TOWARDS AN ECONOMIC THEORY OF MEANING AND LANGUAGE

Gábor Fáth and Miklos Sarvary*

ABSTRACT: We present a model in which abstract concepts of a language acquire meaning as the result of competition between heterogeneous interacting agents in a community. We argue that bounded rationality requires individuals to use a reduced number of abstract concepts to represent the rich reality of the world. The meanings of these concepts emerge as a trade-off between two objectives: (i) agents want to use concepts that are best adopted to their idiosyncratic preferences and characteristic distribution of choice alternatives, (ii) agents seek to share concepts to benefit from communication. Agents play a non-cooperative game, whose Nash equilibrium determines the collective meanings of concepts in the population, constituting together the community's language. Analysis of the possible Nash equilibria and the evolutionary game dynamics shed light on interesting theoretical questions such as the origins of meaning, the coherence of language, the language-culture relationship, and Whorf's hypothesis on linguistic relativism.

KEYWORDS: evolution of language, mental representation, bounded rationality, game theory, Nash equilibrium, dimension reduction, PCA, agent heterogeneity

1 Introduction

Human language is a unique trait that clearly sets us apart from animals. Part of the human language system is biological, i.e., hard wired in us by millions of years of evolution. Speech organs, for instance, clearly belong to this category. Other parts of the language system are the result of more or less conscious cognitive processes such as learning, or deliberate social interactions. The creation of new words by a community is the result of complex interactions between its members. The subject of this paper is language in this latter sense, i.e., language as a conscious/social process.

Roughly speaking, language consists of words and rules. Words are linguistic signals referring to concepts, the collection of which constitutes the so-called mental dictionary or mental lexicon. We use the term "word" to denote a *listeme*: any string of linguistic elements (e.g., morphemes, words, or composite expressions) that is associated with a particular meaning. (In this general sense idioms, for instance, are also listemes.) This paper focuses on the emergence of the mental dictionary, i.e., it seeks to explain how words acquire meaning.

*Gábor Fáth is a Senior Research Fellow at the Research Institute for Solid State Physics and Optics of the Hungarian Academy of Sciences, P. O. Box 49, Budapest, 1525, Hungary, Tel: +36 1 392 2222, E-mail: fath@szfki.hu. Miklos Sarvary is Associate Professor of Marketing at INSEAD, Bd. de Constance, 77305, Fontainebleau, France, Tel: +33 1 60 71 26 05, E-mail: miklos.sarvary@insead.edu. This work was supported by INSEAD Foundation. Gábor Fáth also acknowledges support from the Hungarian Scientific Research Fund (OTKA) under grant Nos. T43330 and T47003 and by the Hungarian National R&D Program (NKFP) under grant No. 2/051/2004.

Our central thesis is that meanings emerge in language through an *economic* process, in the sense of competition between agents. We study intelligent maximizing agents with well-defined preferences facing a series of decisions. Agents have an overwhelming amount of information (physical perceptions) concerning their decision alternatives. However, with restricted cognitive resources and limited communication bandwidth, agents have to use a reduced number of abstract concepts (words) to represent the rich reality of the world. In this environment, agents' "choices" (of the meanings) of concepts is driven by the objectives of keeping representation error minimal and communication effective. The former means using concepts that provide the "best" description of the world in the sense of leading to optimal decisions given the agents' preferences. Since preferences may differ across agents, optimal concepts will differ too. However, as agents want to communicate with each other (e.g., to gather information from each other or to cooperate) they also need to share their representations (concepts). Trading off these two objectives (representation and communication) agents essentially play a non-cooperative game (Language Game), whose Nash equilibrium is a set of partially shared meanings that we interpret as "language". We are interested in the conditions under which such an equilibrium exists, as well as the level of coherence in the resulting words/concepts. We analyze the properties of the Language Game's evolutionary dynamics, and discuss the multiplicity and stability of its equilibria. Our ultimate goal is to understand the emerging mapping between agents' perceptions and the (abstract) concepts (words) of language. In particular, an important question we investigate is how individual preferences influence agents' mental representations of the world.

The analytic results show that the Language Game always has at least one Nash equilibrium, although it may not be unique. In the resulting language(s), concepts have partial coherence across agents, the strength of which is an increasing function of the exogenous benefit from communication. However, although the average coherence of concepts across agents increases monotonically with the strength of communication, the coherence of concepts can differ significantly. Analysis of the Language Game's evolutionary dynamics shows that it always has at least one fixed point that is also a Nash equilibrium. This equilibrium is also dynamically stable against small collective deviations (an attractor). However, not all attractors of the Language Game's dynamics are Nash equilibria and not all Nash equilibria are dynamically accessible. The number of potential equilibria increases with the number of agents. Which concepts emerge in equilibrium is sensitive to initial conditions, i.e., language exhibits strong path-dependence. These patterns have interesting implications for the interpretation of natural languages, their differences and relative stability across cultures and the psychology of mental representations.

2 Relevant Literature

There are two literature streams that formally model the language formation process and, in particular, the endogenous emergence of meaning in certain signals. In the "cheap-talk" literature of rational game theory [1, 2, 3], agents are properly defined in terms of preferences and interact strategically in a specific game with incomplete information. The purpose of language construction is to better coordinate on an efficient outcome of this game. It is shown that under certain conditions (non-costly) signals become associated with specific meanings and the use of the so-constructed language helps coordination. In models of evolutionary biology [4, 5, 6, 7] agents are boundedly

rational and use different heuristics. The association of specific signals to given meanings emerge through imitation or reproduction. It is shown that a shared mental lexicon (mapping between signals and objects of reality) can dynamically emerge if heuristics leading to successful communication have a higher chance to survive.

Our framework differs from these literature streams in several ways. First, in contrast with existing models, in our case, the social process of language construction involves explicit conflict of interest between agents. We assume that language has a strong influence on decision making (Rubinstein [8], Chapter 4). In fact, optimal decision making - in the sense of optimal choice between alternatives - is the driving force behind language formation. Language is used to describe choice alternatives with a limited number of concepts. As a result, agents with different preferences may not want to attribute the same meanings to the concepts. Yet, if they want to communicate with other members they have to largely agree on the meanings. These, potentially opposing objectives result in a competitive game, that we call Language Game. A related issue is that traditional evolutionary approaches assume that the language construction process is "unconscious", in the sense that agents' heuristics in associating specific signals to given objects survive through imitation or reproduction. In other words, in these models agents do not actively promote certain meanings to be used. In reality however, communities do construct their language in every day's life and their members do so consciously. They add words to the mental lexicon and modify the meanings of existing ones. Consistently with this picture, in our framework, agents are strategic in influencing the formation of a common language.

Second, we seek to build a *general* model of language, in the sense of language being context-independent. In other words, we are interested in the emergence of concepts/words that are used across a large number of different contexts. In existing evolutionary approaches, concepts do not really emerge endogenously. Rather, concepts and signals are given and the issue is to reach agreement over the mental lexicon providing the mapping between them [9, 10]. In other words, these models do not explicitly address *which* concepts should be named by signals (words) and why concepts may be similar or different across agents and/or communities. As Rubinstein [8] points out in his reflections on language, even in cheap-talk games, the relevant concepts are already defined by the underlying game. As such, the messages are limited to certain well-defined objects. Rubinstein argues [8]: "A persuasive explanation of the emergence of linguistic concepts requires a much more general setting" (p. 34.). Our goal is precisely to model such a general (context-independent) setting.

Two recent papers are closest to our work in addressing the endogenous emergence of meanings with conflict of interest between agents. Battigalli and Maggi [11] construct a model of language to build a theory of contract incompleteness. A contract uses language to partition the set of events and associate it to the contracting parties' obligations. A more "precise" - hence more costly - language results in more complete contracts. The paper explains various forms and degree of contract incompleteness by the cost of the language used. In a different formulation, Cremer *et al.* [12] develop a model where parts (departments) in an organization use different partitions of the space of signals to develop an internal code adapted to each department's objectives (see also, Wernerfelt [13]). Conflict between departments may arise when internal codes differ. The focus of the paper is the organizational structures that emerge as a result of such conflict or the need to make conflict disappear.

Our work has several points of departure from these papers. Instead of contracts and

organizational structures, our focus is the process of language formation itself and the features of the resulting equilibrium languages. In this respect, in our framework, the meanings of concepts are not discrete in a language community. As in reality, people do not completely agree or completely disagree with each other on the meaning of a word and this does not (completely) prevent communication. Words may have somewhat different meanings for different people even if there is rough agreement between them. Existing approaches do not allow for this eventuality. In contrast, our model captures this continuous aspect of meaning in the mental dictionary at the outset and - as will become clear later - even in equilibrium. In fact, an interesting question we ask is: to what extent is there agreement between agents on the meaning of words in a language?

Our model structure also differs fundamentally from the above papers'. In particular, we do not define words as a partition of the state space, but rather as weighted averages of the states. As will become clear later, this allows us to make a distinction between *language* and *culture*, that existing models have a hard time to incorporate.

Finally, our model is closely related to psychology, which is concerned by the mental representation of concepts and language's link to decision making in general. Psychologists broadly see mental representations as a result of a clustering procedure that gives birth to hierarchies of categories with more abstract concepts belonging to higher level categories [14, 15]. Objects and concepts are shown to belong to categories to various degrees, ranked by how "typical" members they are of the given category. More recent research in categorization acknowledges that such categorical structures are "ad hoc" and are closely linked to decision making in that they are defined by the decision maker's goal [16, 17]. Our framework is consistent with this view, in the sense that "categories" emerge endogenously in a language equilibrium, partially guided by decision making. Specifically, concepts are assumed to be weighted averages of the elementary (microscopic) signals that agents perceive about the world (we assume that these perceptions are identical across agents). Naturally, the optimal weighting scheme will reflect similarities and contrasts between the signals that we can associate with the "objective" structure of the world. However, the weights will also reflect the agent's preference structure, which we assume to be idiosyncratic, i.e. heterogenous across agents. In sum, our framework explicitly describes how mental representations are influenced by three fundamental "inputs" for language: (i) the correlation structure of the physical world, (ii) agents' (decision makers') individual preferences and (iii) the rate of communication. In this way, it also sheds light on numerous debates in psycholinguistics, such as the well-known Whorf hypothesis on linguistic relativism or the relationship between culture and language.

3 The Language Game

We define "language" as the Nash equilibrium of the Language Game. In the Language Game, agents choose a finite set of concepts (words) trading off two objectives: (i) to best "represent" their vast perceptions (available data) about the world and (ii) to share the concepts across each other, i.e., to associate - to the extent possible - the same meaning to any given word.

So defined, the Language Game has three important characteristics. First of all, language should provide an *accurate representation* of the complex world. The underlying assumption is that agents regularly face decisions, that is, choices between alternatives. Efficient decision making requires that choice alternatives be evaluated as accurately

as possible. The structure of language providing the necessary "mental representation" stems from two principal sources. Language needs to reflect the objective structure of the natural world, and at the same time, the subjective structure of the agents' preferences. For example, a hiker may prefer to use different symbols to draw the map of a region than a botanist, not because they perceive nature differently, but because they would like to distinguish different objects: while the hiker only needs to tell apart forest from clearing, the botanist needs to represent the dominant types of plants. Conversely, the botanist may be less interested in the details of the terrain's morphology. Their preferences are different.Traditionally, cognitive sciences tend to focus on how perceptions of reality influence the development of languages. *In contrast, in this paper, we are primarily concerned with the effect of agents' heterogenous preferences.*

The second characteristic is that the complex perceptions of agents are to be represented by a *finite set of optimally chosen concepts*. This is motivated by our assumptions that agents are boundedly rational and their communication bandwidth is limited. Both of these drive agents (in an evolutionary sense) towards a well-organized and compressed representation of the world. Bounded rationality, which can manifest itself in the form of limited memory space, time constraints on learning the correct use of concepts, or rapidly increasing processing cost as a function of model complexity, cannot let the number of concepts (words) to rise too high. Similarly, biology limits human communication bandwidth to a relatively low rate of words per second. Thus, efficient communication also requires the compression of information along a small number of highly significant concepts. Therefore, the typical size and organization of the human mental lexicon arises from the balance between two competing goals: to increase representational accuracy and decrease model complexity (see Chater and Vitnyi [18] for a short introduction on the importance of model simplicity). In order to simplify our model we will assume that the number of concepts in the language is fixed, and will only concentrate on the optimal choice of these concepts.

Finally, the third characteristic of a Language Game is that agents need to align their concepts if they want to profit from *communication*. The underlying, natural assumption is that the more concepts are similar (words have similar meanings), the more they are useful to transfer information about the state of the world (choice alternatives, etc.) between agents. We will assume that society imposes an exogenous "pressure" to communicate, which will contribute to agents' utility according to the actual alignment of their concepts. Notice, that if agents are different in terms of preferences, then the different aspects of language (accurate mental representation and communication) put agents in conflict with one another. A priori, each agent would prefer to use a different language for efficient decision making. Since agents also want to benefit from communication, it is in their interest to deform the collectively shared meanings in such a way that these better reflect their own preferred view of the world. Language, defined as the collection of concepts, emerges as an equilibrium of the Language Game.

3.1 Mental representation with a finite number of concepts

We consider I agents, each restricted to use a number K of concepts only. Agents divide the world into a number X of *decision contexts* in which they evaluate decision alternatives according to their personal preferences. We assume that alternatives are characterized by their objective (physical) attributes $\boldsymbol{a} = \{a_1, \ldots, a_D\}$, that are common knowledge. In other words, agents are homogenous in terms of their perceptions about

the world. Agent i's *objective payoff* from choosing alternative $\boldsymbol{a}$ in context x is denoted by $\pi_i^{(x)}(\boldsymbol{a})$. In the following, we assume for simplicity that the objective payoff is a *linear* function of the attributes

$$\pi_i^{(x)}(\boldsymbol{a}) = \boldsymbol{\omega}_i^{(x)} \cdot \boldsymbol{a}, \tag{1}$$

in which the vector of coefficients $\boldsymbol{\omega}_i^{(x)}$ will be called the agent's *preference vector in context* x. The collection of vectors $\{\boldsymbol{\omega}_i^{(x)}\}_{x=1}^{X}$ defines the agent's objective (biological or otherwise acquired) preferences in all possible decision contexts. These parameters are assumed to be fixed in the model. In other words, each agent's preference structure is described by a very large number XD of fixed parameters. As opposed to $\boldsymbol{a}$, the parameters $\{\boldsymbol{\omega}_i^{(x)}\}_{x=1}^{X}$ are agent specific, i.e., we consider *agent heterogeneity in preferences*.

If they knew their preference vectors, the agents would be able to calculate the objective payoffs of their decision alternatives in the assumed linear world, and make the best possible decisions whenever they face a decision problem. However, although the preference vectors $\boldsymbol{\omega}_i^{(x)}$ are well defined in theory, we have good reason to suppose that the numbers they represent are not directly available for the agents. On the one hand, the agents need a good deal of experimentation (learning) to find out how a given attribute impacts their payoffs. As an example think about food allergy, which may be an innate condition in the patient, but requires a painful "learning process" through bad choices of food alternatives to recognize and diagnose. When the number of attributes and contexts is large, there is a problem with the sparsity of the "training samples" (cf., the "powerty of stimulus" argument in linguistics [18]). On the other hand, a perfect knowledge of $\boldsymbol{\omega}_i^{(x)}$ would require a detailed understanding of the effect of all physical attributes on the payoffs: a total number of XD parameters in the linear model. We can assume that this is beyond the agent's mental capacity, or the required cognitive complexity is too costly for the agent.

What remains for the agent to reduce the cost of complexity and to have a model which better generalizes from sparse data is to try to invent a simplifying scheme, a so called *mental representation*. The mental representation is an *approximate* mapping from microscopic attributes to payoffs. Given a decision alternative $\boldsymbol{a}$ and using the mental representation, the agent arrives at an *approximate payoff*, $\tilde{\pi}_i^{(x)}(\boldsymbol{a})$, which is not equal but close to the exact payoff $\pi_i^{(x)}(\boldsymbol{a})$. We assume that there exists some learning mechanism, which improves the agent's mental representation by collecting experience on previous choices and their success rate. Eventually, as a result of learning, the approximate payoffs get as close to the objective ones as allowed by the structural constraints of the mental representation. The mental representation gets optimized within its limits.

In the following we suppose that the agent's mental representation has a fixed architecture with a number of free parameters to optimize for. In particular, we assume a two-level hierarchical organization, in which the processing of the input $\boldsymbol{a}$ is done in two steps: (i) evaluating the alternative along concepts, and (ii) weighting the concept scores to arrive at the associated payoff. The structure of this mental representation is depicted on Fig. 1. It defines three layers: the *input layer*, where the microscopic attributes of decision alternatives are perceived, the *middle layer* composed of abstract concepts, and the *output layer* representing the context- and agent-specific payoffs associated with alternatives.

The hierarchical structure depicted in Fig. 1 formally resembles a linear, two-level (concept vectors, mental weights) neural network. Note, however, that it is not a mi-

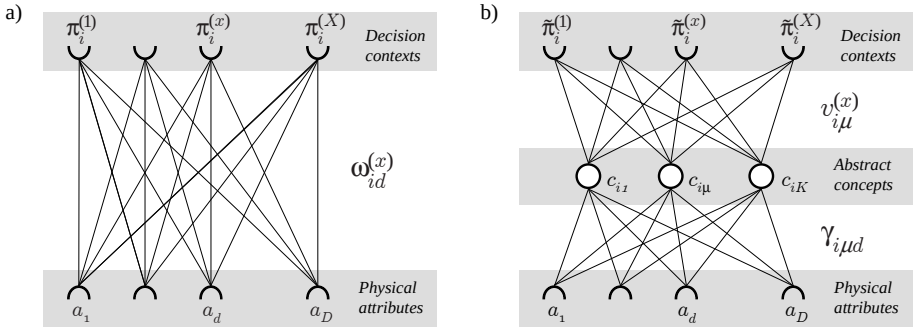

Figure 1: Evaluating decision alternatives under bounded rationality using a finite number of concepts. (a) Structure of reality reflecting idiosyncratic preferences, (b) structure of the mental model (both assumed linear).

croscopic neural network model of a cognitive function, but a phenomenological (macroscopic) model of a generic decision making strategy under bounded rationality. This structure is based on research in psychology, asserting that the human mind is a "feature detector" that can only perceive the aspect of reality which it has a concept for [19].

Each concept μ in the middle layer is a real valued function $\boldsymbol{a} \to c_{i\mu}(\boldsymbol{a})$ (the agent index i represents that the meaning of concepts may vary from agent to agent). We assume that the number of concepts, i.e., the size of the (mental) dictionary, is the same fixed number K for all agents with $K << D$. Thus, the first step implies *dimension reduction*, a mapping $\mathbb{R}^D \to \mathbb{R}^K$. The second step involves calculating the approximate payoff in a given decision context as a function of the concept scores, $\tilde{\pi}_i^{(x)} = p_i^{(x)}(c_{i1}, c_{i2}, \ldots, c_{iK})$, where $p_i^{(x)}$ is an appropriate function $\mathbb{R}^K \to \mathbb{R}$ called the *mental model* of agent i in context x. Each agent possesses a number X of such mental models, one for each decision context.

At the lowest level on Figure 1, agents are homogeneous and have identical perception of reality. In the highest layer they are heterogeneous and have individual payoff functions based on their individual preferences. The middle layer with abstract concepts shows partial coherence, whose measure, as we will see, is determined by the strength of social interactions. In other words, agents more or less agree in the meanings of the concepts, but there is no perfect consensus. This is the layer we intend to monitor across society for observing the emergence and evolution of a shared language.

In general, both the concept functions $c_{i\mu}$ and the mental models $p_i^{(x)}$ can be nonlinear functions. However, in order to keep the model simple, in the following, we will assume that all these functions are *linear*. This "linear mind" assumption and the former "linear reality" assumption are crude approximations but ensure some analytical results and thus useful insight. Accordingly, we assume that concepts are linear mappings defined by *concept vectors* $\{\boldsymbol{\gamma}_{i\mu}\}_{\mu=1}^K$, i.e.,

$$c_{i\mu}(\boldsymbol{a}) = \boldsymbol{\gamma}_{i\mu} \cdot \boldsymbol{a}, \tag{2}$$

and that the mental models are linear, too,

$$\tilde{\pi}_i^{(x)}(\boldsymbol{a}) = \sum_{\mu=1}^{K} v_{i\mu}^{(x)} c_{i\mu}, \tag{3}$$

where the coefficients $v_{i\mu}^{(x)}$ will be called *mental weights*. Putting the two steps together, the *approximate payoff* is bilinear in the $\boldsymbol{\gamma}$ and v parameters

$$\tilde{\pi}_i^{(x)}(\boldsymbol{a}) = \sum_{\mu=1}^{K} v_{i\mu}^{(x)} \boldsymbol{\gamma}_{i\mu} \cdot \boldsymbol{a}. \tag{4}$$

We emphasize that, as opposed to $v_{i\mu}^{(x)}$, the concepts are context-independent (there is no superscript (x) for $c_{i\mu}$), i.e., they have the same "meaning" in all decision contexts for a given agent.

Why is this cognitive architecture preferred? In this scheme the complexity, i.e., the number of variables defining an agent is $(X + D)K$, which can be much less than the total number of parameters describing the world, XD. However, there is a price to pay for this "bounded rationality": due to the reduction of dimensionality, $K << D$, the approximate payoff $\tilde{\pi}_i^{(x)}$ deviates from the objective payoff $\pi_i^{(x)}$. In fact we can think of K as the optimal number of concepts which arises endogenously in the trade-off of precision vs. model (representation) complexity. Nevertheless, as the exact value of K is not essential to our analysis, we simplify the model by assuming that K is exogenously fixed.

In our model, standard grammatical categories like nouns and adjectives are confounded. Each decision alternative receives a score on a concept, which can be associated with a noun or an adjective alike. For example, a stool will receive a high score on the "chair-ness" concept. This feature is consistent with empirical work on people's mental representations [14, 16, 15], which shows that, while people have a tendency to cluster things in distinct categories, membership in a category is not rigid but can be represented with a so-called "graded structure" where a member is measured on how typical it is for that category. (This is also the fundamental concept behind "fuzzy logic".)

Finally, a three-layer structure with one concept layer is clearly a simplified model of language, which could be better described with many layers, each representing a different level of abstraction. However, we can easily replace the lowest (physical) layer by a layer of concepts that are (for all practical purposes) completely agreed upon by agents. For example, there is strong agreement between people on the meanings of concepts/words like "chair", "table", "fork". There is much less agreement however, on the meanings of abstract words like "truth", or "God". We are interested in understanding how the meanings of such abstract concepts emerge.

Let us illustrate the model with a further example. Consider a headhunter seeking candidates (alternatives) for job openings (contexts). The headhunter possesses a large amount of raw data about the candidates in the form of CVs, test results, photos, recommendation letters, certificates, interview recordings, etc. (physical attributes, $\boldsymbol{a}$), and would like to use these to direct the right candidate to the right job. In theory, each job is associated with a complicated function (objective payoff), mapping candidates (described by $\boldsymbol{a}$) to payoff of the candidate for the given job ($\pi^{(x)}(\boldsymbol{a})$). Without going into the intimate details of this complex relationship, the headhunter can "summarize" the candidate profiles along a small number of suitably chosen concepts such as "level of education", "expertise", "communication skills", "physical appearance", etc, which she can (easily) extract from the attributes. Moreover, she classifies openings into typical job categories (say, musician, scientist, executive, etc.) for which she already possesses a weighting scheme (mental weights) along the concepts used. The question is how to find the most efficient set of abstract concepts to minimize her efforts but maximize her

matchmaking efficiency. Obviously, headhunters should optimize their concepts for the general genre of cliental they work for. Those specialized in hunting movie actors will apply rather different concepts (language) than those recruiting corporate executives.

3.2 Representation error

Each agents' goal is to find the best possible set of concepts and mental weights that minimize the error of the mental representation under the constraint that only a finite number K of concepts can be used. The natural measure of agent i's *representation error* is the variance of the payoff deviation over all decision contexts,

$$E_i^{\mathrm{REP}} = \sum_{x=1}^{X} \left\langle \left[\pi_i^{(x)}(\boldsymbol{a}) - \tilde{\pi}_i^{(x)}(\boldsymbol{a}) \right]^2 \right\rangle_{ix} . \tag{5}$$

Such a quadratic error function equally penalizes positive and negative deviations of the predicted (approximate) payoffs from the objective (exact) ones.

In Eq. (5), $\langle f(\boldsymbol{a}) \rangle_{ix} = \int f(\boldsymbol{a})\, \rho_i^{(x)}(\boldsymbol{a})\, d\boldsymbol{a}$ denotes average over the occurrences of alternatives, which is characterized by the probability density $\rho_i^{(x)}(\boldsymbol{a})$. In a realistic setup the distribution of alternatives can be context and agent dependent. Some alternatives may occur with different probabilities (maybe with zero probability) for some agents and/or in some decision contexts. In the following we omit context dependence, but keep a possible agent dependence, and assume $\rho_i^{(x)}(\boldsymbol{a}) = \rho_i(\boldsymbol{a})$ for all x. This simplifies the forthcoming analysis without losing essential features. We assume that the attributes are centered and their correlation structure is represented by the context-independent *covariance matrix* $\boldsymbol{A}_i$,

$$\langle a_d \rangle_{ix} = 0, \quad \langle a_d a_{d'} \rangle_{ix} = [\boldsymbol{A}_i]_{dd'} \quad \forall x. \tag{6}$$

Using Eq. (6) it is easy to see that the representation error becomes

$$E_i^{\mathrm{REP}} = \sum_{x=1}^{X} \left(\boldsymbol{\omega}_i^{(x)} - \sum_{\nu=1}^{K} v_{i\nu}^{(x)} \boldsymbol{\gamma}_{i\nu} \right) \cdot \boldsymbol{A}_i \left(\boldsymbol{\omega}_i^{(x)} - \sum_{\nu=1}^{K} v_{i\nu}^{(x)} \boldsymbol{\gamma}_{i\nu} \right) . \tag{7}$$

The agent's goal is to minimize his/her error E_i^{REP} by optimally choosing the concept vectors $\boldsymbol{\gamma}_{i\nu}$ and mental weights $v_{i\nu}^{(x)}$. Recall that $\boldsymbol{\omega}_i^{(x)}$ and $\boldsymbol{A}_i$ are assumed fixed in the model.

As formulated so far, concept vectors and mental weights are both dynamic variables. However, it is reasonable to think about concepts as "slow" variables, changing noticeably on the scale of decades or centuries, partly because they are shared across agents. In contrast, the mental weights are agent-specific and adapt to the existing concepts in months or years. Certainly, learning to use correctly a concept is much faster than inventing a new, collectively successful concept. Thus, in the following, we optimize the mental weights, assuming that they accommodate to the slow variables very shortly, and only keep concept vectors as dynamic variables.

Given the $\boldsymbol{\gamma}$ vectors the optimal value of $v_{i\mu}^{(x)}$ follows from the solution to the equation $\partial E_i^{\mathrm{REP}} / \partial v_{i\mu}^{(x)} = 0$. From this condition, and assuming that the $K \times K$ symmetric matrix

$$G_{i\nu\mu} = \boldsymbol{\gamma}_{i\nu} \cdot \boldsymbol{A}_i \boldsymbol{\gamma}_{i\mu} \tag{8}$$

is invertible (note that the concepts are not necessarily orthogonal), the optimal mental weights turn out to be:

$$v_{i\mu}^{(x)} = \sum_{\nu=1}^{K} [G_i^{-1}]_{\mu\nu} \boldsymbol{\gamma}_{i\nu} \cdot \boldsymbol{A}_i \boldsymbol{\omega}_i^{(x)}. \tag{9}$$

Writing this back to Eq. (7) we can write the error now as a function of the $\boldsymbol{\gamma}$'s only:

$$E_i^{\mathrm{REP}} = \sum_{x=1}^{X} \boldsymbol{\omega}_i^{(x)} \cdot \boldsymbol{A}_i \boldsymbol{\omega}_i^{(x)} - \sum_{x=1}^{X} \sum_{\mu,\nu=1}^{K} \left(\boldsymbol{\omega}_i^{(x)} \cdot \boldsymbol{A}_i \boldsymbol{\gamma}_{i\mu} \right) [G_i^{-1}]_{\mu\nu} \left(\boldsymbol{\omega}_i^{(x)} \cdot \boldsymbol{A}_i \boldsymbol{\gamma}_{i\nu} \right). \tag{10}$$

The first term is an uninteresting constant which can be neglected. We introduce a *representation utility* U_i^{REP} as the negative of the second term. This can be cast in a more compact form:

$$U_i^{\mathrm{REP}}(\boldsymbol{\Gamma}_i) = \mathrm{Tr}\left(\boldsymbol{\Gamma}_i^T \boldsymbol{A}_i^T \boldsymbol{B}_i \boldsymbol{A}_i \boldsymbol{\Gamma}_i \boldsymbol{G}_i^{-1} \right), \tag{11}$$

where Tr is the trace of the matrix, $\boldsymbol{\Gamma}_i \equiv [\boldsymbol{\gamma}_{i1} | \boldsymbol{\gamma}_{i2} | \ldots | \boldsymbol{\gamma}_{iK}]$ is the agent's *language matrix* formed from the concept vectors as columns, and

$$[\boldsymbol{B}_i]_{dd'} = \sum_x \omega_{id}^{(x)} \omega_{id'}^{(x)} \tag{12}$$

is the agent's *preference matrix*. The representation utility U_i^{REP} is a function of the agent's language matrix $\boldsymbol{\Gamma}_i$.

The representation utility is maximal if the concept vectors are chosen optimally. However, even before trying to solve this optimization problem it is immediately clear that the solution cannot be unique. Indeed, due to the linearity of the model the representation utility is invariant for a redefinition of the concept vectors in any (possibly nonorthogonal) ways, provided that the new vectors span the same K dimensional subspace.

Lemma 1 *Let $\boldsymbol{R}$ be an arbitrary real, nonsingular, $K \times K$ matrix. For the transformation $\widetilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\boldsymbol{R}$ the representation utility is invariant:*

$$U^{\mathrm{REP}}(\widetilde{\boldsymbol{\Gamma}}) = U^{\mathrm{REP}}(\boldsymbol{\Gamma}). \tag{13}$$

Proof: Since $\boldsymbol{R}$ is not singular, $\boldsymbol{R}^{-1}$ exists. The metric tensor transforms as

$$\widetilde{\boldsymbol{G}}_x = \boldsymbol{R}^T \boldsymbol{G}_x \boldsymbol{R}, \tag{14}$$

and its inverse becomes

$$\widetilde{\boldsymbol{G}}_x^{-1} = \boldsymbol{R}^{-1} \boldsymbol{G}_x^{-1} (\boldsymbol{R}^T)^{-1}. \tag{15}$$

Using this and the cyclic property of the trace,

$$\begin{aligned} U^{\mathrm{REP}} &= \mathrm{Tr}\left[\widetilde{\boldsymbol{\Gamma}}^T \boldsymbol{A}^T \boldsymbol{B} \boldsymbol{A} \widetilde{\boldsymbol{\Gamma}} \widetilde{\boldsymbol{G}}^{-1} \right] = \mathrm{Tr}\left[\boldsymbol{R}^T \boldsymbol{\Gamma}^T \boldsymbol{A}^T \boldsymbol{B} \boldsymbol{A} \boldsymbol{\Gamma} \boldsymbol{R} \boldsymbol{R}^{-1} \boldsymbol{G}^{-1} (\boldsymbol{R}^T)^{-1} \right] \\ &= \mathrm{Tr}\left[\boldsymbol{\Gamma}^T \boldsymbol{A}^T \boldsymbol{B} \boldsymbol{A} \boldsymbol{\Gamma} \boldsymbol{G}^{-1} \right], \end{aligned} \tag{16}$$

as claimed. □

In natural languages concepts are not always fully independent, but there is a tendency to describe reality along more or less uncorrelated dimensions (synonyms are

exceptions and we do not consider this possibility here). For example, identifying positions in 2D space we can use concepts like "left–right" and "front–rear" or alternatively "North–South" and "East–West", but we hardly use correlated pairs like "North-South" and "North-West–South-East", although this would be theoretically possible. (In this example, the word-pairs like "left–right" are understood as a single concept for a coordinate axis.) The likely reasons are the mental difficulty to process correlated variables and/or the increased sensitivity for noise of the mental model when concepts trying to span the relevant subspace are strongly correlated. It seems that real mental models involve a (nonlinear) cost term related to the correlation of concepts, and this cost term brings about an effective "repulsion" for concepts.

Having the liberty of Lemma 1 to choose a basis freely in the optimal subspace we can mimic this effect by requiring that the concept vectors be exactly uncorrelated. The correlation of concept μ with concept ν is defined as $\langle c_\mu c_\nu \rangle_x = \langle (\boldsymbol{\gamma}_\mu \cdot \boldsymbol{a})(\boldsymbol{\gamma}_\nu \cdot \boldsymbol{a}) \rangle_x = \boldsymbol{\gamma}_\mu \cdot \boldsymbol{A} \boldsymbol{\gamma}_\nu$, thus the agent's language is uncorrelated if

$$\boldsymbol{G}_i = \boldsymbol{\Gamma}_i^T \boldsymbol{A}_i \boldsymbol{\Gamma}_i = \mathbf{1}. \tag{17}$$

This condition can be used as a constraint in the optimization problem. This constraint reduces the degeneracy of the optimum stated by Lemma 1, but not fully, as will be discussed in the sequel. Given the constraint we have $\boldsymbol{G}_i^{-1} = \mathbf{1}$, and the representation utility simplifies to

$$U_i^{\mathrm{REP}}(\boldsymbol{\Gamma}_i) = \mathrm{Tr}[\boldsymbol{\Gamma}_i^T \boldsymbol{A}_i^T \boldsymbol{W}_i \boldsymbol{\Gamma}_i]. \tag{18}$$

where we have introduced for later convenience the generically non-symmetric matrix

$$\boldsymbol{W}_i \equiv \boldsymbol{B}_i \boldsymbol{A}_i, \tag{19}$$

the so-called *world matrix*. $\boldsymbol{W}_i$ represents in a concise form agent i's overall relationship to the world. It encompasses the agent's perception of structure in the occurrences of decision alternatives ($\boldsymbol{A}_i$) and his/her subjective preferences ($\boldsymbol{B}_i$).

3.3 Communication between agents

So far, we have only talked about how using a finite number of concepts affects decision making. Agents have a second important objective, namely to communicate. We assume that communication between pairs of agents occurs on the level of concepts. In other words, agents cannot communicate the values of the large number of physical attributes associated with an alternative, but can only provide the corresponding – relatively small number of – concept scores. The basic idea is that communication cannot operate on the level of attributes due to limited bandwidth, nor on the level of payoffs due to substantial heterogeneity in preferences, which is anticipated by the agents. For example, it is not possible to describe all details of a flower to a person trying to purchase flowers on the phone, but it is not too informative to say "I like it" or "it is beautiful" either unless agents have very similar preferences (tastes).

It is obvious that if agents i and j use a somewhat different definition (different meaning) for concept μ, then their communication involving this concept introduces some misunderstanding. We can assume that misunderstanding, in general, implies disutility for the agents, whose amount depends on how different the two concept vectors $\boldsymbol{\gamma}_{i\mu}$ and $\boldsymbol{\gamma}_{j\mu}$ are. The lowest order (bilinear) measure of the misunderstanding error is related to the overlap of concept vectors. Thus $\boldsymbol{\gamma}_{i\mu} \cdot \boldsymbol{\gamma}_{j\mu}$ can be used as a reasonable

measure of the average communication benefit for the agents. When the two concepts are exactly identical, $\boldsymbol{\gamma}_{i\mu} = \boldsymbol{\gamma}_{j\mu}$, communication utility is maximal. We write i's benefit from communication with agent j as

$$U_{ij}^{\mathrm{COM}} = \sum_{\mu} C_{ij\mu}\, \boldsymbol{\gamma}_{i\mu} \cdot \boldsymbol{\gamma}_{j\mu}, \tag{20}$$

with $C_{ij\mu}$ denoting the importance of concept μ in the communication of agents i and j. In the following, we restrict our attention to the simple case $C_{ij\,\mu} = c/(I-1)$ for all agent pairs identically, where $I-1$ is the number of agents to communicate with, and c is the exogenous *rate of communication* in the community. The denominator is introduced to obtain a meaningful limit when $I \to \infty$.

The contribution of all communications to agent i's utility is

$$U_i^{\mathrm{COM}} = c\,\frac{1}{I-1}\sum_{j\neq i}^{I} \mathrm{Tr}(\boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_j). \tag{21}$$

As is formulated above, communication benefit is a symmetric function, $U_{ij}^{\mathrm{COM}} = U_{ji}^{\mathrm{COM}}$. Indeed, it is reasonable to postulate that the benefit of communication is distributed symmetrically between the two agents involved. Communication is typically a role game in which the roles of being a sender (speaker) or a receiver (hearer) interchanges from time to time. Both sender and receiver can benefit in a single communication act: depending on the content of the message the sender can generate profit by influencing the receiver, or the receiver can have benefit by getting information from the sender. On the long run, benefit accumulates on both sides.

Collecting the representation and communications terms together the overall utility for language of agent i is $U_i = U_i^{\mathrm{REP}} + U_i^{\mathrm{COM}}$, which takes the form

$$U_i \;\; = \;\; \mathrm{Tr}\left(\boldsymbol{\Gamma}_i^T \boldsymbol{A}_i^T \boldsymbol{B}_i \boldsymbol{A}_i \boldsymbol{\Gamma}_i\right) + c\,\frac{1}{I-1}\sum_{j\neq i}^{I} \mathrm{Tr}\left(\boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_j\right), \tag{22}$$

with the nonlinear constraint

$$\boldsymbol{\Gamma}_i^T \boldsymbol{A}_i \boldsymbol{\Gamma}_i = \mathbf{1}. \tag{23}$$

As defined by Eqs. (22) and (23), we have a coupled and constrained maximization problem for the individual language matrices $\boldsymbol{\Gamma}_i$, $i = 1, \ldots, I$.

Again, even without solving the problem, it is clear that the maximizing solution will not be unique. Let us introduce the shorthand notations $\boldsymbol{\Gamma}_{\mathrm{all}} \equiv (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \ldots, \boldsymbol{\Gamma}_I)$ (all concept vectors in the problem) and $\boldsymbol{\Gamma}_{-i} \equiv \boldsymbol{\Gamma}_{\mathrm{all}} \backslash \boldsymbol{\Gamma}_i$ (all concept vectors but those of agent i) for later convenience. Clearly $\boldsymbol{\Gamma}_{\mathrm{all}} = (\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}_{-i})$ for any i. As the following lemma asserts, a *collective orthogonal rotation* in the subspaces spanned by the concept vectors leaves all U_i invariant:

Lemma 2 *An identical collective rotation of the concept vectors for all agents*

$$\forall\, i \quad \boldsymbol{\Gamma}_i' = \boldsymbol{\Gamma}_i \boldsymbol{O}, \qquad \boldsymbol{O}\boldsymbol{O}^T = \mathbf{1}, \tag{24}$$

leaves U_i and the constraints invariant,

$$U_i(\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}_{-i}) = U_i(\boldsymbol{\Gamma}_i', \boldsymbol{\Gamma}_{-i}'), \qquad \boldsymbol{\Gamma}_i'^T \boldsymbol{A}_i \boldsymbol{\Gamma}_i' = \mathbf{1}. \tag{25}$$

Proof: It follows from Lemma 1 that U^{REP} is invariant, so it remains to prove that U^{COM} is also invariant. This boils down to show that $\sum_\mu \boldsymbol{\gamma}_{i\mu} \cdot \boldsymbol{\gamma}_{j\mu} = \mathrm{Tr}(\boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_j)$ is invariant. Indeed, we have

$$\mathrm{Tr}(\boldsymbol{\Gamma}_i'^T \boldsymbol{\Gamma}_j') = \mathrm{Tr}(\boldsymbol{O}^T \boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_j \boldsymbol{O}) = \mathrm{Tr}(\boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_j), \tag{26}$$

where we have used the cyclic property of the trace and that $\boldsymbol{O}$ is orthogonal.

As for the constraint, we have

$$\boldsymbol{\Gamma}_i'^T \boldsymbol{A}_i \boldsymbol{\Gamma}_i' = \boldsymbol{O}^T \boldsymbol{\Gamma}_i^T \boldsymbol{A}_i \boldsymbol{\Gamma}_i \boldsymbol{O} = \boldsymbol{O}\boldsymbol{O}^T = \mathbf{1}, \tag{27}$$

where Eq. (23) and the orthogonality of $\boldsymbol{O}$ was used. □

The only possibility to get rid of this rotational degeneracy, and fix the concept vectors unambiguously is to add some nonlinearity to the model. We can introduce a further cost term (complexity or structural cost) associated with the distribution of coefficients connecting the concepts to the physical attributes. A standard choice is

$$Q(\boldsymbol{\Gamma}_i) = -\epsilon \sum_\mu q(\boldsymbol{\gamma}_{i\mu}), \qquad q(\boldsymbol{\gamma}) = \sum_{d=1}^{D} \gamma_d^4, \tag{28}$$

but any similar nonlinear function could do just as well. In Eq. (28) ϵ is positive and infinitesimally small, thus it does not deform the subspace itself. Such a cost eliminates the artificial degeneracy arising from the linearity of our model and can be used to select from otherwise degenerate configurations. A possible interpretation is that it is easier to evaluate a concept, which only depends on a small number of relevant attributes, than one, which has more or less equivalent "loadings" on many. In neurophysiological terms a smaller number of synapses are required to approximate this representation. The weak links can be cut without committing large error, thus instead of working with a fully connected (semantic) network, a sparse network can be used as a good approximation. The specific form of Q in Eq. (28) is the standard "quartimax" rotation criterium, widely applied in the theory of PCA and factor analysis [20]. In fact, most of our results can be presented without considering such a nonlinear term – the exceptions will be pointed out in due course.

3.4 Measuring the coherence of meanings

When agents' preferences and probability densities for alternatives differ they will end up using concepts with different meanings. However, as the importance (benefit) of communication increases (c increases), there is a pressure on agents to share coherent meanings. The coherence will only be perfect at $c = \infty$. For a quantitative measure of coherence we introduce two definitions.

Definition 1 (Average meaning and coherence of concepts) *The* average meaning *of concept μ is the population average of the individual concept vectors*

$$\bar{\boldsymbol{\gamma}}_\mu = \frac{1}{I} \sum_{j=1}^{I} \boldsymbol{\gamma}_{j\mu}. \tag{29}$$

The measure of its coherence *is the length, $|\bar{\boldsymbol{\gamma}}_\mu|$.*

When coherence is perfect, i.e., all agents have identical concept vectors for concept μ the coherence (order-) parameter has unit length $|\bar{\gamma}_\mu| = 1$. In the opposite limit $|\bar{\gamma}_\mu| = 0$ we can speak about complete disorder.

It is also useful to define a scalar parameter, which measures the *overall* coherence of the language as a whole.

Definition 2 (Coherence of language) *The overall coherence L of a language is defined to be*

$$L(\mathbf{\Gamma}_{\rm all}) = \left(\frac{1}{I^2} \sum_{i}^{I} \sum_{j>i}^{I} {\rm Tr}\left(\mathbf{\Gamma}_i^T \mathbf{\Gamma}_j \right) \right)^{1/2} . \tag{30}$$

Note that in the limit of large populations, $I \to \infty$, L can be written as a quadratic function of the concept coherences,

$$L(\mathbf{\Gamma}_{\rm all}) = \left(\frac{1}{2} \sum_{\mu=1}^{K} \bar{\gamma}_\mu^2 \right)^{1/2} + \mathcal{O}\left(\frac{1}{I} \right), \tag{31}$$

and, thus, measures naturally the coherence of all the concept vectors in the language.

As it was discussed above, $U_i^{\rm REP} + U_i^{\rm COM}$ is invariant for a collective rotation of the concept vectors, and the infinitesimal nonlinear term Q was introduced to lift this degeneracy. It is important to note that the $\bar{\gamma}_\mu$ vectors are not invariant for such collective rotations, but L as defined in Eq. (30) is.

Lemma 3 *An identical collective rotation of the concept vectors for all agents*

$$\forall\, i \quad \mathbf{\Gamma}_i' = \mathbf{\Gamma}_i \boldsymbol{O}, \qquad \boldsymbol{O}\boldsymbol{O}^T = \mathbf{1}, \tag{32}$$

leaves L invariant,

$$L(\mathbf{\Gamma}_{\rm all}) = L(\mathbf{\Gamma}'_{\rm all}). \tag{33}$$

Proof: As seen in the proof of Lemma 2, each term in Eq. (30) is invariant in itself. □

The fact that collective rotations leave the agents' utilities and the overall coherence of language invariant asks for an interpretation of the invariant subspace spanned by the concept vectors. It is tempting to interpret this subspace as "culture" [21, 22]. Effectively, this interpretation says that we can call two agents culturally identical if one can predict exactly the behavior of the other in all possible decision problems. This doesn't mean that agents would make identical decisions, since their preferences may be different (heterogeneity). However, if their concept subspaces are identical, they can understand/predict each other accurately. In contrast, if the subspaces are misaligned, there is always some prediction error (misunderstanding or cultural incommensurability) between the agents. This is a useful working definition of "culture" because it allows – as in the real world – for the existence of different "languages" within the same culture, i.e., different basis vectors spanning the same subspace. The individual concepts are different across these languages but alternatives can be described identically in each. One could consider a stricter definition of culture, which requires the identity of preferences as well. With this stricter definition however, the left and right political parties in a country would belong to different cultures, which is a somewhat uncomfortable interpretation. The definition of culture has to allow for different preferences.

4 Equilibrium Languages

4.1 Single agent or identical agents

Let us first investigate the properties of our model Eq. (22-23) in the case when there is no communication between agents, i.e., $c = 0$. This is the problem of isolated agents who develop a mental representation of their world on their own. It also corresponds to the case when agents are identical. A solution which maximizes U_i under the constraint is trivially a Nash equilibrium of the Language Game. This limit can serve as a benchmark in the analysis of the more interesting case when multiple, heterogeneous agents interact.

Proposition 1 *Without social interactions the optimal (equilibrium) language solves a Principal Component Analysis (PCA) problem. The optimal concepts span the most significant subspace of the world matrix* $\boldsymbol{W}_i = \boldsymbol{B}_i\boldsymbol{A}_i$*, and within this subspace minimize the nonlinear cost function* Q *(if present).*

Proof: Let us consider first the case $Q = 0$. In the lack of interactions the utility of agent i reads

$$U_i(\boldsymbol{\Gamma}_i) \;\; = \;\; \mathrm{Tr}\left(\boldsymbol{\Gamma}_i^T\boldsymbol{A}_i^T\boldsymbol{W}_i\boldsymbol{\Gamma}_i\right) - \mathrm{Tr}\left(\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{A}_i\,\boldsymbol{\Gamma}_i\boldsymbol{\Lambda}\right). \tag{34}$$

where the second term is the constraint added with Lagrange multipliers. The $K \times K$ symmetric matrix $\boldsymbol{\Lambda}$ is a compact form of the Lagrange multipliers for all the $K(K+1)/2$ components of the constraint. The maximization problem for $\boldsymbol{\Gamma}_i$ in Eq. (34) is a Principal Component Analysis (PCA) problem [23]. Indeed, varying with respect to $\boldsymbol{\Gamma}^T$ we obtain the condition of extremum

$$\boldsymbol{A}_i^T\boldsymbol{W}_i\boldsymbol{\Gamma}_i = \boldsymbol{A}_i\,\boldsymbol{\Gamma}_i\boldsymbol{\Lambda}. \tag{35}$$

Assuming that the symmetric covariance matrix $\boldsymbol{A}_i = \boldsymbol{A}_i^T$ is nonsingular and thus invertible, Eq. (35) is equivalent to

$$\boldsymbol{W}_i\boldsymbol{\Gamma}_i = \boldsymbol{\Gamma}_i\boldsymbol{\Lambda}. \tag{36}$$

This latter states that the K-dimensional subspace spanned by the concept vectors (language matrix) is an invariant subspace of the world matrix $\boldsymbol{W}_i$. The only K-dimensional invariant subspaces are the ones spanned by K of the eigenvectors of $\boldsymbol{W}_i$. The remaining question is how to choose the eigenvectors to maximize the utility.

Let $\lambda_1 \geq \ldots \geq \lambda_n \geq \ldots \geq \lambda_D \geq 0$ denote the eigenvalues of the world matrix in decreasing order, and $\boldsymbol{w}_n$, $n = 1, \ldots, D$, the associated eigenvectors, $\boldsymbol{W}_i\boldsymbol{w}_n = \lambda_n\boldsymbol{w}_n$. The above ordering is possible, since the eigenvalues of $\boldsymbol{W}_i$ are all real and non-negative. (The fact that both $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are symmetric, positive definite is enough to prove this.) If the subspace is spanned by the eigenvectors $\boldsymbol{w}_{n_1}, \boldsymbol{w}_{n_2}, \ldots, \boldsymbol{w}_{n_K}$, i.e., the language matrix $\boldsymbol{\Gamma}$ is constructed from these vectors as columns, the utility in Eq. (18) becomes

$$U_i^{\mathrm{REP}} = \mathrm{Tr}\left(\boldsymbol{\Gamma}_i^T\boldsymbol{A}_i^T\boldsymbol{W}_i\boldsymbol{\Gamma}_i\right) = \mathrm{Tr}\left(\boldsymbol{\Lambda}\right) = \sum_{\mu=1}^{K}\lambda_{n_\mu}, \tag{37}$$

where we have used Eq. (36). This is maximal if $n_\mu = \mu$, i.e., if the eigenvectors chosen in the language matrix are the ones with the K largest eigenvalues. Thus the utility

maximizing language arises as the PCA problem of the world matrix. Having the optimal concepts determined, the mental weights of the representation adapt according to Eq. (9).

Assuming that $\boldsymbol{W}_i$ is full rank, all the eigenvalues are positive, and for $K < D$ the optimal utility $U_i^{\rm opt}$ is strictly below the theoretical maximum ${\rm Tr}\boldsymbol{W}_i = \sum_{\mu=1}^{D} \lambda_{i\mu}$. The representation error due to the reduction of dimensionality is the weight of omitted eigenvectors

$$E_i^{\rm REP} = \sum_{\mu=K+1}^{D} \lambda_{i\mu}. \tag{38}$$

The particular solution we have found is not the only solution which maximizes $U_i^{\rm REP}$ and complies with the constraints. As shown by Lemma 2, any rigid rotation within the subspace spanned by these eigenvectors produces another degenerate solution. (Note that since $K < D$, any configuration that only differs from the reference configuration by a relabeling of the basis vectors or by sign flips of some of the vectors can also be attained by a suitable rigid rotation.) This infinite degeneracy is, however, lifted by the nonlinear cost term Q, when it is added. When Q is infinitesimally small, the solution remains in the "most significant subspace", but the basis vectors get determined, at least up to discrete transformations such as sign flips and relabeling. For finite Q even the subspace gets deformed, nevertheless the nonlinearity helps fixing the optimal concepts in an unambiguous manner. □

The PCA-based optimal mental representation discussed above may be reached by practically any (myopic) utility maximizing learning dynamics. Note that without agent-agent interactions this is a (constrained) optimization problem on a fixed landscape which is quadratic (without the Q term) and thus smooth. There is no danger that the dynamics get stuck in suboptimal local maxima [24].

4.2 Nash equilibria for multiple agents

Let us investigate now agent interactions, and set $c > 0$ and $I > 1$. Communication between agents will deform their individual mental representations away from the PCA solution. The first question we ask is whether this coupled and constrained system has any Nash equilibrium. The proof of existence will boil down to the following fundamental observation:

Proposition 2 *The Language Game is a potential game. There exists a multi-agent potential $V : \mathbb{R}^{IKD} \to \mathbb{R}$ such that for any two strategy configurations $(\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}_{-i})$ and $(\boldsymbol{\Gamma}'_i, \boldsymbol{\Gamma}_{-i})$ we have*

$$U_i(\boldsymbol{\Gamma}'_i, \boldsymbol{\Gamma}_{-i}) - U_i(\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}_{-i}) = V(\boldsymbol{\Gamma}'_i, \boldsymbol{\Gamma}_{-i}) - V(\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}_{-i}). \tag{39}$$

Proof: The proof goes by an explicit construction of the potential

$$\begin{aligned} V(\boldsymbol{\Gamma}_{\rm all}) &= \sum_{i=1}^{I} \left[U_i^{\rm REP}(\boldsymbol{\Gamma}_i) - Q(\boldsymbol{\Gamma}_i) \right] + \sum_{i=1}^{I} \sum_{j>i}^{I} U_{ij}^{\rm COM}(\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}_j) \\ &= \sum_{i=1}^{I} \left[{\rm Tr}\left(\boldsymbol{\Gamma}_i^{\rm T} \boldsymbol{W}_i \boldsymbol{\Gamma}_i \right) - Q(\boldsymbol{\Gamma}_i) \right] + \frac{c}{I-1} \sum_{i=1}^{I} \sum_{j>i}^{I} {\rm Tr}\left(\boldsymbol{\Gamma}_i^{\rm T} \boldsymbol{\Gamma}_j \right), \end{aligned} \tag{40}$$

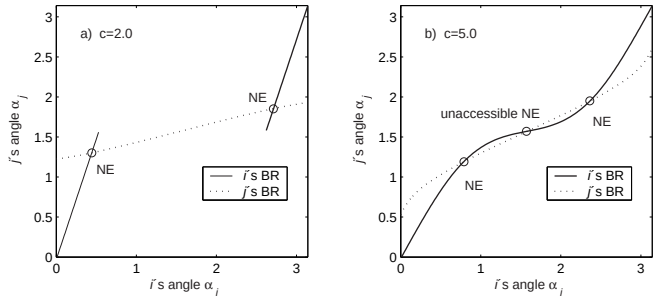

Figure 2: An example: best response (BR) curves for $D = 2$, $X = 1$, $K = 1$ and two agents i and j. The preference vectors are $\boldsymbol{\omega}_i = [1, 0]$, $\boldsymbol{\omega}_j = [0, 1.2]$ and the agents' concept vectors are parameterized by the polar angles $0 \leq \alpha_i, \alpha_j \leq \pi$, resp., $\boldsymbol{\gamma}_i = [\cos\alpha_i, \sin\alpha_i]$, $\boldsymbol{\gamma}_j = [\cos\alpha_j, \sin\alpha_j]$. (a) Two NEs for $c = 2.0$; (b) Three NEs for $c = 5.0$. The one in the middle is a special saddle point of V, and thus unaccessible dynamically.

where we assume that the nonlinear cost term Q is also present in the utility. Obviously the single agent terms are identical on the left and right side of Eq. (39). As for the two-agent terms (COM), the equality holds provided that the interaction is symmetric, i.e., $U_{ij}^{\rm COM} = U_{ji}^{\rm COM}$, which is satisfied by the form in Eq. (21). $\square$

A potential can always be constructed if a game is based on symmetric pair interactions [25]. These games are sometimes also called *partnership games* [26]. Note that it follows from the general expression Eq. (39) that the differential forms

$$\frac{\partial U_i}{\partial \boldsymbol{\Gamma}_i} = \frac{\partial V}{\partial \boldsymbol{\Gamma}_i}, \qquad \frac{\partial^2 U_i}{\partial \boldsymbol{\Gamma}_i \partial \boldsymbol{\Gamma}_j} = \frac{\partial^2 V}{\partial \boldsymbol{\Gamma}_i \partial \boldsymbol{\Gamma}_j} \tag{41}$$

also hold for any i and j. This becomes useful in the sequel. Now it is easy to prove that:

Proposition 3 *The Language Game always has at least one Nash equilibrium.*

Proof: The constraints in Eq. (23) make the multi-agent potential V have compact support, and as such it necessarily takes its global maximum at a point $\boldsymbol{\Gamma}^*_{\rm all} = (\boldsymbol{\Gamma}^*_1, \ldots, \boldsymbol{\Gamma}^*_I)$. (Again, without Q the global maximum would be infinitely degenerate, but Q resolves this degeneracy.) It is easy to see, however, that $\boldsymbol{\Gamma}^*_{\rm all}$ is a Nash equilibrium. Indeed, no agent has an incentive to deviate from this by choosing $\boldsymbol{\Gamma}_i \neq \boldsymbol{\Gamma}^*_i$, since

$$U_i(\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}^*_{-i}) - U_i(\boldsymbol{\Gamma}^*_i, \boldsymbol{\Gamma}^*_{-i}) = V(\{\boldsymbol{\Gamma}_i, \boldsymbol{\Gamma}^*_{-i}\}) - V(\{\boldsymbol{\Gamma}^*_i, \boldsymbol{\Gamma}^*_{-i}\}) \leq 0, \tag{42}$$

where the first equality is assured by Proposition 2, and the second inequality by the fact that $(\boldsymbol{\Gamma}^*_i, \boldsymbol{\Gamma}^*_{-i})$ is a global maximum of V. $\square$

In general, the Nash equilibrium (NE) is not unique. It is typical to have configurations, which are NEs, although they do not maximize the potential V. It is obvious, however, by Eq. (41) that they should necessarily correspond to local extrema of V. This condition is of course not sufficient. Not all local extrema of V are NEs. An example showing the appearance of more than one NEs is shown in Fig. (2) for a simple case: $D = 2$, $X = 1$, $K = 1$ and two agents, $I = 2$. Nash equilibria, which correspond to saddle points of V cannot be accessed dynamically (see later).

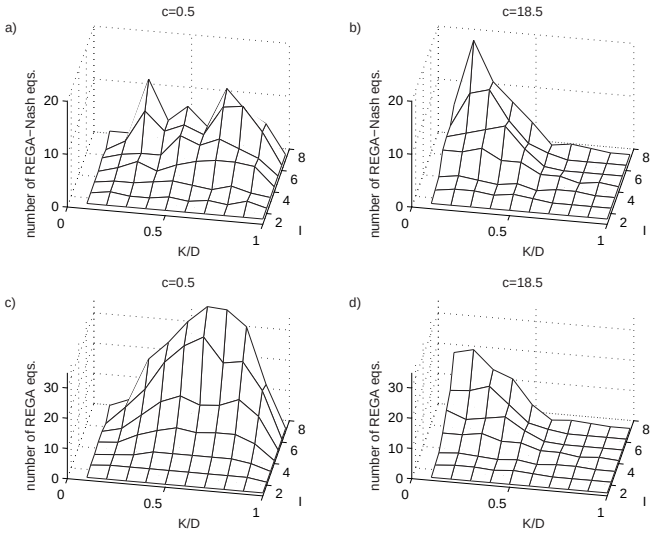

Figure 3: Average number of dynamically accessible Nash equilibria (REGA-Nash equilibria) and stable fixed points (not necessarily NE) of the adjustment dynamics (REGA equilibria) as a function of the relative intelligence K/D and the number of agents I. Subplots (a,c) refer to small communication rate, $c = 0.5$; (b,d) to large communication rate, $c = 18.5$. In the simulations the agent preferences are assumed to be iid Gaussian random vectors, and the structure of alternatives is set $\boldsymbol{A}_i = \mathbf{1}$.

Similar examples can be constructed in higher dimensions and for more agents. The stability of these NEs with respect to the game dynamics will be investigated in Section 5. Some of these NEs turn out to be unstable and thus inaccessible under reasonable evolutionary dynamics. (The possible existence of dynamically inaccessible NEs is a well-known fact of evolutionary game theory. The Folk theorem of evolutionary game theory asserts that under a wide class of dynamics all attractors are NEs, but the converse do not hold. See Cressman [27] and Hofbauer and Sigmund [28] for a formal discussion.) However, even after the omission of these, the non-uniqueness of language equilibria prevails. Thus, our model predicts a strong path-dependence in language evolution, in which the timing and ordering of the appearance of new contexts can play a significant role, and which can be a source of cultural heterogeneity.

The number of dynamically accessible Nash equilibria (to be called REGA-Nash equilibria in the sequel) is a function of the basic model parameters such as the number of agents, the number of concepts, the communication strength, the (heterogeneous) world matrices the agents possess, etc. Figure 3(a-b) illustrates the case when the world matrices are randomly distributed in the population. Random agent properties imply that the number of equilibria is a random variable too. Figure 3(a-b) plots the average number of equilibria determined by a series of simulations. Although there remain considerable fluctuations in the points, especially for small c, the overall picture is rather clear. The number of Nash equilibria can be rather large and typically increases rapidly with the number of agents. This creates a severe coordination problem. An exception for this rule is the region of large c and large K/D. This is the limit when agents are highly intelligent and benefit a lot from communication. In this region we have found that it is typical to have only one or two Nash equilibria, even when the number of agents is rather large. Note that a large communication benefit in itself is not enough

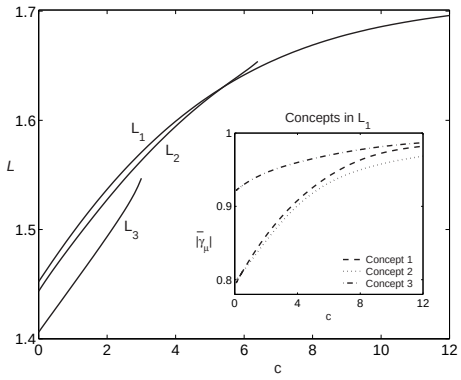

Figure 4: Coherence of language L and individual concepts $|\bar{\gamma}_\mu|$ as a function of the communication strength c for $D = X = 8$, $K = 3$, $I = 4$, and randomly generated, then fixed preferences. L_1, L_2 and L_3 are three different language equilibria appearing for these parameters. Inset shows the coherence of concepts for language L_1.

to reduce the number of NEs, since this only implies that concept vectors of different agents be highly parallel without explicitly defining what they should be. In the case of $K << D$ (low intelligence), the number of Nash equilibria seems to proliferate even when agent utility is dominated by the communication benefit. Our simulation results, although very limited in scope, may indicate a qualitative change in the behavior of the model ("a phase transition") on the phase plain intelligence vs communication rate – the investigation of which is left for a future study.

Many of the equilibria we have found only exist in a certain range of c, and disappear (become unstable) at some critical values. Figure 4(a) demonstrates this by showing the overall language coherence $L = L(c)$ in equilibrium. All curves in the figure denote a different Nash equilibrium. As is seen from Fig. 4(a), the coherence in each equilibrium increases monotonically as a function of c. This is intuitive, since we expect that more communication, i.e., a higher value for the external communication rate, c implies enhanced coherence of the language utilized. In fact we can prove the following:

Proposition 4 *In (each possible) equilibrium the overall coherence $L = L(c)$ of the language is a monotonically increasing function of the communication rate, c.*

Proof: Let us rewrite the interaction part of the potential with the help of L^2. Applying Eq. (30) we get

$$\frac{c}{I}\sum_{i=1}^{I}\sum_{j>i}^{I}\mathrm{Tr}\left(\mathbf{\Gamma}_i^{\mathrm{T}}\mathbf{\Gamma}_j\right) = cI\,L^2. \tag{43}$$

With this the potential has the formal structure

$$V(\mathbf{\Gamma}_{\mathrm{all}}) = V_0(\mathbf{\Gamma}_{\mathrm{all}}) + cI\,L^2(\mathbf{\Gamma}_{\mathrm{all}}), \tag{44}$$

where V_0 represents all single-agent terms, and the c term collects all communication terms. At a dynamically accessible NE, $\mathbf{\Gamma}^*_{\mathrm{all}}$, the potential V necessarily takes its local maximum. As c changes the equilibrium configuration adapts analytically (except for bifurcation points, which are beyond our consideration). The two terms in V compete: for small c it is the maximum of V_0 which determines $\mathbf{\Gamma}^*_{\mathrm{all}}$, whereas for large c it is

the second term. As c increases the balance of importance swings towards the second term, and hence $\mathbf{\Gamma}^*_{\text{all}}$ gets closer and closer to the individual maximizing configuration of this term, meaning that L^{2*} monotonically increases. The rigorous formulation of this argument is delegated, in form of a lemma, to the Appendix. □

Proposition 4 ensures that language as a whole becomes more coherent when the rate of communication increases, at least until the equilibrium (which is followed analytically as a function of c) exists. However, as Fig. 4 illustrates the landscape structure of V can be such that local maxima arise and disappear by varying c. Thus, certain language equilibria may lose stability and disappear in a bifurcation process for some critical value of c. This occurs to L_2 and L_3 in the example presented in Fig. 4.

Although we have found that language as a whole becomes more coherent for increasing communication rate, it is not obviously true for every single concept in the language. The system is coupled in an intimate way, and the concepts themselves get determined by the nonlinear term Q. Even though we could not prove this rigorously, in all our simulations we have found that for all μ, $\bar{\gamma}_\mu$ increases monotonically with c (see the inset of Fig. 4 for an illustration).

5 Social Dynamics

So far, we have only discussed the Nash equilibria of the Language Game. However, we can view language as a dynamic, evolutionary problem in which agents perpetually adapt their mental representations to the changing environment and to each other. Adaptation occurs through a trial-and-error procedure in which the test configuration the agent considers necessarily deviates from his actual (reference) representation. Since the parameter space is enormous, a large (random) change very likely makes the representation worse, and thus will be rejected eventually. It is reasonable to assume a search heuristic, which concentrates on small (local) deformations and a set of obvious discrete transformations, while ad hoc, larger scale deformations of the representation are only tested very rarely. Evolutionary dynamics acts as an equilibrium selection method, which can solve the coordination problem related to multiple equilibria found above.

5.1 REGA dynamics

In the following, we assume a myopic adjustment dynamics in which agents slowly deform their concepts in order to optimize them in a local sense to the natural (perceptual) and social environment, maintaining the assumed constraints. A possible continuous time evolution in this spirit is along the steepest ascent of the utility (*gradient adjustment dynamics*), which reads

$$\frac{\delta \mathbf{\Gamma}_i}{\delta t} = \text{const}\, \boldsymbol{P}_i \frac{\partial U_i}{\partial \mathbf{\Gamma}_i}, \tag{45}$$

where $\boldsymbol{P}_i$ is an adequate projector, which projects the bare gradient $\partial U_i/\partial \mathbf{\Gamma}_i$ (meant by components) into the tangent space allowed by the constraint Eq. (23). It is customary to call $\boldsymbol{P}_i \partial U_i/\partial \mathbf{\Gamma}_i$ the *projected gradient*, which assures that $\mathbf{\Gamma}_i(t)$ continues to respect the constraints for all t.

The gradient dynamics is a local search heuristic, which continuously deforms the concept vectors. However, as dictated by the nature of the problem, it seems reasonable to complement this continuous dynamics with a very specific discrete part, namely *sign*

flips, $\boldsymbol{\gamma}_{i\mu} \to -\boldsymbol{\gamma}_{i\mu}$, and *relabeling transformations*, $\boldsymbol{\gamma}_{i\,\mu} \to \boldsymbol{\gamma}_{i\,R\mu}$, where R is a permutation operator for concepts. Sign flips correspond to using, for instance, "fastness" instead of "slowness", or a (directed) South-North axis instead of a North-South axis. Relabeling, in turn, permutes the associations, which relate concepts to signals used in communication. Configurations attainable by such discrete transformations are readily available for the (boundedly rational) agent, and are assumed to be tested perpetually with some finite probability during gradient adjustment. The flipped or permuted configuration is accepted and replaces the reference configuration, if it increases the agent's utility, whereas it is discarded and the continuous part of the dynamics continues with the reference configuration, if not.

The basic idea is that, as concepts slowly deform due to continuous adjustment, the agent may realize that a proper sign flip or relabeling can vastly improve his/her communication efficiency while leaving his representation error intact. The above discrete processes, which will be referred to collectively as *rematching transformations* (rematching concepts and their linguistic signals used in communication), help to avoid spurious configurations/fixed points, which could be amended trivially by adequately permuting (relabeling) the player's concepts. We will refer to the above dynamics (continuous and discrete parts together) as the "rematching enabled gradient adjustment" (REGA) dynamics. Note that during a REGA iteration step the K-dimensional language subspace only changes slightly. Although rematching transformations make seemingly large configuration changes, they keep the subspace invariant. As such, these rematching transformations can still be considered "myopic" adjustments because they occur within the K-dimensional subspace defined by the concept vectors.

5.2 Fixed points and stability

Let us consider now the potential fixed points of the REGA dynamics.

Proposition 5 *From all initial conditions the REGA dynamics of the Language Game converges to a fixed point.*

Proof: The multi-agent potential V acts as a Lyapunov function for REGA, in the sense that V increases in all iteration steps. This is trivial for the continuous (gradient) part, and also holds for the discrete (rematching) part by Eq. (39). As such the $t \to \infty$ limit of the dynamics is necessarily a fixed point. □

Fixed points of the REGA dynamics can be sorted according to their stability properties. In particular, we define REGA equilibria, as the *stable* fixed points of the dynamics:

Definition 3 (REGA Equilibrium) *A certain choice of concept vectors by agents in the society will be called a REGA equilibrium if this configuration is (1) a fixed point of the REGA dynamics, (2) it is asymptotically stable against infinitesimal individual and collective deviations, and (3) stable against individual and collective rematching transformations.*

Proposition 6 *There exists at least one REGA equilibrium of the Language Game.*

Proof: We are going to prove that the global maximum $\boldsymbol{\Gamma}^*_{\text{all}}$ of the multi-agent potential V defined in Eq. (40) satisfies the definition of the REGA equilibrium.

By definition a REGA equilibrium is a fixed points of the REGA dynamics, i.e., a point where the projected gradient $\boldsymbol{P}_i \partial U_i / \partial \boldsymbol{\Gamma}_i$ vanishes for all i. However, by Eq. (41)

this condition is equivalent to the first order condition for local maximum of V. The global maximum $\mathbf{\Gamma}^*_{\rm all}$ of V, which necessarily exists since the support of V is compact, is a point which satisfies this. This is a fixed point of the REGA dynamics. The global potential V may have several local extrema, which are all fixed points of the REGA dynamics. Some of these may be asymptotically stable some may be unstable with respect to infinitesimal deviations. Only the stable fixed points are attainable by the evolutionary dynamics, and only these will be called REGA equilibria.

In order to prove the local stability of $\mathbf{\Gamma}^*_{\rm all}$ against infinitesimally small, not necessarily unilateral but supposedly collective deviations, we have to show that the associated *bordered Hessian* is negative semi-definite at that point [29]. The bordered Hessian is a supermatrix, composed from the second derivatives of the individual utilities $\partial^2 U_i/\partial\mathbf{\Gamma}_i\partial\mathbf{\Gamma}_j$ as a submatrix (the ordinary Hessian), and submatrices formed by the first derivatives of the constraints. However, as follows from Eq. (41), the Hessian piece is the same as the Hessian of the global potential. It follows that the bordered Hessian in question is identical to the bordered Hessian of the potential at $\mathbf{\Gamma}^*_{\rm all}$. Since $\mathbf{\Gamma}^*_{\rm all}$ is the (global) maximum of the potential (satisfying all constraints) its bordered Hessian in necessarily negative semi-definite, proving the assertion.

Lastly, we have to prove that no player can improve his utility by any rematching of his concepts. Let $\tilde{\mathbf{\Gamma}}^*_{\rm all}$ denote a configuration obtained by arbitrary, independent rematching of the concepts. Since $\mathbf{\Gamma}^*_{\rm all}$ is the global maximum rematching cannot improve the global potential

$$V(\tilde{\mathbf{\Gamma}}^*_{\rm all}) \leq V(\mathbf{\Gamma}^*_{\rm all}). \tag{46}$$

On the other hand, as rematching leaves invariant the representation error $U_i^{\rm REP}(\tilde{\mathbf{\Gamma}}^*_i) = U_i^{\rm REP}(\mathbf{\Gamma}^*_i)$ and the complexity cost $Q_i(\tilde{\mathbf{\Gamma}}^*_i) = Q_i(\mathbf{\Gamma}^*_i)$ we can write

$$\begin{aligned} U_i(\tilde{\mathbf{\Gamma}}^*_i, \mathbf{\Gamma}^*_{-i}) - U_i(\mathbf{\Gamma}^*_i, \mathbf{\Gamma}^*_{-i}) &= U_i^{\rm COM}(\tilde{\mathbf{\Gamma}}^*_i, \mathbf{\Gamma}^*_{-i}) - U_i^{\rm COM}(\mathbf{\Gamma}^*_i, \mathbf{\Gamma}^*_{-i}) = \\ & V(\tilde{\mathbf{\Gamma}}^*_{\rm all}) - V(\mathbf{\Gamma}^*_{\rm all}) \leq 0 \end{aligned} \tag{47}$$

where for the last inequality we have used Eq. (46). □

The REGA equilibrium concept differs from the Nash equilibrium concept in two respects. First, a huge amount of otherwise possible global strategy options are excluded from the ones "tested" due to bounded rationality, and second, the equilibrium is required to be stable against collective local deviations too, and not just against unilateral deviations. The rationale behind excluding most of the parameter space from the accessible strategies is its practical infinity with respect to the capacities of the human mind, $K/D \to 0$. Search for better response is necessarily heuristic, focusing essentially on local improvements. The emergence of drastically new successful concepts is necessarily a non-systematic, trial and error mechanism, which has much lower probability.

Note also that none of the two equilibrium concepts implies the other. There are Nash equilibria which are fixed points of the dynamics, but not REGA equilibria because they are unstable against collective deviations – examples are shown in Fig. 2(b). Such points are necessarily saddle points, which cannot be attained dynamically from generic initial conditions. Being exactly on the stable manifold of a saddle point has zero probability, and even if this were the case initially, any noise in the dynamics would finally drive the system away from such a fixed point.

Corollary 1 *A Nash equilibrium of the Language Game that is not a REGA equilibrium at the same time (Nash-only equilibrium), is dynamically inaccessible and thus, cannot be interpreted as language.*

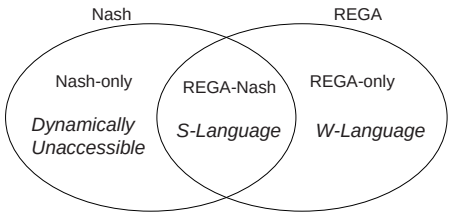

Figure 5: Equilibria and their interpretation as language.

As such, we should focus on REGA equilibria. As demonstrated by the simulation results of Fig. 3(c,d) their number can be rather high. These can additionally satisfy the requirements for a Nash equilibrium too, but this is not necessary. REGA equilibria, which are not Nash equilibria (REGA-only) are unstable against large-scale (global) unilateral deviations. For such fixed points there are agents who would have better response than the actual. However, as assumed above, locating a better response globally is a low-probability event due to the large barrier separating the two configurations. Consequently, the system can stay in such a "metastable" equilibrium for a very long time.

Definition 4 (Weak form of language) *A (metastable) REGA-only equilibrium of the Language Game is called a "Weak form of language" (W-language).*

Finally, equilibria which are both REGA and Nash (REGA-Nash equilibria) have superior stability properties and play a distinguished role. The following definition formally describes such an equilibrium and a proposition states their existence in the Language Game.

Definition 5 (Strong form of language) *A REGA-Nash equilibrium of the Language Game is called a "Strong form of language" (S-language).*

Proposition 7 *The Language Game has at least one REGA-Nash equilibrium (S-language).*

Proof: In the proof of Proposition 3 we have seen that $\mathbf{\Gamma}^*_{\rm all}$, the global maximum of the multi-agent potential V, is a Nash equilibrium, and at Proposition 6 that it is a REGA equilibrium. Consequently, $\mathbf{\Gamma}^*_{\rm all}$ satisfies the requirements of the proposition. □

The different kinds of equilibria, their relationship, and interpretation as language is depicted in Fig. 5. As our simulation results indicate there can be a large number of REGA-Nash equilibria, especially when the population is very heterogeneous and communication has little significance (c small). See Fig. 3 for a quantitative analysis. For less heterogeneity and larger communication strength we have found that almost all REGA equilibria are indeed REGA-Nash equilibria.

The stability properties of the different kinds of equilibria are summarized in Fig. 6. The possible perturbations are categorized according to their scale (local vs. global) indicating the size of the deviation tested, and according to the number of agents they involve (single agent vs. multiple agents). By definition, Nash equilibria are stable against single agent perturbations irrespective whether they are small (local) or large (global). Also, our definition of the REGA equilibrium implies stability with respect to any kind of (single or multiple agent) small (i.e., local) perturbations. As we see REGA-Nash equilibria are the most stable, only their multi-agent global stability remains undetermined.

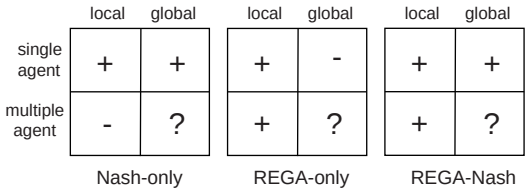

Figure 6: Stability properties of equilibria.

6 Discussion and Conclusion

We have presented a model based on the assumption that agents make choices between alternatives by first describing them on a finite set of concepts. In turn, agents will have an incentive to promote concepts that better fit their preferences. This feature of our model resonates to an old problem in linguistics and philosophy called *linguistic relativism*, which essentially states that the language we speak will influence (or even determine) our thoughts and judgements. As Whorf [30] writes: "We dissect nature along lines laid down by our native language." (See also Wittgenstein [31].) The debate is not whether linguistic relativism is true or false but rather to what extent it is true. Quoting Paul Kay, Ross [32] advances a moderate view: "[Although it may be correct that the languages people speak mold their thought] It is unlikely that the various languages of the world are so different that the ways their speakers think is incommensurable".

Our framework helps predict when linguistic relativism may be important. In our model, language is determined by two key inputs: the structure of the physical world (captured in $\boldsymbol{A}$) and agents' subjective preferences (captured in $\boldsymbol{B}$). As we mentioned however, the three-layer structure of Figure 1 represents the formation of concepts at only one level of abstraction. If we were to study the formation of more abstract concepts, then the bottom layer (the $\boldsymbol{a}$-s) would be concepts themselves that emerged from a previous Language Game. This structure suggests sequential development of more abstract layers in language: simple concepts describing the physical environment emerge first and provide the basis for the evolution of more abstract concepts in a series of subsequent Language Games. The consequence is that linguistic relativism will depend on the concepts' level of abstraction or complexity. When concepts name simple objects that we all perceive identically (because of our biological design), the objective structure of reality will have a major influence on them, and this layer of the equilibrium language is likely to be similar across isolated societies. At higher levels of abstraction, the $\boldsymbol{a}$-s of our model are concepts themselves and as such are less likely to be identical across isolated groups of people as the path-dependence of the previous Language Game has already introduced idiosyncratic structure in them. As we move away from the actual perceptual basis for concepts, heterogenous preferences and the path-dependent nature of language evolution may lead to very different concepts across isolated groups of people. For example, while we have no problem naming furnitures across highly different cultures and make the translation easily between their languages in this domain, our notions of abstract things such as "God" or "ethics" are hard to translate across different cultures.

The well-known Gavagai problem nicely illustrates this point: a foreigner landing on an isolated island tries to infer the meaning of the word "gavagai" shouted by a local pointing to a rabbit crossing the path. While, *a priori* he has no reason to associate the

meaning of "gavagai" with rabbit (it could also be "dinner" or "white" among others) this is the most intuitive meaning. This is because the structure in the data is strong and our differences in terms of payoffs of alternatives are relatively small when it comes to describing the material world around us. Clearly, if the local were to say "gavagai" when pointing to a weird statue representing a creature combining human and animal features, the translation would be extremely hard, if at all possible.

Data in two domains, colors and numbers, also seems to be consistent with the above pattern. Color is a relatively concrete domain that is directly linked to perceptions, i.e., it has low level of abstraction. It turns out that natural languages differ wildly in the (number of) basic colors that they name. Yet, research shows (see, e.g., Ross [32]) that this does not lead to linguistic relativism: people remember and use colors to the same degree of sophistication across cultures. The situation seems to be the opposite when it comes to numbers, which represent a domain with an arguably higher level of abstraction. In a recent study by Peter Gordon of Columbia University [33], it is shown that people using languages that do not have words for numbers higher than 2 have problems comparing and remembering quantities. Linguistic relativism definitely applies in this - more abstract - domain. In sum, empirical patterns suggest that linguistic relativism is likely to be strong in abstract domains of language, while it is weak in domains that are closer to our perceptions, such as concepts describing our physical environment. Our model provides an explanation for why this may happen.

The model is also consistent with psychology's view on how concepts are structured in our mind, namely *categorization theory* and *graded structures* [16]. It is intuitive that language should reflect the structure of the world. This results from humans' innate capacity to categorize things based on similarity, resulting in categories with well defined hierarchies called graded structures. It is also known however, that categories are 'ad hoc' in the sense that similarities are evaluated in relation to goals. As such, objects may potentially belong to any category but with a different weight. In our model, concepts are built from reality in the same way: all components of $\boldsymbol{a}$ may contribute to any of the concepts but their weights will differ wildly depending on people's preferences (goals).

An interesting new insight from our model is the interpretation of the relationship between language and culture. We interpret language as the collection of concepts while culture is the subspace defined by the concepts. Different sets of concepts can define the same subspace, but concepts defining different subspaces cannot be mapped into each other without a large error. This structure suggests that cultural difference doesn't come from disagreement between groups of people or differences in preferences, but rather from the fact that cultures dissect the world along different dimensions. In other words, the concepts used by one culture cannot be mapped accurately into the concepts of the other (there is no, or little possibility for translation) resulting in poor communication across cultures. As mentioned earlier this is only likely for communication in abstract domains (e.g. ethics or esthetics). Furthermore, the model is also consistent with the fact that for any given culture, there might be dozens of languages with notable differences in their concepts. Similarly, the model also accounts for the fact, that individuals may slightly disagree on the meanings of the words within the same language. Despite these differences, communication is possible because choice alternatives can be identically described with multiple sets of concepts covering the same sub-space. In other words, accurate translation and, as a result efficient communication is possible across these languages.

Finally, our study of language's dynamic evolution also provides insights with regard

to empirical observations on the evolution of societies. Archeologists suspect that the so-called *cultural explosion* dating back to some 50 thousand years is somehow linked to the evolution of language (see, e.g., Mithen [34]). At about this date, human evolution seems to become much faster and proper cultures with abstract concepts (e.g. cults or religions) appear in the data. This pattern is consistent with our model. Our analysis of the Language Game's dynamics shows that concepts gradually emerge over time. Once an equilibrium is reached, it represents a consensus across the members of society, which is instantly available for the next generation. This consensus provides a layer for the development of more abstract concepts. While the speed of evolution before language was probably determined by the speed of biological evolution (as evidenced by the gradual increase in brain size, for example), after the cultural explosion, the speed of evolution is primarily determined by the speed of the dynamic *social process* governing the emergence of concepts. This speed may well exceed that of biological evolution.

As our goal was to develop a rather general theory of language, we had to introduce many technical simplifications. We considered a linear language system with well-defined constraints and we haven't explicitly modelled different levels of abstraction. In our analysis, we have only considered pure strategy equilibria, a strictly fixed number of concepts and we used one particular adjustment dynamics. In various extensions we have tried to explore the importance of some of these assumptions, while we argued for the strong validity of others. For example, we have ruled out mixed strategies as they cannot be interpreted for language. We have also explored other dynamics and found that our convergence results hold for other potential improving dynamics too as is generally believed (see Ermoliev and Flam [35], Hofbauer and Sigmund [26]). On a more general level, we have entirely ignored syntax from our analysis based on the broadly accepted argument that grammar is an "innate" capacity of humans, a result of biological evolution [36]. It is hard to imagine however, that the social evolution of concepts happens independently from grammar. We leave this and other interesting questions related to language for future research. However, we believe that to the extent language formation/usage is a social process, economics is likely to have an important role in explaining the related phenomena.

Appendix: Proof of Proposition 4

The proof boils down to the following Lemma:

Lemma 4 *Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ denote variables satisfying some constraints $g_j(\boldsymbol{x}) = 0$, $j = 1, \ldots, m$. Let $f(\boldsymbol{x}, c) = a(\boldsymbol{x}) + c\, b(\boldsymbol{x})$, with a, b continuously differentiable and $c \in \mathbb{R}$ a parameter, be a function that takes its local maximum at a point $\boldsymbol{x}^* = \boldsymbol{x}^*(c) = \operatorname{argmax}_x f(\boldsymbol{x}, c)$. Introducing $b^* = b(\boldsymbol{x}^*(c))$, we have*

$$\frac{db^*}{dc} \geq 0. \tag{48}$$

Proof: Let us introduce a set of lagrange multipliers $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^m$ to treat the constraints, and define the Lagrangian function

$$L(\boldsymbol{\lambda}, \boldsymbol{x}) = a(\boldsymbol{x}) + c\, b(\boldsymbol{x}) + \sum_j^m \lambda_j g_j(\boldsymbol{x}). \tag{49}$$

Using the chain rule we can write

$$\frac{db^*}{dc} = \frac{\partial b}{\partial \boldsymbol{x}^*} \cdot \frac{\partial \boldsymbol{x}^*}{\partial c}. \tag{50}$$

and the task is to calculate the vector $\partial \boldsymbol{x}^*/\partial c$. A straightforward way is to proceed by a series expansion around a point $c_0 \to \boldsymbol{x}_0^*, \boldsymbol{\lambda}_0^*$. When c is perturbed $c = c_0 + \epsilon$, the local maximum and the value of the Lagrange multipliers shift too, $\boldsymbol{x}^* = \boldsymbol{x}_0^* + \epsilon\,\boldsymbol{\alpha}$, $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_0^* + \epsilon\,\boldsymbol{\beta}$. In order to determine $\boldsymbol{\alpha} \equiv \partial \boldsymbol{x}^*/\partial c$ and $\boldsymbol{\beta}$, we introduce general deviation variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ as

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{x}_0^* + \epsilon\,\boldsymbol{\alpha} + \boldsymbol{\eta} \\ \boldsymbol{\lambda} &= \boldsymbol{\lambda}_0^* + \epsilon\,\boldsymbol{\beta} + \boldsymbol{\xi} \end{aligned} \tag{51}$$

and require that at the solution of the constrained optimization problem $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ should vanish; this will provide the necessary equations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

The Lagrangian can be expanded in a Taylor series in $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$. It is enough to consider terms $\mathcal{O}(\boldsymbol{\eta})$ and $\mathcal{O}(\boldsymbol{\xi})$ which should vanish. Moreover, these terms can be further expanded in ϵ. The coefficient of the $\mathcal{O}(\boldsymbol{\eta})$ term reads

$$\left[\frac{\partial a}{\partial \boldsymbol{x}} + c_0\,\frac{\partial b}{\partial \boldsymbol{x}} + \sum_j \lambda_{0j}^* \frac{\partial g_j}{\partial \boldsymbol{x}}\right] + \epsilon\Big[\frac{\partial b}{\partial \boldsymbol{x}} + \boldsymbol{G_x}\boldsymbol{\beta} + \boldsymbol{L_{xx}}\boldsymbol{\alpha}\Big] + \mathcal{O}(\epsilon^2) \tag{52}$$

where $[\boldsymbol{G_x}]_{ji} = \partial g_j/\partial x_i$ is the matrix of the first order derivatives of the constraints and $[\boldsymbol{L}]_{ii'} = \partial^2 L/\partial x_i \partial x_{i'}$ is the Hessian matrix at the point c_0. The first square bracket vanishes as this is the first order condition at c_0. The second square bracket implies

$$\frac{\partial b}{\partial \boldsymbol{x}} + \boldsymbol{G_x}\boldsymbol{\beta} + \boldsymbol{L_{xx}}\boldsymbol{\alpha} = 0. \tag{53}$$

Similarly, the coefficient of the $\mathcal{O}(\boldsymbol{\xi})$ term can be expanded in ϵ, and we obtain

$$\boldsymbol{G_x}\boldsymbol{\beta} = 0. \tag{54}$$

Equations (53) and (54) gives

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} = -\begin{pmatrix} \mathbf{0} & \boldsymbol{G_x} \\ \boldsymbol{G_x}^T & \boldsymbol{L_{xx}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \partial b/\partial \boldsymbol{x} \end{pmatrix} \tag{55}$$

where the hipermatrix is usually called the "bordered Hessian". This is necessarily negative semi-definite since we are at a local maximum. This allows us to express $\partial b^*/\partial c$ as a quadratic form

$$\frac{\partial b^*}{\partial c} = -\begin{pmatrix} \mathbf{0} \\ \partial b/\partial \boldsymbol{x} \end{pmatrix}^T \begin{pmatrix} \mathbf{0} & \boldsymbol{G_x} \\ \boldsymbol{G_x}^T & \boldsymbol{L_{xx}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \partial b/\partial \boldsymbol{x} \end{pmatrix} \geq 0 \tag{56}$$

which is thus necessarily non-negative. □

Proposition 4 follows as a direct corollary.

References

[1] Blume A., Kim Y.-G., Sobel J. Evolutionary stability in games of communication. *Games Econ. Behav.* 1993; 5(4):547–75.

[2] Blume A., DeJong D. V., Kim Y.-G., Sprinkle G. B. Experimental evidence on the evolution of meaning of messages in sender-receiver games. *Am. Econ. Rev.* 1998; 88(5):1323–40.

[3] Lachmann M., Számadó S., Bergstrom C. T. Cost and conflict in animal signals and human language. *P. Natl. Acad. Sci.* 2001; 98(23):13189–13194.

[4] Steels L. Self-organizing vocabularies. In Langton C., editor, *Proceedings of Artificial Life V* Nara, Japan 1996;.

[5] Nowak M. A., Plotkin J. B., Krakauer D. C. The evolutionary language game. *J. Theor. Biol.* 1999; 200:147–62.

[6] Nowak M. A. Evolutionary biology of language. *Phil. Trans. R. Soc. Lond. B* 2000; 355:1615–1622.

[7] Pinker S. Survival of the clearest. *Nature* 2000; 404:441–442.

[8] Rubinstein A. *Economics and Language.* Cambridge University Press Cambridge, UK 2000.

[9] Hurford J., Studdert-Kennedy M., Knight C. *Approaches to the Evolution of Language.* Cambridge University Press Cambridge 1998.

[10] Plotkin J. B. Nowak M. A. Language evolution and information theory. *J. Theor. Biol.* 1999; 205:147–59.

[11] Battigalli P. Maggi G. Rigidity, discretion and the cost of writing contracts. *Am. Econ. Rev.* 2002; 92:798–817.

[12] Cremer J., Garicano L., Prat A. Codes in organizations. Working Paper, MIT and CEPR 2003.

[13] Wernerfelt B. Organizational languages. *J. Econ. Manage. Strat.* 2004; 13(3):461–472.

[14] Armstrong S. L., Gleitman L. R., Gleitman H. On what some concepts might not be. *Cognition* 1983; 13:263–308.

[15] Croft W. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information.* Univ. of Chicago Press Chicago 1991.

[16] Barsalou L. Ad hoc categories. *Mem. Cognition* 1983; 11(3):211–227.

[17] Barsalou L. Ideals, central tendency and frequency of instantiation as determinants of graded structure in categories. *J. Exp. Psychol. Learn.* 1985; 11(4):629–654.

[18] Chater N. Vitnyi P. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* 2003; 7:19–22.

[19] Mérő L. *Ways of thinking. The limits of rational thought and artificial intelligence.* World Scientific New Jersey 1990.

[20] Jennrich R. I. A simple general procedure for orthogonal rotation. *Psychometrika* 2001; 66:289–306.

[21] Fáth G. Sarvary M. A renormalization group theory of cultural evolution. *Physica A* 2005; 348:611–629.

[22] Fáth G. Sarvary M. Cultural evolution in a population of heterogeneous agents. In Namatame A., Kaizouji T., Aruka Y., editors, *Economics and Heterogeneous Interacting Agents* Lect. Notes Econ. Math. Springer 2005;.

[23] Jolliffe I. T. *Principal Component Analysis.* Springer New York 1986.

[24] Baldi P. Hornik K. Learning in linear networks: a survey. *IEEE T. Neural Networ.* 1995; 6:837–858.

[25] Monderer D. Shapley L. S. Potential games. *Games Econ. Behav.* 1996; 14:124–143.

[26] Hofbauer J. Sigmund K. *Evolutionary Games and Population Dynamics.* Cambridge University Press Cambridge 1998.

[27] Cressman R. *Evolutionary dynamics and extensive form games.* MIT Press Cambridge, Mass. 2003.

[28] Hofbauer J. Sigmund K. Evolutionary game dynamics. *Bull. Am. Math. Soc.* 2003; 40(4):479–519.

[29] Simon C. Blume L. *Mathematics for Economists.* W. W. Norton & Company New York 1994.

[30] Whorf B. L. Science and linguistics. In Carroll J., editor, *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf* pages 213–14. MIT Press Cambridge, MA 1956;. (reprinted).

[31] Wittgenstein L. *Philosophical Investigations.* Blackwell Oxford 1953.

[32] Ross P. E. Draining the language out of color. *Sci. Am.* 2004; April.

[33] Gordon P. Numerical cognition without words: Evidence from amazonia. *Science* 2004; 306:496–499.

[34] Mithen S. *The Prehistory of the Mind.* Thames and Hudson London 1996.

[35] Ermoliev Y. M. Flam S. Learning in potential games. Working Papers ir97022, International Institute for Applied Systems Analysis 1997.

[36] Pinker S. *The Language Instinct: How The Mind Creates Language.* Morrow New York, NY 1995.