Semantic normativity and coordination games

Social externalism deflated

Semantic Normativity and Coordination Games: Social Externalism Deflated

Daniel Lassiter

Department of Linguistics, New York University
Institute of Philosophy, School of Advanced Study, University of London

[Final version, 1 Nov 2010. In press at *Croatian Journal of Philosophy*]

Abstract. Individualists and externalists about language take themselves to be disagreeing about the basic subject matter of the study of language. Are linguistic facts are *really* facts about individuals, or *really* facts about language use in a community?

The right answer to this question, I argue, is 'Yes'. Both individualistic and social facts are crucial to a complete understanding of human language. The relationship between the theories inspired by these facts is analogous to the relationship between anatomy and ecology, or between micro- and macro-economics: both types of facts are important objects of study in their own right, but we need a theory that accounts for the complex relationship between the two. I argue that modern extensions of the signaling-games approach of Lewis (1969) do just this, defusing the conflict while preserving the core positive insights of both sides of this debate.

The upshot is that arguments for social externalism and the normativity of meaning pose no threat to individualist explanations and can be accounted for within a naturalistic theory of language. A complete theory of language will make crucial reference to individualistic facts, but go further by examining language users' interactions in a systematic way.

Keywords: Individualism, internalism, externalism, normativity of meaning, coordination games

1. Introduction

What kinds of facts form the subject matter of a theory of language? Here are some proposals. According to *individualists*, linguistic facts are facts about individuals' behavior and/or mental states. Individualists are further divided on the question of whether behavioral facts are relevant to a linguistic description: *internalist* individualists, notably Chomsky, claim that the proper subject matter of a scientific theory of language is the internal psychological states of individuals, and that overt behavior is at best indirect evidence regarding the true object of study.

In the opposite corner, *externalists* argue that the appropriate subject matter of a theory of language crucially (perhaps exclusively) involves facts external to individuals' skin. According to *social* externalists, linguistic facts – at least, the

interesting ones – deal with the aggregate linguistic behavior of a community. Social externalism is closely related to the Wittgensteinian notion of *semantic normativity*, the claim that facts about meaning are inherently normative. On this approach, questions of meaning are not questions about how a word **is** used, but about how a word **ought** to be used given facts about the use of others in the community of which a speaker is part.[1]

I will try to articulate a way of viewing this question which cuts across these well-worn divisions. Essentially, the idea is this: both individualistic and social facts are real and important, and they play different but interlocking roles in a complete theory of language. Zooming in on questions of meaning, as much of the philosophical debate does, the idea is that there is not one concept of 'meaning' at play, but two: *individual meaning* and *social meaning*. Individual meaning depends directly on psychological facts, and plays an important role in explaining speakers' and interpreters' linguistic behavior in an immediate sense – their choice of *strategy*, in game-theoretic jargon. Social meaning and other social aspects of language use are emergent properties of the complex systems formed by groups of individuals in interaction. The two notions are systematically related, but the large-scale behavior of a complex system of this type is not generally predictable from the properties of individuals and descriptions of their local interactions. As a result, we expect to find patterns of behavior emerging at the macro-level that are not predictable from the micro-level, as we in fact do.

The idea that I am describing is not really new: it is implicit in the practice of many sociolinguists, as well as some linguists and philosophers whose perspective on language is informed by game theory. But the consequences of this perspective for the internalism-externalism debate have not been fully appreciated. In particular, many externalists have doubted that individual idiosyncrasies have any interest at all for the study of language. But even if we were to allow (rather dubiously) that 'community' behavior is the only thing that a theory of language should deal with, this perspective would still be too limited because minor idiosyncrasies can have large-scale effects down the line. On the opposite extreme, some internalists have argued vociferously that facts about language outside the skin are not the sort of thing that we can even try to give a scientific account of, either because issues of language use are too complex or because they are somehow supposed to be beyond the scope of scientific inquiry (e.g., Chomsky 1995, Lohndal & Narita 2009). I have dealt with this sort of 'argument from personal incredulity' elsewhere (Lassiter 2010a). Here I will assume that the approach I take stands or falls on its own merits, and not relative to some predetermined classification of strategies for inquiry into 'scientific' and 'unscientific' categories.

A few words about what I will not be doing here. First, I will deal almost exclusively with questions about meaning, simply because this has been the focus of the debate in the past. The game-theoretic model applies equally well to grammatical facts, however. Second, I will say nothing about environmental externalism, simply because it is a much more difficult topic, and one for which

[1] A further type, *environmental* or *physical* externalism (Putnam 1975), will not be discussed here. Social and environmental externalism are, at least prima facie, independent theses.

theoretical and conceptual tools do not lie ready at hand as they do in the case of social externalism. Third, I will say very little about the idea that mental content is externally individuated, as my main interest is in the philosophy of linguistics. (However, there is a real question whether externalism as interpreted here is able to support mental content externalism.)

2. Motivations for Externalism

2.1 A Quick Argument for Social Externalism

Our starting point is Burge's (1979) well-known argument for social externalism. Simplifying a good bit, here is the story. Jim goes to the doctor, complaining, 'I have arthritis in my thigh'. The doctor replies, 'You must mean "rheumatism"; arthritis is an inflammation of the joints'. The question that we are meant to ask is: was Jim's utterance 'I have arthritis in my thigh' **false**?

Most people – at least, those who do have a stake in the debate at hand – answer 'yes'. If truth-value intuitions are to be taken seriously as data, then this is a serious problem for individualism about linguistic meaning, for the following reason.[2] In order to evaluate Jim's utterance for truth or falsity, we must (i) supply it with a syntactic parse, (ii) assign denotations to the leaves of the tree, (iii) calculate the denotation of the sentence, and (iv) take account of relevant contextual factors. The first two steps of this process require interpreting Jim's utterance as belonging to some language *L*. Now assume that the identity of *L* is determined exclusively by individualistic facts about Jim, and that the parse is unambiguous. Since (let us suppose) Jim believes that 'arthritis' means the same as 'inflammation of the soft tissue', the obvious choice of *L* will be a language that assigns to the linguistic item 'arthritis' the same content as the complex 'inflammation of the soft tissue'. But assuming that Jim thinks that the words of the latter phrase mean the same that we do, and that *L* reflects this fact, his utterance is true in *L*, since he does in fact have an inflammation of the soft tissue. As a result, we should find an unambiguous judgment that Jim's utterance is true. By reductio, our assumption that the choice of *L* is determined exclusively by individualistic facts about Jim is false; non-individualistic facts must play at least some role in determining what language Jim is speaking.

Burge concludes from this thought-experiment that the meaning of 'arthritis' in Jim's mouth is determined by how the word is used in Jim's community, whether or not he is fully aware of the community's usage. Furthermore, Burge claims, the example does not rely on any special features of the word 'arthritis': the example could be replicated with any linguistic item. If Burge is right here, then internalism is in trouble. To be fair, internalists have some ways out: for example, they might deny that truth-value intuitions count as data, or deny that the intuitions are reliable in this case. However, the former reply would endanger the whole enterprise of

[2] Burge (1979) takes the story I am describing as a problem for individualism about mental content as well. However, as Kim (2005) points out, the basic intuition elicited by Burge's story is about word meaning; the move to mental content externalism requires a further step of reasoning about which we may have separate doubts. As a result I will stick to questions of meaning here.

semantics, while the latter takes on the burden of explaining what it is that makes this example special. Without a specific criterion for distinguishing reliable and unreliable intuitions, it is unclear what force such a response would have.

2.2. Social Externalism and Semantic Normativity

Burge (1979) is primarily concerned with what determines the meanings of words, and concludes that community usage, not individual usage or individuals' psychological states, does this job. This argument is closely aligned with Kripke's (1982) famous 'plus'/'quus' argument purporting to show that meaning is deeply *normative* (although the two are of course somewhat different in form and scope). In particular, they share the conclusion that community usage is the fundamental factor determining what words mean.

Kripke's argument runs roughly as follows. Suppose I ask you: 'What is 68 plus 57?' You think briefly and answer: '68 plus 57 equals 125.' Kripke's surprising claim is that there is no individualistic fact about you that makes this the correct answer, as opposed to the answer '68 plus 57 equals 5.' The reason is that we can imagine two denotations for 'plus', one where 'plus' denotes standard addition '+', and one where 'plus' denotes '⊕', where $x \oplus y = x + y$ if $x,y < 57$, and $x \oplus y = 5$ otherwise. Kripke asks: what are the facts by virtue of which, when I said 'plus', I means '+' and not '⊕'?

Questions of past arithmetical history are obviously not crucial here; no matter how much arithmetic you have done, we could invent an example with the same form to make Kripke's point. Moreover, appealing to dispositions will not help. If someone had a disposition to make arithmetical mistakes when adding large numbers (as most of us do), we would not conclude from this that the speaker means something different by 'plus' than we do, but that they sometimes make mistakes. Kripke's conclusion is that the meaning of 'plus' – and, by extension, other words of the language – is a normative fact: not a fact about what people **do**, or what they **would** do in appropriate circumstances, but about what they **should** do if they are to use the language correctly.

To make sense of using a language 'correctly', Kripke appeals to the notion of belonging to a community:

> The entire 'game' that we have described – that the community attributes a concept to an individual so long as he exhibits sufficient conformity, under test circumstances, to the behavior of a community – would lose its point outside of a community that generally agrees on its practices.
>
> (Kripke 1982: 96)

The conclusion, then, is generally consonant with Burge's: facts about meaning rely crucially on facts about the community in which a speaker is embedded.

In the following I will argue that the game-theoretic perspective on language use, in a certain sense, validates the conclusion that meaning is normative. However, this conclusion is much less radical than it appears: the relevant social norms are an instance of a much less terrifying theoretical construct, conventions. First, though, I

will briefly review some problems for the notion of 'community' that Burge and Kripke appeal to.

3. Linguistic Communities?

Social externalist theories of language usually rely heavily on the notion of a 'linguistic community'. The existence of these objects is generally taken for granted. However, most linguists consider the notion of a linguistic community too ill-defined to do any theoretical work. In particular, in many cases around the world we find *dialect continua*, unbroken low-level variation over large social or geographical spaces in which, at the extremes, we find two or more groups of speakers who clearly do not speak the same language. With no natural boundaries, any attempt to divide these individuals into 'communities' would be completely arbitrary. Furthermore, linguistic variation is so ubiquitous that and, no matter how broadly or narrowly we demarcate linguistic communities, we will still find internal variation, even down to the level of individuals. Hudson (2001) discusses the concept of the linguistic community in some detail, concluding that we have no choice but to 'give up any attempt to find objective and absolute criteria for defining speech communities'. This issue is discussed in more detail with further references by Lassiter (2008), who concludes that 'terms such as "language", "dialect", and "speech community" cannot be defined precisely without doing violence to the empirical facts of human language'. I think it is fair to say that this is the opinion of a large majority of linguists.

What is the significance of this fact for social-externalist theories? Chomsky, for one, concludes that social externalism is straightforwardly falsified:

> [Externalism] crucially relies on a notion of 'common, public language' that remains mysterious. … [O]rdinary usage provides no notion of 'shared public language' that comes even close to meeting the requirements of empirical inquiry or serious philosophical reflection on language and its use, and no more adequate notion has been proposed.
>
> (Chomsky 2000: 155,158)

This seems to me to be to be correct, at least as far as the usual presentations of social externalism go. The following sentence, however, is less obvious:

> Nor is there an explanatory gap that would be filled by inventing such a notion, as far as is known.

I think that there **is** an explanatory gap: restricting ourselves to a internalist account of language prevents us from accounting for Burge's and Kripke's problems discussed in section 2. No internal facts about Jim will account for how his use of 'arthritis' to mean 'inflammation of the joints or soft tissue' can be *wrong* – after all, as Burge points out, if others around Jim used 'arthritis' in the way that he does, our truth-value judgments would be different. The same goes for Kripke's argument for the normativity of meaning.

What we need is an account that resolves these problems without making reference to anything like a 'common, public language'. The game-theoretic approach that I will present does just this: notions like 'social norm', 'social meaning', and 'correctness' have a natural explication in terms of optimal strategies in a coordination game, with no reliance on anything like a 'community' or a 'public language'. All we need are the most basic concepts of coordination games: individuals adopt strategies and encounter other individuals who also adopt strategies. The 'correctness' of a strategy is determined relative to its success, measured by its usefulness to agents in accomplishing (non-linguistic) goals.

4. Convention and Coordination

Coordination games were introduced in Schelling (1960) and used in an effort to explain the origins of linguistic meaning by Lewis (1969). Lewis was particularly concerned to respond to skeptical arguments by Quine (1936), who rejected the 'platitude' that language was conventional. According to Quine, typical conventions are established by some kind of agreement, typically by using language. For language to be conventional, there would have to have been a prior language in which the convention was formed, which just pushes the problem back further. Quine concludes that "[t]he sober truth is that our use of language conforms to regularities – and that is all" (Lewis 1969, p.2, summarizing Quine).

Lewis' contribution was to give an account of convention that did not rely on prior agreement – essentially, the beginnings of a theory of the evolution of conventions. The theory is quite general, although many of the central cases of interest involve conventions of language. According to Lewis (1975), a regularity R is a *convention* in a population P if and only if, within P,

1. Everyone conforms to R.
2. Everyone believes that the others conform to R.
3. The belief that the others conform to R gives everyone good and decisive reason to conform to R himself.
4. There is a general preference for general conformity to R rather than slightly-less-than-general conformity …
5. R is not the only possible regularity meeting the last two conditions.
6. (1-5) are common knowledge.

It is clear, in light of the discussion in the previous section, that these conditions – invoking universal conformity within a well-defined population as they do – are much too strong. A notion of linguistic 'convention' that is usable for us will need to allow for variation, and it cannot rely on 'populations' being extrinsically given. As we will see, however, these features of Lewis' analysis can be discarded without loss of content once we interpret the analysis game-theoretically and enrich the model with some independently motivated features.

The simplest game-theoretic interpretation of Lewis' definition of convention involves a *repeated coordination game* with players drawn at random from a population. We can think of these games as abstract descriptions of situations in

which both participants, the *players*, have preferences over states of the world and want to choose an *action* which will maximize their *utility* (a numerical description of their preferences). However, in any non-trivial game, the utility of a particular action depends on what action the other player chooses. For example, if you and I want to meet for coffee, and neither of us cares much where, then our interests are well-served if we go to the same coffee shop, and thwarted if we do not. Or again, driving on the right side of the road is a good idea if the next person you pass going the other way happens to be driving on their right; otherwise, the consequences can be dire. A strategic form representation of a game like this is given in Figure 1.

		Player 2	
		Right	Left
Player 1	Right	**1** **1**	**0** **0**
	Left	**0** **0**	**1** **1**

Figure 1. The driving game.

The boxes to the right of 'Player 1' represent this player's two possible actions, *Drive on the right* and *Drive on the left*; the boxes below 'Player 2' represent his possible actions (which happen, in this game, to be the same set). In each box, the number in the bottom left represents the payoff to player 1 (p_1) if that pair of actions is chosen, and the number in the top right represents the payoff to player 2 (p_2).

Note that, for any pair of strategies $<p_1(a), p_2(a')>$ in the driving game, the choice of action a is an optimal strategy for player 1 if and only if the choice of action a' is an optimal choice for player p_2. Moreover, the driving game is non-trivial in that, for each player, there are multiple actions which might be optimal depending on which action the other player takes.

Lewis suggests that games like this can be used to give an account of the formation of linguistic conventions as well, using what he calls *signaling games*. I illustrate using a well-known example from animal communication.[3] Vervet monkeys have a call 'pyow' which warns of an approaching leopard, and a call 'hack' which warns of an eagle overhead. If a leopard is coming, the best thing to do is to climb a tree if one is available; if an eagle is coming, the best thing to do is freeze in place and hope that the eagle does not see you. We may assume that the players have some interest in each others' well-being, so that the player who spots an eagle or leopard is better off if he gives an accurate warning (though perhaps not so drastically affected if he fails as the other player is).

[3] Signaling games are also used widely in linguistic pragmatics: see Jäger (2008) and Franke (2008, ch.1) for good overviews. These applications differ somewhat from our current concerns because they generally take the linguistic meanings of words and sentences to be given extrinsically, and use game-theoretic tools to extract further information from the choice of signals. Here, we are interested in using game theoretic tools to illuminate the process of choosing an interpretation in the first place, as Lewis (1969) was. Needless to say, the two uses of game theory are not in competition: Jäger (2008), for instance, treats the use of game theory in the evolution of language side-by-side with its use to model on-line pragmatic inference.

Leopard present:		Monkey Two	
		Climb Tree	Freeze
Monkey One	"pyow"	**1** **5**	**-1** **-5**
	"hack"	**1** **5**	**-1** **-5**

Figure 2a. Subgame if leopard present

Eagle present:		Monkey Two	
		Climb Tree	Freeze
Monkey One	"pyow"	**-1** **-5**	**1** **5**
	"hack"	**-1** **-5**	**1** **5**

Figure 2b. Subgame if eagle present

Figure 2. The vervet-predator game

The obvious difference between the game in Figure 2 and the driving game is that, if we know how the world really is – here, which predator is present – the choice of message does not matter at all: Monkey Two's best choice is to take the appropriate action for the predator at hand. This remains a coordination game, however, because of the information asymmetry between the participants: only one of them knows which subgame they are really in. Monkey One knows whether there is a leopard or an eagle approaching, and wants to signal to Monkey Two which of these is the case so that Monkey Two can react appropriately. The *choice of signaling system* does not matter, as long as it is agreed on. If Monkey Two knows that Monkey One has adopted the strategy *if leopard, 'pyow'; if eagle, 'hack'*, then he can infer upon hearing 'hack' that the game is Figure 2b and act appropriately; and likewise with the inverse strategy.

This idea can be made more precise using a simplified signaling game model. A signaling game consists of a *sender* S who knows which state of the world from a set of relevant options $W = \{w_1, \ldots w_k\}$ is actual; a set of *messages* $M = \{m_1, \ldots m_n\}$ from which the sender chooses; and a *receiver* R who chooses from a set of possible actions $A = \{a_1, \ldots a_m\}$. We assume in the simple model that messages are costless, i.e. that sender's choice of message does not affect either player's payoffs except in its possible influence on receiver's action. A (pure) sender strategy s for this game is a function $W \rightarrow M$ which determines, for each relevant state, which message the sender will send. A (pure) receiver strategy r is a function $M \rightarrow A$ which determines which action the receiver will take given each message that the sender might send. A *strategy profile* is a pair $\langle s,r \rangle$ of a sender strategy s and a receiver strategy r. A strategy profile completely characterizes the players' behavior in a round of play. (For simplicity, we assume that S knows with certainty which state is actual, although this assumption is not crucial. Later we will see *mixed* strategies as well, which are just probability distributions over pure strategies.)

Even for this simple model, it is non-trivial to assume that S and R will converge on a strategy profile which is useful to both of them: there are sixteen possible strategy profiles in this game, and messages do not carry any information in most of them. For instance, one possible sender strategy in the vervet-predator

game is to send the message 'pyow' no matter which type of predator is present. Likewise, a possible receiver strategy in this game is to ignore the message and climb a tree no matter what.

	Sender strategy	Receiver strategy
Profile 1.	{<Leopard, 'pyow'>, <Eagle, 'pyow'>}	{<'pyow', freeze>, 'hack', climb tree>}
Profile 2.	{<Leopard, 'pyow'>, <Eagle, 'hack'>}	{<'pyow', climb tree>, 'hack', climb tree>}

Figure 3. Some possible strategy profiles in the vervet-predator game

A *signaling system*, as Lewis defines it, is a strategy profile in which S sends a different signal for each state of the world and R takes the best possible action in each state. For instance, in the vervet-predator game, there are two possible signaling systems, given in Fig. 4.

	Sender strategy	Receiver strategy
Profile 1.	{<Leopard, 'pyow'>, <Eagle, 'hack'>}	{<'pyow', climb tree>, 'hack', freeze>}
Profile 2.	{<Leopard, 'hack>, <Eagle, 'pyow'>}	{<'pyow', freeze>, 'hack', climb tree>}

Figure 4. Lewisian signaling systems in the vervet-predator game

In this game, the two signaling systems are the only strategy profiles in which messages convey useful information. In many other strategy profiles, messages carry no information at all; in still others, messages convey perverse information, in the sense that R always takes the action which is *worst* for him (and for S).

5. Formation of Conventions

Signaling games, as we have interpreted them, are a model of language choice: senders (speakers) and receivers (listeners) each choose a strategy and, if their strategies overlap sufficiently and in situationally appropriate ways, each receives some benefit. The basic question that we need to ask here is: how do players manage to converge on strategy profiles in which messages convey information?

There are many proposed 'solution concepts' in the game-theoretic literature. For example, if everyone in a population were to choose *Right* in a repeated driving game, and everyone were to do so because she knew that the others would also do so, this regularity of behavior would not only meet Lewis' definition of a convention, but would also be a strict Nash equilibrium: each player would be worse off if she were to deviate unilaterally from the norm. What is less clear, though, is what this fact adds to our understanding of the *formation* of conventions. What we need is not just an account that tells us which strategy

profiles are locally optimal for each player, but one that tells us how players come to choose the strategy profile that they do. As the economist H. P. Young puts it,

> Neoclassical economics describes the way the world looks once the dust has settled; we are interested in how the dust goes about settling. This is not an idle issue, since the business of settling may have considerable bearing on how things look afterwards.
>
> (Young 1998: 4)

Lewis' work shares this feature with neoclassical economics: it does not really offer a clear account of how conventions are formed – that is, how speakers coordinate on a choice of language – in the first place.

We can resolve these issues by making use of the *evolutionary* interpretation of the signaling-games model. In this section and the following I will review, briefly and informally, some of the crucial features of evolutionary game theory and how they are useful for a theory of language choice; formal details and much more extensive discussion can be found in references cited.

The "problem of convention formation", according to Young (1996:108), has the following schematic form:

> A number of individuals face a one-shot game that is played repeatedly by different players drawn from a large population. … [T]he players are boundedly rational and only partially informed. They do not have perfect foresight, they do not know in detail the structure of the process they are engaged in, and they do not know why other players are acting the way they are. Instead, they use simple rules of thumb to adjust their behavior based on their information about what other players are currently doing or have done in previous periods. Furthermore, there are unexplained variations in their behavior that play a role analogous to mutations in biological evolution. These three elements – local interaction between individuals, boundedly rational responses to the perceived environment, and random perturbations – define an "evolutionary game dynamic."

Young intends this as a description of the process by which social institutions such as monetary systems, forms of dress, and rules of the road are formed from the bottom up by the interaction of large numbers of agents. As Young suggests and Skyrms (2004) considers in detail, linguistic conventions are presumably formed and maintained by similar processes: large numbers of agents who act locally on the basis of limited information interact to produce an institution which is surprisingly uniform despite the lack of oversight.

There is a large literature making use of analytical results and computational simulations to examine precisely what conditions allow for the possibility of conventions of various kinds (see Skyrms 2004, 2010 for an overview). The basic result is that a convention can be established in a repeated signaling game as long as two conditions are met. First, there must be differential survival and reproduction: in other words, it cannot be the case that all strategies are equally successful.

Second, players must be able to get and use information (possibly very weak) about what other players will do, for instance, by observing their past behavior and guessing that future behavior will be correlated.

When these conditions are fulfilled, signaling systems tend to evolve spontaneously, even though the players may have very little intelligence or insight into the structure of the game. For example, we may suppose that all players initially randomize over the space of possible strategies (i.e., if there are *n* possible sender strategies, a sender S plays a mixed strategy in which each sender strategy *s* is chosen with probability *1/n*). On each round, two players are chosen at random from the population and they play two signaling games, each playing the role of sender once. Occasionally, two players will encounter each other whose strategies happen to be successful against each other. When strategy *s* is successful against strategy *r*, both sender and receiver update their mixed strategies to increase the probability with which *s* or *r* is played according to some pre-determined rule. As a result, when senders playing *s* meet appropriately matched receivers playing *r*, *s* and *r* tend to become more frequently played. Eventually we may see a feedback loop in which *s* and *r* are further strengthened, to the point that the entire population are playing strategy *s* and *r* with probability 1. This is as close as we can get, in an evolutionary setting, to a Lewisian convention, and it appears as the result of simulations under a wide array of starting conditions, learning rules, and assumptions about game structure and error rates (Skyrms 2004).

6. Linguistic Diversity

The result just described is encouraging because it provides a model of language choice in which large-scale patterns emerge from the bottom up from diverse behavior of locally optimizing agents. No monolithic 'public language' is assumed, and no enforcement mechanism is needed, but patterns of behavior which resemble the ordinary notion of a public language emerge spontaneously. Before returning to the questions of linguistic normativity and correctness that motivated our discussion, I want to address a worry that I mentioned earlier with regard to Lewis' notion of convention.

Lewis' definitions rely on universal (or near-universal) conformity to a pattern of behavior within a pre-defined population. We inherited these predictions in our model in the previous section, where a population of players is given in advance, and universal or near-universal conformity is the outcome. But, as we already saw, the real facts of linguistic diversity make these features of the model undesirable: except in exceptional circumstances, there are no well-defined populations of language users in the world who interact only with each other. Furthermore, linguistic diversity is often widespread and structured, rather than being scattered and inherently unstable: Weinreich et al. (1968) call this the property of "orderly heterogeneity". They, and much following work in sociolinguistics, show that structured variation is a characteristic of language use in all human societies.

It turns out, though, that the prediction of universal conformity is a contingent feature of the way we set up the model, and in particular of the

assumption that interaction is random. In the real world, we obviously do not encounter other individuals at random – typically, we encounter a few individuals very often, some occasionally, and most never. A more realistic treatment, then, is in terms of social networks, a construct which has been influential in sociolinguistics since Milroy (1980) and plays a major role in recent game-theoretic analyses of social behavior, including language.

When we add social networks to a model of the type described above, we get several benefits. First, rather than universal conformity, we often get local clustering of diverse strategies, although the distribution of variation depends on the details of the network structure assumed (Skyrms 2004; Goyal 2006). Second, if we restrict attention to small-world networks, the network type most plausible as a description of human social interaction, formation of local conventions is much more rapid and reliable than in games without networks described above (Young 1998; Wagner 2009).

I hasten to add that enriching our model with network structure does not resolve all of our problems with linguistic variation. We still predict that variation is an inefficiency which would be eliminated if there were some way for an appropriate change to be propagated through the network. From a sociolinguistic perspective, however, this is plainly wrong: some forms of linguistic variation are robust and positively useful. This occurs in particular when there are alternative ways of 'saying the same thing' (Labov 2001), in which case the choice of variants often carries social meaning in itself (Le Page & Tabouret-Keller 1985; Eckert 2000, 2008). It is not entirely clear, however, whether these facts create a serious problem for the model at hand. Quite possibly this issue can be resolved by enriching the model with more signals and with interlocutors' desire to communicate social information in addition to information about states of the world: see Lassiter (2010b) for more discussion.

7. Linguistic Norms as Conventions

Tools from evolutionary game theory provide us with a ready-made explanation of how linguistic conventions can arise spontaneously, without the need for a 'public language' to be given in advance or to serve as an enforcer of usage. Most of the time, agents' desire to communicate with each other is sufficient motivation for them to converge linguistically. We also have an account of how linguistic diversity can emerge within complex networks, and a rough sense of how linguistic diversity can be maintained and utilized at the local level.

These latter features eliminate the need to appeal to an extrinsically given 'linguistic community' of dubious empirical status in our analyses of linguistic phenomena – including, I will argue, correct usage and the normativity of meaning. Instead, we can call upon agents' local connections within a network to do the job that the community was supposed to do in Kripke's and Burge's approaches.

There are many different ideas in the philosophical literature about what counts as a 'norm'. In the economics literature, however, it is common to treat 'norm' and 'convention' as closely related or even synonymous (Harsanyi 1968; Young 1996, 1998). For example, driving on the right or left as appropriate is

referred to interchangeably as a 'convention' or a 'norm'. Bicchieri (2006) argues for a classification into *social norms*, which typically require some enforcement mechanism and 'most often apply when there is a tension between individual and collective gains'; and *descriptive norms* such as clothing choice, which are presumably followed because they are individually beneficial, and from which individual deviation is not usually enforced by punishment. I do not want to get involved in the debate about whether all norms fit into this classification; however, I offer up a hypothesis about the sense in which linguistic meaning can rightly be understood to be normative:

> *To say that meaning is normative is nothing more – or less – than to say that it is conventional, in the sense of the evolutionary analysis given above.*

My claim is that linguistic conventions are descriptive norms in Bicchieri's sense, and that this construal is enough to make sense of the normativity of meaning.[4] If this is correct, then we do not need anything more mysterious than a coordination convention of the type discussed above to explain Burge's and Kripke's examples motivating externalism and the normativity of meaning. In particular, since we have seen that coordination conventions do not need to be grounded in 'linguistic communities' or 'public languages', it follows that meaning can be normative in the favored sense without there being any well-defined community with respect to whose usage it is normative. This is enough to answer Chomsky's objection: semantic normativity and linguistic correctness are consequences of the decentralized interactions of individual agents, no public languages needed.

Let's consider Burge's 'arthritis' story first. We can model this interaction straightforwardly as a signaling game between the patient, Jim, and his doctor. Jim has rheumatism; he says to his doctor, 'I have arthritis'. If the doctor understands that Jim has a joint disease and treats him accordingly, this is a bad outcome for both doctor and patient.

[4] After I gave the talk on which this paper is based, I heard Giacomo Sillari give a talk entitled 'Rule-Following as Coordination' in which he made a very similar claim. Unfortunately nothing has been published, and I have not had the opportunity to hear more, but I suspect that the general idea is congenial to my argument.

J		Doctor treats	
		Joint disease	*Soft tissue disease*
Jim says	'arthritis'	**1** **1**	**0** **0**
	'rheumatism'	**1** **1**	**0** **0**

Fig. 5a. Subgame if Jim has a joint (J) disease

ST		Doctor treats	
		Joint disease	*Soft tissue disease*
Jim says	'arthritis'	**0** **0**	**1** **1**
	'rheumatism'	**0** **0**	**1** **1**

Fig. 5b. Subgame if Jim has a soft tissue (ST) disease

Figure 5. The arthritis game

The possible signaling systems in this game are:

	Sender strategy	Receiver strategy
Profile 1.	{<J, 'arthritis'>, <ST, 'rheumatism'>}	{<'arthritis', treat J>, <'rheumatism', treat ST>}
Profile 2.	{<J, 'rheumatism'>, <ST, 'arthritis'>}	{<'rheumatism', treat J>, 'arthritis', treat ST>}

Figure 6. Lewisian signaling systems in the arthritis game

Given his past experiences with patients, the doctor has every reason to expect to expect Jim to adopt the sender strategy in Profile 1, with the result that the doctor's best choice is the receiver strategy in profile 1. However, Burge stipulates that Jim adopts a different strategy, say the sender strategy in profile 2. The resulting strategy profile is obviously defective, for it leads to the sub-optimal outcome where Jim has rheumatism but says 'arthritis', leading the doctor to treat him for arthritis, with a net payoff of zero for each. In general terms, this sub-optimality is due to the fact that the strategy profile that the players have adopted here is not a Lewisian signaling system.

Our intuition that Jim has made a mistake in the arthritis game is due to two factors, on this interpretation. First, the strategy profile consisting of Jim's strategy and the doctor's is not a signaling system. Second, the doctor's strategy *is* optimal relative the usage of other individuals with whom he has been or is likely to be in contact, while Jim's strategy is not. This makes the interesting prediction that, if Jim and the doctor come from different social networks in which their respective usages are commonplace, we are not inclined to interpret Jim's usage as incorrect, but rather as a dialectal difference. Burge hints at this conclusion, and I argue that this is the correct prediction in some detail in Lassiter (2008).

Note that this interpretation of the arthritis scenario does not rely on any philosophically weighty notion of 'correctness', nor on the assumption that there is a relevant social norm from which Jim's deviation from general usage will be punished. This way of looking at Burge's arthritis scenario does not fit nicely into either the usual construal of social externalism. However, it is not an individualist

theory either: rather, understanding this scenario game-theoretically requires taking into account both individualistic and social facts. The fact that Jim said 'arthritis' while intending to communicate what we would normally describe by 'rheumatism' – that is, Jim's choice of individual strategy – has an immediate explanation in terms of Jim's psychology. However, the fact that Jim failed to *communicate* 'rheumatism' is what triggers our intuition that he made a mistake. This can only be explained by facts that are external to Jim – here, facts about the doctor's psychology which account for *his* choice of strategy. Both of these sets of facts are ultimately to be accounted for in terms of previous interactions between these agents and others, tying them in with other members of their respective social networks. I conclude that any explanation of the unfolding of this game which neglects either individual strategy or social structure is doomed to be incomplete: neither purely individualist nor purely externalist explanations will do the job.

Extending the account to Kripke's 'plus'/'quus' story is a bit trickier, but still possible. Clearly, in cases where the two diverge, we can appeal to mis-coordination. Suppose the sender in some game uses '99 plus 87' intending to refer to the function *99 + 87* while the receiver interprets this as referring to the function *99* $\oplus$ *87* (so that she calculates the result as *5*). Depending on what the possible actions in the game are, this may lead to consequences ranging from an undeserved bad grade on a homework assignment or the receiver's mislearning arithmetic, to a spaceship crashing. This is a plausible example of a signaling system gone wrong.

Somewhat less clear are cases in which sender and receiver adopt a strategy profile which appear to be a signaling system by Lewis' definition, but is intuitively a case of accidental coordination. For example, if a sender says '5 plus 7' intending to refer to the function *5 + 7*, while receiver understands *5* $\oplus$ *7*, both will get the same result, and the difference in the procedure by which the result was computed is not reflected in any payoff-relevant action. We might think that this qualifies as a signaling system, since the receiver takes the optimal action. However, this appearance is incorrect: the conclusion would only follow if this were a very strange game in which people take the same sum repeatedly (and it would not be an unreasonable prediction in this case, to my mind).

However, the game which we implicitly take these players to be engaged in involves taking arbitrary sums, and the interpretive strategy which the receiver is adopting here is not optimal given the strategy which the sender is presumably adopting. Imagine that sender and receiver are a teacher and a student respectively, and the game involves dictating an arithmetic exam which the student then attempts to answer. States of the world for a given play of the game are equivalent here to test questions which the teacher transmits. Imagine that the teacher adopts the (infinite) strategy in Figure 7a while the student adopts the (infinite) strategy in Figure 7b. (These strategies obviously have more compact representations, but I write them out to emphasize the similarity to strategies in games that we have already analyzed.)

…
<Question = *5 + 7,* say 'five plus seven'>
<Question = *2 + 4,* say 'two plus four>
<Question = *47 + 99,* say 'forty-seven plus ninety-nine'>
…

Figure 7a. Teacher strategies in the arithmetic game

…
<Teacher says 'five plus seven', compute $5 \oplus 7$>
<Teacher says 'two plus four', compute $2 \oplus 4$>
<Teacher says 'forty-seven plus ninety-nine', compute $47 \oplus 99$>
…

Figure 7b. Student strategies in the arithmetic game

Suppose that student and teacher both get 1 point when the student answers a question correctly, and 0 otherwise. Then the strategy profile consisting of the teacher's strategy in Figure 7a and the student's strategy in Figure 7b is not a Lewisian signaling system. In particular, if the question happens to be *47 + 99*, then the student will compute compute $47 \oplus 99$ and return the answer *5,* which is not the optimal action. The fact that this is not a signaling system means that the strategy profile is defective even though, depending on which questions are asked, this defect may never be noticed.

Note that this account of Kripke's 'plus'/'quus' problem differs in an important way from the dispositional account that Kripke correctly dismisses. On that account, facts about dispositions to overt behavior are supposed to distinguish '+' from '$\oplus$', whether or not the dispositions are ever realized. Kripke points out, however, that this account predicts wrongly that a disposition to make arithmetical errors also disqualifies one from meaning '+' by 'plus'. On the account that I am suggesting, however, strategies are not the same thing as dispositions to overt behavior. Strategies are ultimately psychological facts about agents, and the route by which they are translated into overt behavior may well be error-prone. (Gintis 2009 speaks suggestively of a 'competence-performance' distinction in this domain.) A disposition to make arithmetical errors may exist even if one's strategy is error-free. As a result, Kripke's objection to the dispositional account does not apply to our game-theoretic account of semantic normativity.

8. Overview and Outlook

Many linguists (and some philosophers) are hostile to social externalism and semantic normativity because they appear to rely on problematic empirical claims about linguistic communities and public languages, or because they invoke unfamiliar and perhaps ill-defined concepts such as 'correctness' and 'normativity'.

I have tried to argue that an independently motivated game-theoretic approach to linguistic conventions yields a plausible account of the most influential arguments for social externalism and semantic normativity, and does this without dubious empirical commitments or invoking new theoretical concepts. At the same time, although I have not attempted to draw out these consequences in detail, this account does not seem to share the radical philosophical consequences which have been claimed to follow from semantic externalism. I do not doubt that committed partisans on both sides of this debate will find my account lacking. However, I hope that I have at least made the case that it is a non-trivial question whether there really is a deep theoretical or philosophical conflict to be found in the internalism-externalism debate.

On the account that I prefer, individualistic and externalistic facts find their appropriate place at different levels of explanation, and linguistic theory should seek to discover what happens at each of these levels, and what the principles are that connect them. I suggest that individual meaning – or, more generally, individual sociolinguistic competence – should be indentified with what game theorists call *individual strategy*, interpreted as a psychological state of individual language users. (This might resemble the steady states of a probabilistic model of grammatical learning such as Yang 2002, though it would be somewhat more general.) Innate constraints on the space of possible grammars and word meanings, if any, determine the space of possible strategies that an agent can adopt.

Individualistic/internalistic considerations such as these play a vital role in the gigantic coordination game that is human language. But there is more. Without paying attention to facts about how individuals interact – that is, externalist facts – we artificially limit the scope of our theory of language. I do not think that there is anything wrong with doing this: if internalist facts are all that a particular theorist is interested in, she should of course feel free to worry about these facts exclusively. But it is disingenuous to pretend, as some internalists still do, that there is nothing more that can be said in a scientifically respectable way about human language. The science of complex systems is far too well-developed today for this to be a credible methodological restriction.

According to Epstein (2007), the fundamental question of social science is 'how ensembles achieve functionalities that their constituencies lack'. That is, according to Epstein, social science is, or should be, interested in *emergent* phenomena, properties of complex systems which arise from decentralized interactions but cannot be predicted by inspecting these interactions at a local level. Human language is a complex system, and conventions of language are emergent phenomena in precisely this sense. Experience influences individuals' psychology; psychological facts interact with performance systems to generate behavior. Individual behavior, combined with facts about network structure and the nature of the game being played, determines which large-scale properties will emerge, both in terms of the distribution of strategies and the diachronic properties of the system. To be sure, the state of the system at higher levels of analysis cannot be directly read off of the strategy choices of any particular agent in the system. Nevertheless, the ensemble of individual strategies and the way that these strategies interact fully

determines the state of the system. A complete picture of human language will make reference to facts collected at all levels of description.

References

Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge University Press.

Burge, Tyler. 1979. Individualism and the mental. In P. French , T. Uehling , and H. Wettstein (eds.), *Midwest Studies in Philosophy, vol. IV: Studies in Metaphysics.* Minneapolis: University of Minnesota Press.

Chomsky , Noam. 1986. *Knowledge of Language*. New York: Praeger .

Chomsky, Noam. 1995. Language and nature. *Mind* 104: 1-61.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge, UK: Cambridge University Press.

Eckert, Penelope. 2000. *Linguistic Variation as Social Practice*. Oxford, U.K.: Blackwell.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4): 453-476.

Epstein, Samuel. 2007. *Generative Social Science: Studies in Agent-Based Computational Modeling.* Princeton: Princeton University Press.

Franke, Michael. 2008. *Signal to Act: Game Theory in Pragmatics.* Ph.D. thesis, Institute for Logic, Language, and Computation, Universiteit van Amsterdam. [http://student.science.uva.nl/~mfranke/Papers/Franke_PhD_thesis.pdf]

Gintis, Herbert. 2009. *The Bounds of Reason.* Princeton: Princeton University Press.

Goyal, Sanjeev. 2007. *Connections: An Introduction to the Economics of Networks.* Princeton: Princeton University Press.

Harsanyi, John. 1968. Individualistic and functionalistic explanations in light of game theory: The Example of Social Status. In I. Lakatos and A. Musgrave eds., *Problems in the Philosophy of Science*. Amsterdam: North Holland.

Hudson, R. A. 2001. *Sociolinguistics*. Cambridge, UK: Cambridge University Press.

Jäger, Gerhard. 2008. Applications of game theory in linguistics. *Language and Linguistics Compass* 2: 1-16.

Kim, Jaegwon. 2005. *Philosophy of Mind.* Westview Press.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Labov, William. 2001. *Principles of Linguistic Change, Vol. II: Social Factors.* Oxford: Blackwell.

Lassiter, Daniel. 2008. Semantic externalism, language variation, and sociolinguistic accommodation. *Mind and Language* 23(5): 607-633.

Lassiter, Daniel. 2010a. Where is the conflict between internalism and externalism? A reply to Lohndal & Narita (2009). *Biolinguistics* 4(1): 138-148.

Lassiter, Daniel. 2010b. Strategic interaction as a bridge between inter- and intra-individual variation. Manuscript, New York University and Institute of Philosophy.

Le Page, R. B., and Andrée Tabouret-Keller. 1985. *Acts of Identity: Creole-Based Approaches to Language and Ethnicity.* Cambridge, UK: Cambridge University

Press.
Lewis, David. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Lewis, David. 1975. Languages and language. In K. Gunderson (ed.), *Language, Mind and Knowledge*. Minneapolis: University of Minnesota Press.
Lohndal, Terje and Hiroki Narita. 2009. Internalism as methodology. *Biolinguistics* 3(4): 321-331.
Milroy, Leslie. 1980. *Language and Social Networks.* Oxford: Blackwell.
Nowak, Martin. 2006. *Evolutionary Dynamics: Exploring the Equations of Life.* Cambridge, MA: Belknap Press of Harvard University Press.
Putnam , Hilary. 1975. The meaning of 'meaning'. In *Mind, Language and Reality: Philosophical Papers* vol. 2. Cambridge: Cambridge University Press.
Quine, W. V. O. 1936. Truth by convention. In O. H. Lee (ed.), *Philsophical Essays for A. N. Whitehead*. New York: Longmans.
Schelling, Thomas. 1960. *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.
Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure.* Cambridge, UK: Cambridge University Press.
Skryms, Brian. 2010. *Signals: Evolution, Learning, and Information.* Oxford: Oxford University Press.
Wagner, Elliot. 2009. Communication and structured correlation. *Erkenntnis.*
Weinreich, Uriel, William Labov and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkeil (eds.), *Directions for historical linguistics: A symposium.* Austin: University of Texas Press. 95-188.
Young, H. Peyton. 1996. The economics of convention. *Journal of Economic Perspectives* 10(2): 105-122.
Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions.* Princeton: Princeton University Press.