_{ii} = \sigma_i$. Then a language is Nash if the supports of its P, $P\Lambda$ and Q matrices coincide and if each row* (*column*) *of its $P\Lambda$ (Q) matrix contains at most two distinct values, one of which is zero.*

Example 5.7. For $N = 2$, $M = 3$ and σ_i given by $(1/2, 1/4, 1/4)$, the language

$$P = \begin{pmatrix} 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}, \qquad Q = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 1/3 & 0 & 2/3 \end{pmatrix},$$

$$P\Lambda = \begin{pmatrix} 1/4 & 1/4 & 0 \\ 1/4 & 0 & 1/4 \end{pmatrix}$$

is a Nash equilibrium.

Next, let us generalize our results about the weak ESS. As we saw in the previous section, for uniform σ-distributions weak ESS may contain isolated synonyms and homonyms. This means that the evolutionary system can sometimes get stuck in a sub-optimal state where the average communicative efficiency is smaller than one; this is a consequence of having homonyms in a language and/or not being able to express all meanings. Below we show that ambiguous languages can be stable only in the degenerate case of uniform probability distribution, σ. As soon as we lift this degeneracy, homonyms disappear from the language!

Theorem 5.8. *For non-uniform σ-distributions, the language μ is a weak ESS if and only if the only kind of synonyms it has are isolated synonyms. It cannot contain homonyms.*

Proof. First we will show that a Nash language for non-uniform σ-distributions cannot contain isolated homonyms.

Let us assume that there exists a Nash language, μ, such that it contains a string of l isolated homonyms. The fact that μ is Nash means that μ is the best response to itself, i.e., $F(\mu,\mu) = \sup_{\mu'} F(\mu,\mu')$. Let us follow the algorithm of Section 3.2.2 to construct the X matrix. In order for the X matrix to contain the sting of homonyms, the Q^* matrix must contain them, which means that the $P\Lambda$ matrix must contain a string of l identical elements, say, $(a,\ldots,a)$. This in turn means that the P matrix must contain a string $(a/\sigma_1,\ldots,a/\sigma_l)$. Since the values $\sigma_1,\ldots,\sigma_l$ are not all the same, this means that the elements below and above the string $(a/\sigma_1,\ldots,a/\sigma_l)$ of the matrix P cannot be all identically equal to zero (remember that the columns of P must sum up to one). Therefore, the matrix μ must contain non-zero elements below or above this string, i.e., the homonyms cannot be isolated, which is a contradiction.

Now, we need to show that for any language, μ'', which contains syno-homonyms, another language $\tilde{\mu}$ can be found such that $F(\mu'',\tilde{\mu}) = F(\mu'',\mu'')$ and $F(\tilde{\mu},\tilde{\mu}) > F(\mu'',\mu'')$. We proceed with building the language $\tilde{\mu}$ exactly as in the proof of Theorem 5.6. The difference emerges when we consider the function $F(\tilde{\mu},\tilde{\mu})$. Before, its value did not depend on the values of the elements which corresponded to isolated homonyms. Say, if we had a string $\alpha_1,\ldots,\alpha_l$ in the $\widetilde{Q}$ matrix, the corresponding elements entered as $\sum_{i=1}^{l}\alpha_i$, which is equal to one. Now, they enter as a linear combination $\sum_{i=1}^{l}\sigma_i\alpha_i$, and in order to maximize their contribution, we would have to take $\alpha_k = 1$ and $\alpha_j = 0$ for $j \neq k$, where σ_k is the largest of σ_i. Of course, this means that the support of the resulting matrix $\widetilde{Q}$ is smaller than the support of Q'', so we cannot use this language as a maximizer. However, we can take $\sigma_k = 1-\varepsilon$ and $\sigma_j = \varepsilon/(l-1)$. By choosing ε to be small enough, we can always find the language $\tilde{\mu}$ satisfying (27).

We conclude that the only weak ESS are languages which may not contain homonyms. □

Note that Nash equilibria in the case of non-uniform σ-distributions may contain isolated synonyms. The important difference is that isolated synonyms do not introduce any ambiguity in the language. We can conclude that in the case of general distributions, the ESS and weak ESS of the system correspond to the states with perfect coherence, i.e., no ambiguities may be present in the language. The only possible source of reduction of the average communicability function may come from poverty, i.e., the absence of certain meanings from the language.

Remark 5.9. Theorems 5.6 and 5.8 can be proven also if we do *not* assume that the matrices P and Q are connected through a common matrix μ, and the proofs are not much longer than the ones presented here.

6. Conclusions

We have considered a system of linguistic agents, each characterized by a language, i.e., a measure μ on the (signal$\times$meaning) space. The mutual intelligibility of such agents can be characterized in a natural way as a simple probability to transmit signals successfully both ways. We have studied the problem of optimizing the mutual intelligibility of linguistic agents in a shared environment.

It turned out that, for a given language μ_0, another language can be found which leads to the mutual intelligibility higher than the one achieved with μ_0 itself. (The exceptions are the languages which correspond to Nash equilibria, for instance, permutation languages.) Moreover, a family of languages, μ_ε, exists which leads to the optimization of intelligibility as $\varepsilon \to 0$. We have identified an algorithm to construct such languages with and without external noise in the system.

The results are of consequence for learning theory. It is apparent that in order to maximize intelligibility, the offspring is better off learning the "best response" languages found here rather than simply copying the language of their parents (or the population). We have identified some algorithms that can be used for this learning task and calculated their efficiency.

From the evolutionary prospective, we can identify all the languages which correspond to evolutionary stable strategies in a language game. It turns out that the strict ESS (i.e., the stable equilibria of the system) are languages which relate signals to meanings in a one-to-one way. The weak ESS (the neutral equilibria in dynamics) may contain isolated synonyms, but never homonyms, which means that there cannot be ambiguity in language (the exception is the degenerate case where all meanings occur with exactly the same frequency).

Appendix A. Infinite matrices

Here we present an algorithmic proof of Theorem 2 for infinite languages. The main difficulty here is that the best decoder and the best encoder cannot be defined in the same way as they were for finite matrices, formulas (8) and (9). In the present case we need to show that an equivalent of matrices P^* and Q^* can be constructed. We will prove the following

Lemma A.1. *For any ε_0, there exists an integer $\mathcal{N}$ and a pair $(\widetilde{P}^*, \widetilde{Q}^*)$ of $\mathcal{N}\times\mathcal{N}$ matrices such that*

$$\left| \sup_{\mu} \tfrac{1}{2} \sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q^0_{kl} - \tfrac{1}{2} \sum_{l=1}^{\mathcal{N}} \sigma_l \sum_{k=1}^{\mathcal{N}} \widetilde{P}^*_{kl} Q^0_{kl} \right| < \varepsilon_0, \tag{A.1}$$

$$\left| \sup_{\mu} \tfrac{1}{2} \sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} P^0_{kl} Q_{kl} - \tfrac{1}{2} \sum_{l=1}^{\mathcal{N}} \sigma_l \sum_{k=1}^{\mathcal{N}} P^0_{kl} \widetilde{Q}^*_{kl} \right| < \varepsilon_0. \tag{A.2}$$

Remark A.2. The matrix $\widetilde{P}^*$ can be said to be within ε_0 of the best decoder, and the matrix $\widetilde{Q}^*$ is within ε_0 of the best encoder. Finding these matrices is equivalent to controlling the behavior of, respectively, the second and the first terms in the expression for $F(\mu_0, \mu)$:

$$F(\mu_0, \mu) = \tfrac{1}{2} \sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} \left[P_{kl}^0 Q_{kl} + P_{kl} Q_{kl}^0 \right]. \tag{A.3}$$

Proof. First of all, for any language μ, $\forall\, \varepsilon_1 > 0$, there exists an integer $\mathcal{L}$ such that

$$\left| \tfrac{1}{2} \sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} P_{kl}^0 Q_{kl} - \tfrac{1}{2} \sum_{l=1}^{\mu} \sigma_l \sum_{k=1}^{\infty} P_{kl}^0 Q_{kl} \right| < \varepsilon_1, \tag{A.4}$$

$$\left| \tfrac{1}{2} \sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q_{kl}^0 - \tfrac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q_{kl}^0 \right| < \varepsilon_1. \tag{A.5}$$

This is because σ_l is a measure and the "tails", $\frac{1}{2} \sum_{l=\mathcal{L}}^{\infty} \sigma_l \sum_{k=1}^{\infty} P_{kl}^0 Q_{kl}$ and $\frac{1}{2} \sum_{l=\mathcal{L}}^{\infty} \sigma_l \times \sum_{k=1}^{\infty} P_{kl} Q_{kl}^0$ can always be made small enough by adjusting $\mathcal{L}$. The same inequalities hold for the supremum values of all terms.

Let us concentrate on the first term of $F(L_0, L)$. For any language L, $\forall \varepsilon_1 > 0$, $\exists \mathcal{K}_1$ such that

$$\left| \tfrac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\infty} P_{kl}^0 Q_{kl} - \tfrac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_1} P_{kl}^0 Q_{kl} \right| < \varepsilon_1, \tag{A.6}$$

because P_{kl}^0 is a measure in the index k. Thus the behavior of the first term of $F(L_0, L)$ can be controlled at infinity, and limiting the range of l by $\mathcal{L}$ and the range of k by $\mathcal{K}_1$ only introduces an error smaller than $2\varepsilon_1$.

For $1 \leqslant k \leqslant \mathcal{K}_1$ let us define $\tilde{l}(k)$ such that $P_{k,\tilde{l}(k)}^0 = \max_l P_{k,l}^0$ (here we assume for simplicity that the property of unique maxima holds). Then we can set the "nearly best decoder" $\widetilde{Q}^*$ to be

$$\widetilde{Q}_{kl}^* = \begin{cases} 1, & l = \tilde{l}(k), \\ 0, & \text{otherwise.} \end{cases} \tag{A.7}$$

Clearly we have $\sup_L \frac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_1} P_{kl}^0 Q_{kl} = \frac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_1} P_{kl}^0 \widetilde{Q}_{kl}^*$. Therefore, we obtain

$$\left| \sup_L \tfrac{1}{2} \sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} P_{kl}^0 Q_{kl} - \tfrac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_1} P_{kl}^0 \widetilde{Q}_{kl}^* \right| < 2\varepsilon_1. \tag{A.8}$$

Next, we turn to the second term of $F(\mu_0, \mu)$. Its behavior is harder to control in the k direction because Q_{kl}^0 is *not* a measure with respect to the index k. However, we can still approach $\sup_L \frac{1}{2} \sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q_{ml}^0$ using the following construction. We note that $\forall l$, $1 \leqslant l \leqslant \mathcal{L}$, the sequence $Q_{1,l}^0, Q^{2,l}, \ldots, Q_{i,l}^0, \ldots$ is contained between 0 and 1. Therefore,

we can find such $\tilde{k}(l) < \infty$ that $|Q^0_{\tilde{k}(l),l} - \sup_k Q^0_{kl}| < \varepsilon_1$. Now we define the "nearly best encoder" as follows:

$$\widetilde{P}^*_{kl} = \begin{cases} 1, & k = \tilde{k}(l), \\ 0, & \text{otherwise.} \end{cases} \tag{A.9}$$

We have: $\sup_\mu \frac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q^0_{kl} = \frac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sup_k Q^0_{kl}$. Therefore, if we set $\mathcal{K}_2 = \max_i(\tilde{k}(i))$, we obtain

$$\left| \sup_\mu \tfrac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q^0_{kl} - \tfrac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_2} \widetilde{P}^*_{kl} Q^0_{kl} \right| < \varepsilon_1. \tag{A.10}$$

By combining this with inequality (A.5), we get

$$\left| \sup_\mu \tfrac{1}{2}\sum_{l=1}^{\infty} \sigma_l \sum_{k=1}^{\infty} P_{kl} Q^0_{kl} - \tfrac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_2} \widetilde{P}^*_{kl} Q^0_{kl} \right| < 2\varepsilon_1. \tag{A.11}$$

Next, let us take $\mathcal{N} = \max(\mathcal{L}, \mathcal{K}_1, \mathcal{K}_2)$. It is possible to show that

$$\left| \tfrac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_1} P^0_{kl} \widetilde{Q}^*_{kl} - \tfrac{1}{2}\sum_{l=1}^{\mathcal{N}} \sigma_l \sum_{k=1}^{\mathcal{N}} P^0_{kl} \widetilde{Q}^*_{kl} \right| < \varepsilon_1, \tag{A.12}$$

$$\left| \tfrac{1}{2}\sum_{l=1}^{\mathcal{L}} \sigma_l \sum_{k=1}^{\mathcal{K}_2} \widetilde{P}^*_{kl} Q^0_{kl} - \tfrac{1}{2}\sum_{l=1}^{\mathcal{N}} \sigma_l \sum_{k=1}^{\mathcal{N}} \widetilde{P}^*_{kl} Q^0_{kl} \right| < \varepsilon_1. \tag{A.13}$$

Finally, we set $\varepsilon_1 \equiv \varepsilon_0/3$. Combining formulas (A.8) and (A.13), we obtain inequality (A.2). Combining formulas (A.11) and (A.12) we obtain inequality (A.1). □

Now we present a proof of Theorem 3.8 for infinite matrices.

Proof. Following the algorithm for finite matrices developed in Section 3.1, let us construct a family of $\mathcal{N} \times \mathcal{N}$ languages, L^ε, such that

$$\left| F(\mu_0, \mu^\varepsilon) - \tfrac{1}{2}\sum_{l=1}^{\mathcal{N}} \sigma_l \sum_{k=1}^{\mathcal{N}} \left(P^0_{kl} \widetilde{Q}^*_{kl} + \widetilde{P}^*_{kl} Q^0_{kl} \right) \right| < \varepsilon_0. \tag{A.14}$$

Combining this with inequalities (A.1) and (A.2) of Lemma A.1 we obtain:

$$\left| \sup_\mu F(\mu_0, \mu) - F(\mu_0, \mu^\varepsilon) \right| < 3\varepsilon_0. \tag{A.15}$$

Thus we conclude that the family of languages μ_ε satisfies the conditions of Theorem 3.8. A generalization to the case when the language μ_0 does not have the property of unique maxima is straightforward. □

Appendix B. Noisy channel

From the discussion of Section 3 it is clear that "perfect" languages, i.e., those whose association matrix is a permutation matrix, have communicability $F(\mu_0, \mu_0) = 1$.

Therefore, the best language to communicate with a perfect language is that language itself. However, imagine that an agent needs to communicate with a perfect language user (say μ_0) across a noisy channel. What is the optimal language μ_* for such communication?

In this section we will use the same three assumption as in Section 3.1. We consider only memoryless transmission media and therefore introduce the $M \times N$ matrix T such that T_{ij} is the probability that signal j is received by the listener given that signal i was conveyed by the speaker. We have

$$\sum_{j=1}^{n} T_{ij} = 1.$$

Now the F function can be rewritten in the following way:

$$F(\mu_0, \mu) = \tfrac{1}{2}\big[\mathrm{tr}\big(P^0 T^{\mathrm{T}} Q^{\mathrm{T}}\big) + \mathrm{tr}\big(P T^{\mathrm{T}} (Q^0)^{\mathrm{T}}\big)\big].$$

Let us introduce the effective encoding and decoding matrices of language μ_0:

$$\widetilde{P}^0 = P^0 T^{\mathrm{T}}, \quad \widetilde{Q}^0 = Q^0 T.$$

We obtain:

$$F(\mu_0, \mu) = \tfrac{1}{2}\big[\mathrm{tr}\big(\widetilde{P}^0 (Q)^{\mathrm{T}}\big) + \mathrm{tr}\big(P (\widetilde{Q}^0)^{\mathrm{T}}\big)\big]. \tag{B.1}$$

This definition is formally very similar to the definition with noiseless transmission, except the matrices $\widetilde{P}^0$ and $\widetilde{Q}^0$ are not necessarily related through a common association matrix. In the case when P^0 and Q^0 are identity matrices we have

$$\widetilde{P}^0 = T^{\mathrm{T}}, \qquad \widetilde{Q}^0 = T.$$

Given the matrix T, we would like to optimize the function $F(\mu_0, \mu)$ over all languages μ.

Let us maximize the two terms in expression (B.1) separately. The best encoder, Q^*, is given by picking out the maximum elements in each row of the matrix $\widetilde{P}^0$, i.e., in the matrix T^{T}. The best decoder, P^*, is given by picking out the maximum elements in each column of the matrix $\widetilde{Q}^0$, i.e., in the matrix T. Therefore, we have

$$P^* = (Q^*)^{\mathrm{T}}. \tag{B.2}$$

If for a language, μ_*, $P = P^*$ and $Q = Q^*$, then $F(\mu_0, \mu_*) = \sup_\mu F(\mu_0, \mu)$. In general, it is not possible to find such a language. However, under certain restrictions on the T matrix, we can approach the desired communicability.

We will say that a matrix T is *row-wise diagonally dominant*, if for all $1 \leqslant i \leqslant M$,

$$T_{ii} > T_{ij}, \quad \forall j \neq i.$$

We can prove the following

Theorem B.1. *If μ_0 is a permutation language and T is diagonally dominated row-wise, then* $\sup_\mu F(\mu_0, \mu) = 1/(2M)\,\mathrm{tr}(P^0 T^{\mathrm{T}} (Q^*)^{\mathrm{T}} + P^* T^{\mathrm{T}} (Q^0)^{\mathrm{T}})$.

The proof follows the same logics as the one given in Section 3.1. The key factor again is that there are no closed loops in the auxiliary matrix combining the positive entries of P^* and Q^*. This is established by

Lemma B.2. *If μ_0 is a permutation language and the T matrix is row-wise diagonally dominant, then there can be no closed loops in the auxiliary matrix.*

Proof. Consider a closed loop with (α_1, β_1) going to (α_1, β_2) ultimately to (α_K, β_K) and finally back to (α_1, β_1). Without loss of generality, we can assume that the node (α_1, β_1) corresponds to a 1 in the Q^* matrix. Immediately, it follows that

$$P_{\alpha_1\beta_1} > P_{\alpha_1\beta} \quad \forall\beta.$$

But since $P = T^{\mathrm{T}}$, we have that

$$T_{\beta_1\alpha_1} > T_{\beta\alpha_1} \quad \forall\beta. \tag{B.3}$$

Now consider (α_1, β_2). For this node, Q^* has a corresponding entry of 0 and therefore P^* has a corresponding entry of 1. Since P^* is obtained by taking maxima of columns of Q, we have

$$Q^0_{\alpha_1\beta_2} > Q^0_{\alpha\beta_2} \quad \forall\alpha,$$

and since $Q = T$, we have that

$$T_{\alpha_1\beta_2} > T_{\alpha\beta_2}\forall\alpha. \tag{B.4}$$

Matrix T is row-wise diagonally dominant, and therefore

$$T_{ii} > T_{ik} \quad \forall k \neq i.$$

Thus, from Eqs. (B.3) and (B.4) and the diagonal dominance property, we have

$$T_{\beta_1\beta_1} > T_{\beta_1\alpha_1} > T_{\alpha_1\alpha_1} > T_{\alpha_1\beta_2} > T_{\beta_2\beta_2}.$$

Now continue from (α_2, β_2) and use the same logics. We get,

$$T_{\beta_2\beta_2} > T_{\beta_2\alpha_2} > T_{\alpha_2\alpha_2} > T_{\alpha_2\beta_3} > T_{\beta_3\beta_3}.$$

This can be repeated to eventually obtain

$$T_{\beta_{K-1}\beta_{K-1}} > T_{\beta_K\beta_K},$$

and finally,

$$T_{\beta_K\beta_K} > T_{\beta_1\beta_1}.$$

This leads to a contradiction. □

From Lemma B.2 it follows, that if T is dominated by its diagonal, we can approach the $\sup_\mu F(\mu_0, \mu)$ arbitrarily close by choosing an appropriate language μ. The proof of Theorem B.1 is now straightforward. What is interesting is that the languages which have a high communicability with μ_0 are not necessarily identity matrices. Here is

Example B.3. Consider the following 3×3 T matrix:

$$T = \begin{pmatrix} 0.46 & 0.45 & 0.09 \\ 0.3 & 0.4 & 0.3 \\ 0.47 & 0.05 & 0.48 \end{pmatrix}.$$

For this matrix, we have

$$P^* = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad Q^* = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, the auxiliary matrix X combined out of positive elements of the P^* and Q^* matrices, is given by

$$X = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

It is symmetrical and contains no closed loops, and therefore we can construct the following family of languages:

$$A^\varepsilon = \begin{pmatrix} 0 & \varepsilon^2 & \varepsilon \\ \varepsilon^2 & 0 & 0 \\ \varepsilon & 0 & 1 \end{pmatrix}.$$

As ε tends to zero, the language A^ε tends to the best response to the perfect language μ_0 with the noisy channel T.

Finally, we note that if μ_0 is not a permutation language and T is row-wise diagonally dominated , then $\sup_\mu F(\mu_0, \mu) \leqslant 1/(2M)\,\mathrm{tr}(P^0 T^{\mathrm{T}}(Q^*)^{\mathrm{T}} + P^* T^{\mathrm{T}}(Q^0)^{\mathrm{T}})$, and the inequality can be strict, as is demonstrated by

Example B.4. The language μ_0 and the transition matrix, T, are given by

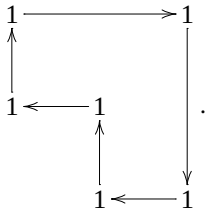
$$\mu_0 \propto \begin{pmatrix} 0.78 & 0.03 & 0.58 \\ 0.72 & 0.94 & 0.20 \\ 0.34 & 0.62 & 0.40 \end{pmatrix}, \qquad T = \begin{pmatrix} 0.72 & 0.00 & 0.28 \\ 0.28 & 0.43 & 0.29 \\ 0.02 & 0.41 & 0.57 \end{pmatrix}.$$

It turns out that the "best decoder" and the "best encoder" in this case are given by

$$P^* = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad Q^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which leads to the following auxiliary matrix with a closed loop:

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \Rightarrow$$



This suggests that finding the best encoder and the best decoder does not help us optimize the communicability function.

References

[1] C. Boutilier, Y. Shoham, M.P. Wellman, Economic principles of multi-agent systems (editorial), Artificial Intelligence 94 (1–2) (1997) 1–6.

[2] E.J. Briscoe, Language Evolution through Language Acquisition: Formal and Computational Models, Cambridge University Press, Cambridge, 2000.

[3] E. Charniak, Statistical Language Learning, MIT Press, Cambridge, MA, 1993.

[4] D.L. Cheney, R.M. Seyfarth, How Monkeys See the World: Inside the Mind of Another Species, University of Chicago Press, Chicago, IL, 1990.

[5] N.A. Chomsky, Language and the Problems of Knowledge, MIT Press, Cambridge, MA, 1986.

[6] D.R. Dowty, R.E. Wall, S. Peters, Introduction to Montague Semantics, Kluwer Academic, Dordrecht, 1980.

[7] A.L. Gorin, S.E. Levinson, A.N. Gertner, Adaptive acquisition of spoken language, in: Proc. ICASSP'91, Toronto, ON, 1991, pp. 805–808.

[8] M.D. Hauser, The Evolution of Communication, MIT Press, Cambridge, MA, 1997.

[9] J. Hofbauer, K. Sigmund, Evolutionary Games and Replicator Dynamics, Cambridge University Press, Cambridge, 1998.

[10] J.R. Hurford, Biological evolution of the Saussurean sign as a component of the language acquisition device, Lingua 77 (1989) 187–222.

[11] C. Isbell, C. Shelton, M. Kearns, S. Singh, P. Stone, A social reinforcement learning agent, in: Proc. Agents 2001, Montreal, QB, 2001, pp. 377–384.

[12] S. Kirby, Syntax out of learning: The cultural evolution of structured communication in a population of induction algorithms, in: D. Floreano, J.-D. Nicoud, F. Mondada (Eds.), Advances in Artificial Life, 5th European Conference, Lausanne, Switzerland, in: Lecture Notes in Computer Science, vol. 1674, Springer, Berlin, 1999, pp. 694–703.

[13] N.L. Komarova, P. Niyogi, M.A. Nowak, Evolutionary dynamics of grammar acquisition, J. Theor. Biol. 209 (1) (2001) 43–59.

[14] N.L. Komarova, M.A. Nowak, Evolutionary dynamics of the lexical matrix, Bull. Math. Biol. 63 (3) (2001) 451–485.

[15] J.M. Macedonia, C.S. Evans, Variation among mammalian alarm call systems and the problem of meaning in animal signals, Ethnol. 93 (1993) 177–197.

[16] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[17] J. Maynard Smith, Evolution and the Theory of Games, Cambridge University Press, Cambridge, UK, 1982.

[18] J. Maynard Smith, G. Price, The logic of animal conflict, Nature 246 (1973) 15–18.

[19] G.A. Miller, The Science of Words, Scientific American Library, New York, 1996.

[20] M. Oliphant, The dilemma of Saussurean communication, BioSystems 37 (1–2) (1996) 31–38.

[21] M. Oliphant, Formal approaches to innate and learned communication: Laying the foundations for language, PhD Thesis, Univ. of California, San Diego, CA, 1997.

[22] M. Oliphant, The learning barrier: Moving from innate to learned systems of communication, Adaptive Behavior 7 (1999) 371–384.

[23] M. Oliphant, J. Batali, Learning and the emergence of coordinated communication, Center for Research on Language Newsletter 11 (1) (1997).

[24] M.A. Nowak, An evolutionarily stable strategy may be inaccessible, J. Theor. Biol. 142 (1990) 237–241; M.A. Nowak, Stochastic strategies in the prisoners dilemma, Theor. Pop. Biol. 38 (1990) 93–112.

[25] M.A. Nowak, N.L. Komarova, P. Niyogi, Evolution of universal grammar, Science 291 (2001) 114–118.

[26] T. Regier, B. Corrigan, R. Cabasan, A. Woodward, M. Gasser, L. Smith, The emergence of words, in: Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Edinburgh, 2001.

[27] T. Regier, Emergent constraints on word-learing: A computational review, Trends in Cognitive Sciences 7 (2003) 263–268.

[28] F. de Saussure, in: C. Bally, A. Sechehaye (Eds.), Course in General Linguistics, Duckworth, London, 1983, Translated and annotated by Roy Harris.

[29] K. Smith, The cultural evolution of communication in a population of neural networks, Connection Sci. 14 (1) (2002) 65–84.

[30] K. Smith, The transmission of language: Models of biological and cultural evolution, PhD Thesis, University of Edinburgh, 2003.
[31] W.J. Smith, The Behavior of Communicating, Harvard University Press, Cambridge, MA, 1977.
[32] W.J. Smith, The behavior of communicating, after twenty years, in: D.H. Owings, M.D. Beecher, N.S. Thompson (Eds.), Perspectives in Ethnology, vol. 10, Plenum Press, New York, 1997, pp. 7–51.
[33] L. Steels, Self-organizing vocabularies, in: C. Langston (Ed.), Proceedings of ALife V, Nara, Japan, 1996.
[34] L. Steels, F. Kaplan, Spontaneous lexicon change, in: Proceedings of COLING-ACL, Montreal, QB, 1998, pp. 1243–1249.
[35] L. Steels, P. Vogt, Grounding adaptive language games in robotic agents, in: P. Husbands, I. Harvey (Eds.), Proceedings of the Fourth European Conference on Artificial Life, MIT Press, Cambridge, MA, 1997.
[36] J.B. Tenenbaum, F. Xu, Word learning as Bayesian inference, in: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Philadelphia, PA, 2000.
[37] F. Tohme, T. Sandholm, Coalition formation processes with belief revision among bounded rational self-interested agents, J. Logic Comput. 9 (6) (1999) 793–815.
[38] P.E. Trapa, M.A. Nowak, Nash equilibria for an evolutionary language game, J. Math. Biol. 41 (2000) 172–188.
[39] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[40] P. Vogt, H. Coumans, Investigating social interaction strategies for bootstrapping lexicon development, J. Artificial Soc. Social Simul. 6 (1) (2003).