Natural Language Equilibrium: Off-Path Conventions I

Philip J. Reny[*]

May 2025

Abstract

We incorporate natural language into games, focusing here on the class of signaling games. The sender, using a commonly understood language, can make cheap-talk statements about the strategy that he is using. Because the sender knows his strategy, any statement that he makes is either true according to its literal meaning or is *intentionally* false or deceptive. It is shown that if the receiver interprets any off-path statement by the sender as true unless it may be seen as a *rational* attempt to deceive, then the only outcomes of the game without language that survive the introduction of language are, generically, those that are stable in the sense of Kohlberg and Mertens (1986). Incorporating language into game theory can thus reap significant benefits, with the potential to significantly refine equilibrium predictions in ways that are more intuitive and more easily justified than when language is absent.

Keywords: natural language, language conventions, language games, stable equilibria, sender-stable equilibria, signaling games.

1 Introduction

We seek a theory of equilibrium behavior for noncooperative games in which a formal role is played by a cheap-talk language that is commonly understood by the players. Our focus here is on the class of signaling games largely because its sender-receiver

[*] Department of Economics, University of Chicago, preny@uchicago.edu. I wish to thank Kyle Bagwell, Dirk Bergemann, Ben Brooks, Eddie Dekel, Songzi Du, Piotr Dworczak, John Geanakopolos, Marina Halac, Emir Kamenica, Frédéric Koessler, Marco LiCalzi, Roger Myerson, Wojciech Olszewski, Motty Perry, Andy Postlewaite, Doron Ravid, Joe Root, Joel Sobel, and Yair Tauman for helpful comments, and the Department of Economics at Northwestern University for their hospitality while I was a visiting scholar there from July 2023 to June 2024. Financial support from the National Science Foundation (SES-2049810) is gratefully acknowledged.

structure provides a significant role for language while also allowing for a theory that is relatively simple to describe. Extending these ideas beyond signaling games would be worthwhile.

To be effective, a language must have conventions in place that support informative communication by identifying a wide range of circumstances in which statements can be made and interpreted truthfully according to their literal meanings. Because pre-existing literal meaning can be overridden by equilibrium usage to the contrary, such conventions may be especially important for statements that are made off the equilibrium path. One particularly salient convention, and the one that we will study here, is that *any off-path statement is interpreted as true unless it may be seen as a rational attempt by the speaker to deceive the listener.*

With this intuitive convention we find that a rich language generically eliminates all equilibrium outcomes of the game without cheap talk that fail to be stable in the sense of Kohlberg and Mertens (1986), and that even a very coarse language yields a powerful refinement of sequential equilibrium that strictly refines Cho and Kreps' (1987) never a weak best reply criterion.

These results for the class of signaling games illustrate the potential benefits of pursuing a broader program to develop a theory of language in games. The benefits include both a theory with sharp predictive power as well as a theory that is based upon intuitive and easily understood conventions for language usage. Language, it seems, may provide a key link connecting strong refinements with simple intuitions.

Like many equilibrium refinements (e.g., Grossman and Perry 1986, Banks and Sobel 1987, and Cho and Kreps 1987), the equilibrium notion that we introduce here, *natural language equilibrium*, is unrestrictive and coincides with sequential equilibrium in sender-receiver games of "pure communication" (Crawford and Sobel 1982, Green and Stokey 2007), where all the sender's actions are payoff irrelevant. Our natural language equilibrium concept has refinement power only when the sender has actions that are payoff-relevant, such as in the education signaling game of Spence (1973), where the only equilibrium that survives the introduction of natural language is the fully separating "Riley" equilibrium (Riley 1979).[1] This shortcoming indicates that the theory put forward here is incomplete and that there is a need for even stronger conventions that, in addition, assign to some statements their literal meaning *even if* they could be interpreted as a rational attempt to deceive. The development of such stronger conventions will be considered in follow-up work.

The classic approach to using language to refine equilibria in sender-receiver games has been to specify conditions under which an out-of-equilibrium statement made by the sender about his private information, or "type", is credible. In this approach, initiated by Farrell (1985, 1993), an out-of-equilibrium statement of the form "my type is in the set $\{t_1, ..., t_k\}$," is called *credible* if, were it believed by the receiver, the receiver would have an optimal reply that, relative to their equilibrium payoffs, is a Pareto improvement for the types in the set but is not a strict improvement for any type outside the set. According to this approach, any credible statement is always believed by the receiver—a stipulation that, in our terminology, is a convention for the language. With this convention, credible statements can allow some sender-types to separate themselves from others and so can sometimes break equilibria that involve pooling. But this approach has been difficult to implement without being in conflict with the existence of an equilibrium (Farrell 1985, 1993; Grossman and Perry 1986; Matthews et. al. 1991, Matthews and Postlewaite 1994), suggesting that the convention for the language that is specified may be too strong.[2] In response, a variety of alternative approaches have been proposed (see Section 7), not all of which resolve the problem of existence.

To reiterate, in the approach taken here, we consider a language convention in which the receiver may doubt the veracity of any off-path statement made by the

[1] This follows from our Theorem 4.7 together with the treatment of Spence (1979) in Cho and Kreps (1987, Section V).

[2] One can interpret the "Stiglitz critique" as calling into question the general sorts of arguments used in this literature to refine equilibria. For the Stiglitz critique itself, see Cho and Kreps 1987, p. 203 and footnote 6 there, and see also footnote 19 below.

sender so long as that statement can be seen as a rational attempt by the sender to deceive him. This simple and intuitive convention avoids the existence problem while still providing strong predictions.

The remainder of the paper is organized as follows. Section 2 contains notational conventions and other preliminary material, and section 3 defines sequential equilibrium in our setting. Section 4 introduces our solution concept "natural language equilibrium," and provides statements of our main results. Several examples are then given in Section 5. In Section 6 we consider the effects of coarsening the language. Section 7 discusses the literature. All proofs can be found in Section 8 with the exception of the proof of Theorem 6.3 which is in the main text. The Supplemental Appendix provides a discussion of various points including the effect of adding more statements to the language, the role of counterfactuals and the introduction of costly messages.

2 Preliminaries

Throughout, if a set is defined by an upper-case letter, e.g., X, then its elements will be denoted by the corresponding lower-case letter x, x', etc. Then, whenever we write x or x', etc., it is implicit that they are in the set that is their upper-case counterpart, X.

Define the *base game* Γ_0 as follows. There are two players, a Sender (S) and a Receiver (R). Nature chooses S's type t from a non-empty finite set T according to a strictly positive probability measure p. After observing t, S chooses an action a from a non-empty finite set A. Then, R, after observing a but not t, chooses an action z from a non-empty finite set Z. Both players can randomize their action choices. After t, a, and z have been chosen, the players receive their payoffs $u_i(t, a, z)$, $i = S, R$, and the game Γ_0 ends.[3]

[3] We have assumed that the action sets A and Z are history-independent. This is for simplicity only. All of our results extend to the case in which these sets are history-dependent.

Next, let M be a non-empty finite or countably-infinite set of cheap-talk messages m, and suppose that, in addition to taking an action a, the sender S is allowed to send R any message m. The messages in M are payoff irrelevant. So if S's type is t and he chooses the action-message (a, m) after observing t, and R chooses z after observing (a, m), then payoffs are $u_i(t, a, z)$, $i = S, R$. Let $\Gamma(M)$ denote this cheap-talk extension of the base game Γ_0.

Notice that the base game Γ_0 is equivalent to the game $\Gamma(M)$ whenever M contains exactly one message because this one message must then be sent by all types, rendering it uninformative. We will occasionally make use of this fact.

3 Sequential Equilibria

In this section, we define sequential equilibrium (Kreps and Wilson, 1982) for the game $\Gamma(M)$. These definitions apply also to the base game Γ_0 by setting M equal to a singleton.

An *assessment* (σ, ρ, β) consists of a *strategy* $\sigma : T \to \Delta(A \times M)$ for S, a *strategy* $\rho : A \times M \to \Delta(Z)$ for R, and a *system of beliefs* $\beta : A \times M \to \Delta(T)$ for R.[4]

Say that an assessment (σ, ρ, β) is a *sequential equilibrium* of $\Gamma(M)$ iff for every t, a, m, z

(i) $\sigma(a, m|t) > 0$ implies $(a, m) \in \arg\max_{a',m'} \sum_{z'} u_S(t, a', z')\rho(z'|a', m')$,

(ii) $\rho(z|a, m) > 0$ implies $z \in \arg\max_{z'} \sum_{t'} \beta(t'|a, m)u_R(t', a, z')$, and

(iii) $\sum_{t'} \sigma(a, m|t')p(t') > 0$ implies $\beta(t|a.m) = \frac{\sigma(a,m|t)p(t)}{\sum_{t'} \sigma(a,m|t')p(t')}$.

Conditions (i) and (ii) are sequential rationality requirements for S and R, respectively, and condition (iii) requires R's beliefs to be Bayes consistent whenever possible.[5]

[4] For any set X, $\Delta(X)$ denotes the set probability distributions over X.

[5] Technically, Kreps and Wilson (1982) define sequential equilibrium only for finite games. The definition given here applies whether M is finite or countably infinite, and coincides with Kreps and Wilson's definition when M is finite because their "consistency" condition on beliefs is equivalent to Bayes' consistency in finite signaling games with or without cheap-talk.

For any sender-type t, for any action-message (a,m), and for any strategy ρ for R, say that (a,m) is a *best reply for t against ρ* iff $(a,m) \in \arg\max_{a',m'} \sum_{z'} u_S(t,a',z')\rho(z'|a',m')$.

Whenever a strategy profile σ for S is understood (as, for example, when an assessment (σ,ρ,β) is under consideration), say that an action-message (a,m) is *off-path (for σ)* iff $\sigma(a,m|t)=0$ for every sender-type t. Otherwise, (a,m) is *on-path*.

4 Natural Language Equilibrium

Consider the signaling game Γ_0. We wish to introduce a commonly understood language in which S can make cheap-talk statements about the strategy that he is using.[6] In doing so we will limit the number of statements in the language to be countably infinite. This entails no substantive loss of generality and eliminates a host of unnecessary technicalities that would otherwise arise.

Let Q be the countably infinite set of all probability vectors in $\Delta(T)$ that have rational coordinates, and, for any $q \in Q$ and for any $t \in T$, let $q(t)$ denote the probability that q assigns to t.

We will let Q represent the countably infinite set of all possible messages/statements in the natural language.[7] With the availability of this language, the players play the signaling game $\Gamma(Q)$.

Each statement in the natural language Q is endowed with a distinct literal meaning. Specifically, for any action taken by the sender (and observed by the receiver), the literal meaning of any accompanying message q is, "I am using a strategy that gives my action and this message positive probability and that, conditional on my action and this message, induces the Bayes posterior q over my types."[8]

[6] Matthews et. al. (1991) introduce "talking strategies" to serve a similar purpose. The term "talking strategy" seems to have first appeared in Park (1997) who studied pure communication games with cheap talk messages but without a natural language.

[7] We use the terms "message" and "statement" interchangeably.

[8] Note the two distinct usages of $q \in Q$ here, once as a representation of (or encoding of) the message itself, and once—within that message—as an element of $\Delta(T)$. Throughout, we will distinguish between these usages by saying "message/statement q" or "beliefs/posterior q" as appropriate.

Notice that for any action a and for any statement q, if the sender takes action a and makes the statement q, then there are always many strategies that S might be using that would make the literal meaning of statement q true, e.g., any strategy $\sigma : T \to \Delta(A \times Q)$ such that $\sigma(a, q|t) := \varepsilon q(t)/p(t)$ for every t, and some $\varepsilon > 0$, where the remainder of the strategy, namely $\sigma(a', q'|t')$ for any t' and for any $(a', q') \neq (a, q)$, can be specified arbitrarily.

Therefore, and importantly, when S takes any action a and makes any statement q, it is always possible that the statement is true (according to its literal meaning), and, if R were to accept it as true, then R's posterior about S's type would be q.

The language Q is rather rich. In particular, it effectively contains all of the $|T|$ statements of the form "My type is t." Indeed, if the probability vector q is a mass point on t and R accepts the statement q as true, then R would place probability one on S's type being t.

Since literal meaning can be overridden by equilibrium usage to the contrary, the effectiveness of a natural language will depend in part on conventions about when to accept off-path statements as true according to their literal meanings, with stronger conventions corresponding to a more effective language. To understand the limits of how effective a language can be in a strategic setting, we should therefore study the strongest possible conventions, i.e., those in which off-path statements are accepted as true according to their literal meanings unless there is "very good" reason not to. But whatever the convention, observe that, because S knows his own strategy, any off-path statement by S about the strategy that he is using is either true or is *intentionally* false in which case we may say that the off-path statement is an *attempt to deceive*.[9,10]

The convention that we will consider here is that R accepts any off-path statement

[9] See Sobel (2020) for proposed definitions of "lying" and "deception" in games. Note that we are silent here on whether the sender's off-path *attempt* to deceive is successful, i.e., whether it is in fact a deception.

[10] See discussion points 7 and 8 in the Supplemental Appendix on the assumption that S knows his own strategy.

by S as being true unless it may be seen as a *rational* attempt to deceive. This leads to the central definition of this paper.

Definition 4.1 Say that (σ, ρ, β) is a *natural language equilibrium (NLE)* for the base game Γ_0 iff it is a sequential equilibrium of $\Gamma(Q)$ and, for any off-path action-message (a, q) and for any pair of distinct types t and t', if $\beta(t|a, q)/\beta(t'|a, q) > q(t)/q(t')$ then (a, q) is a best-reply for t against ρ.[11]

Let us call beliefs that satisfy the condition given in Definition 4.1 *straightforward*; and call them *straightforward at* (a, q) if the condition holds for a given action-message (a, q). Thus, a natural language equilibrium (for Γ_0) is a sequential equilibrium of $\Gamma(Q)$ with straightforward beliefs.

Definition 4.1 says that after any off-path action-message (a, q) by S, if R's beliefs $\beta(\cdot|a, q)$ depart from q—the beliefs that R would have if he accepted as true S's claim that his strategy induces the posterior q—then R's beliefs can give higher relative probability only to types for whom (a, q) is a best reply against ρ. This means that $\beta(\cdot|a, q)$, if not equal to q, is a convex combination of q and some other probability measure on T that gives positive probability only to types for whom (a, q) is a best reply against ρ.[12] Consequently, Definition 4.1 ensures that after any action-message (a, q) by S, R's beliefs weigh precisely two possibilities: either S's statement is true or S's statement may be a rational attempt to deceive.

Remark 4.2 *If (σ, ρ, β) is an NLE, then, by definition, β is straightforward at every off-path (a, q). But in fact β is straightforward at* every (a, q), *off-path or not, because the requisite condition is necessarily satisfied at every on-path (a, q). Indeed, for any*

[11] If either one of the two denominators in the strict inequality is zero, the inequality is to be understood as holding if and only if $\beta(t|a, q)q(t') > \beta(t'|a, q)q(t)$. We maintain this convention throughout the paper.

[12] To see this, denote $\beta(t|a, q)$ by $\beta(t)$ and let $\lambda := \min_{t \in T: q(t)>0} \beta(t)/q(t)$ with t^* a minimizer. Then $\lambda \in [0, 1)$ because β and q are distinct and in $\Delta(T)$. Defining $\gamma(t) := (\beta(t) - \lambda q(t))/(1 - \lambda)$ for each $t \in T$ gives $\gamma \in \Delta(T)$ and $\beta = \lambda q + (1 - \lambda)\gamma$. Moreover, if $\gamma(t) > 0$ then $\beta(t) > \lambda q(t) = \frac{\beta(t^*)}{q(t^*)} q(t)$ and so $\beta(t)/\beta(t^*) > q(t)/q(t^*)$, which implies that (a, q) is a best reply for t against ρ. For a similar observation see Sobel (2020), Proposition 2.

(a, q), if $\beta(t|a,q)/\beta(t'|a,q) > q(t)/q(t')$, then $\beta(t|a,q) > 0$. So if (a,q) is on-path then $\sigma(a,q|t) > 0$ by Bayes consistency, in which case (a,q) is a best reply for t against ρ by sequential rationality.

Importantly, in an NLE, the rationality of any off-path action-message chosen by S is measured against how well it does against R's *equilibrium* strategy, a point of view that resonates with Occam's razor. Indeed, after observing an off-path action-message, R updates his beliefs through some combination of simply accepting as true S's statement about his strategy, and continuing to believe that S is rational, that S knows R's equilibrium strategy, and that S knows how language is used and understood.

Our first result states that natural language equilibria always exist.

Theorem 4.3 *There is at least one natural language equilibrium.*

In preparation for our other results, we introduce some additional terminology. Any terminology for $\Gamma(M)$ applies also to Γ_0 by setting M equal to a singleton.

For a given signaling game with cheap talk $\Gamma(M)$, each strategy profile (σ, ρ) induces a probability distribution over all the endpoints (t, a, m, z) in the game $\Gamma(M)$ as well as a marginal distribution over the payoff-relevant states (t, a, z). Call the distribution over endpoints the *outcome* of the game and call the distribution over payoff-relevant states the *type-action distribution* of the game.[13] If an outcome is the outcome for a Nash equilibrium (σ, ρ), call it a *Nash equilibrium outcome.* If it is the outcome for a sequential equilibrium (σ, ρ, β), call it a *sequential equilibrium outcome*, and so on, and similarly for the type-action distribution. Note that because the endpoints of Γ_0 are the payoff-relevant states (t, a, z), the terms "outcome" and "type-action distribution" are synonymous for Γ_0. We will be especially interested in relating outcomes in Γ_0 with the type-action distributions of natural language equilibria of $\Gamma(Q)$.

[13] I thank Joel Sobel for suggesting the "type-action distribution" terminology.

Each strategy profile also induces interim expected utilities $\pi_S(t)$ for each of S's types t, and an ex-ante expected utility π_R for R.[14] Call the induced expected utility vector, $((\pi_S(t))_{t\in T}, \pi_R) \in \mathbb{R}^{|T|+1}$, the *payoff* for that strategy profile. If a payoff is the payoff for a Nash equilibrium, call it a *Nash equilibrium payoff.* If it is the payoff for a sequential equilibrium call it a *sequential equilibrium payoff* and so on.

For any $\varepsilon \in (0,1)$, for any $\delta_1, \delta_2 \in (0,\varepsilon)$, and for any strategy profile (σ_0, ρ_0) for $\Gamma(M)$, suppose that when the players choose any strategy profile (σ, ρ) in $\Gamma(M)$, the strategy profile that is implemented is instead $((1-\delta_1)\sigma + \delta_1\sigma_0, (1-\delta_2)\rho + \delta_2\rho_0)$. Call the resulting game an *ε-perturbation of* $\Gamma(M)$ *toward* (σ_0, ρ_0). If only the sender's strategy is perturbed, i.e., if $\delta_2 = 0$, call the resulting game an *ε-perturbation of* $\Gamma(M)$ *toward* σ_0.

Let M be a finite set of messages, let $\nu \in \Delta(T\times A\times Z)$ be a type-action distribution of the game $\Gamma(M)$.

Analogous to Kohlberg and Mertens (1986), say that ν is a *stable type-action distribution* of $\Gamma(M)$ iff for any strictly mixed strategy profile, and for any sequence of ε-perturbations of $\Gamma(M)$ toward that strategy profile, there is a corresponding sequence of Nash equilibria of the perturbed games whose type-action distributions converge to ν as ε converges to zero.

Similarly, say that ν is a *sender-stable type-action distribution* of $\Gamma(M)$ iff for any strictly mixed strategy of the sender and for any sequence of ε-perturbations of $\Gamma(M)$ toward that strategy of the sender, there is a corresponding sequence of Nash equilibria of the perturbed games whose type-action distributions converge to ν as ε converges to zero.

Sender-stability is less restrictive than stability because it does not require the given type-action distribution or payoff to be robust against perturbations of the receiver's strategy. But in terms of restricting the receiver's off-path beliefs, sender-stability and stability are equally powerful in that each requires robustness to every

[14] For any strategy profile (σ, ρ) and for any $t \in T$, $\pi_S(t) := \sum_{a,m,z} \sigma(a,m|t)\rho(z|a,m)u_S(t,a,z)$ and $\pi_R := \sum_{t,a,m,z} p(t)\sigma(a,m|t)\rho(z|a,m)u_R(t,a,z)$.

perturbation of the sender's strategy.

Define the phrase *for generic base-game utilities* to mean for all base-game utility vectors $(u_S(t,a,z), u_R(t,a,z))_{(t,a,z)\in T\times A\times Z}$ in $\mathbb{R}^{2(T\times A\times Z)}$ outside some closed set having Lebesgue measure zero.

The next result states that the number of messages in Q that support any natural language equilibrium type-action distribution can be bounded above by $|T\times Z|$, and that the type-action distribution of any sender-stable Nash equilibrium of $\Gamma(T\times Z)$ is a natural language equilibrium type-action distribution.

Theorem 4.4 *Every natural language equilibrium type-action distribution is induced by a natural language equilibrium that uses no more than $|T\times Z|$ messages in Q. Moreover, any sender-stable type-action distribution of $\Gamma(T\times Z)$ is a natural language equilibrium type-action distribution.*

Remark 4.5 *We conjecture that, for generic base-game utilities, only $|T|$ messages are needed to induce any NLE type-action distribution. See Park (1997) for a related result for sequential equilibrium type-action distributions in pure communication games. Using a result by Koessler, Laclau, and Tomala (2024, footnote 5), it can be shown that no more than $|T|+1$ messages are needed to generate any natural language equilibrium payoff $((\pi_S(t))_{t\in T}, \pi_R)$.*

Remark 4.6 *Reny (2024) obtains a characterization of natural language equilbrium payoffs. He shows that $((\pi_S(t))_{t\in T}, \pi_R)$ is the payoff of a natural language equilibrium if—and only if for generic base-game utilities—$((\pi_S(t))_{t\in T}, \pi_R)$ is the payoff of a sender-stable Nash equilibrium of $\Gamma(T)$.*

We should not expect a clear-cut relationship between the set of sequential equilibria of the base game Γ_0—a signaling game without cheap talk—and the set of natural language equilibria after the addition of the language Q. On the one hand, adding to a signaling game any form of cheap talk, including the language Q, expands the set of sequential equilibria. On the other hand, restricting R's beliefs as in a natural language equilibrium, refines the set of sequential equilibria.

Nevertheless, it is natural to ask whether any of the sequential equilibrium outcomes of the base game Γ_0 survive the introduction of natural language. Our next result states that at least one sequential equilibrium outcome of the base game Γ_0 does survive. Moreover, generically, those that survive are precisely the stable outcomes. The proof uses a succinct characterization of Kohlberg-Mertens stable outcomes in generic signaling games (without cheap talk) due independently to Cho and Kreps (1987) and Banks and Sobel (1987).

Theorem 4.7 *At least one sequential equilibrium outcome of the base game is a natural language equilibrium type-action distribution. Moreover, an outcome of the base game is a natural language equilibrium type-action distribution if—and only if for generic base-game utilities—it is a stable outcome of the base game.*

Remark 4.8 *A Corollary of Theorem 4.7 and Lemma 8.3 is that, for generic base-game utilities, every sender-stable outcome of the base game is stable. That is, for the base game, stability and sender-stability generically coincide.*[15]

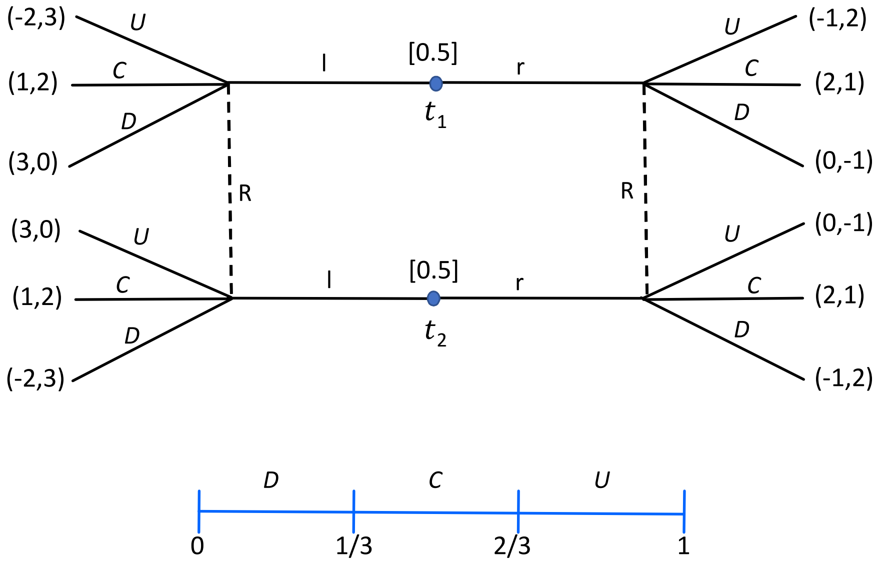
5 Examples

In each example, the game-tree diagram there depicts the base game Γ_0. Dashed lines indicate R's information sets, the first coordinate in each payoff vector is the payoff to S, and the second coordinate is the payoff to R. Each of the examples is generic in the sense required by Theorem 4.7.

Example 5.1 Our first example illustrates how our definitions work and shows how Theorem 4.7 can help to identify the natural language equilibria for a base game.

[15] It is an open question as to whether, for a generic set of base-game utilities, sender-stable outcomes are stable in cheap-talk games. See Blume (1994) for several examples, one of which provides a non-generic pure communication game in which the babbling equilibrium outcome, which is always sender-stable, is not stable.

Consider the base game Γ_0 shown in the upper part of Figure 1. The sender S has two equally-likely types, t_1 and t_2, and two actions, l and r. The receiver R has three actions, U, C, and D.



R's optimal actions given Pr(t_1)

Figure 1. The base game Γ_0 and R's optimal actions.

Notice that when S changes his action from l to r, R's payoff falls by 1 util for each of his three actions. Consequently, R's set of optimal actions depends only on the probability that he assigns to S's type being t_1 and does not depend on the action, l or r, that is taken by S. Of course, in equilibrium, R's beliefs—and therefore also his optimal actions—may depend on whether S chooses l or r.

The lower part of Figure 1 shows R's optimal actions as a function of the probability, denoted by $\Pr(t_1)$, that he assigns to t_1. So, for example, after S chooses l (or r), if R believes that there is probability $2/3$ that S's type is t_1, then R has two optimal actions, U and C.

It is straightforward to show that the base game Γ_0, without cheap talk, has two sequential equilibrium outcomes, one in which both sender-types choose l and then R responds with C, giving payoff $((\pi_S(t_1), \pi_S(t_2)), \pi_R) = ((1,1),2)$, and another in which both sender-types choose r and then R responds with C, giving payoff

$((2,2),1)$. Let us denote these two outcomes by $\nu_{l,C}$ and $\nu_{r,C}$, respectively.[16] Both of these sequential equilibrium outcomes pass some of the most stringent equilibrium refinements, including Cho and Kreps' (1987) never a weak best reply test. However, only one of them is a natural language equilibrium type-action distribution.

The natural language equilibria for Γ_0 are, by definition, the sequential equilibria of the cheap-talk extension $\Gamma(Q)$ that have straightforward beliefs. In the extended game $\Gamma(Q)$, S can choose any of the infinitely many action-message pairs (l,q) or (r,q) for any $q \in Q$, and then the continuation game is exactly as in Figure 1 after the choice of l or r, respectively.

In general, the game $\Gamma(Q)$ can have more sequential equilibrium type-action distributions than the base game Γ_0. In the present example however, this is not the case. The only Nash (or sequential) equilibrium type-action distributions of $\Gamma(Q)$ here are $\nu_{l,C}$ and $\nu_{r,C}$. We will use this fact below.

We now show using the definitions that only $\nu_{r,C}$ is a natural language equilibrium type-action distribution for Γ_0. To establish this, we first display a sequential equilibrium $(\sigma^*, \rho^*, \beta^*)$ of $\Gamma(Q)$ whose beliefs are straightforward and whose type-action distribution is $\nu_{r,C}$. Here is one of many.

Let $p \in Q$ denote the uniform prior over the sender's types, i.e., $p(t_1) = p(t_2) = 1/2$. Define σ^* so that,

$$\sigma^*(r,p|t_1) = \sigma^*(r,p|t_2) = 1.$$

That is, σ^* is defined so that each sender-type chooses the action-message (r,p) with probability one, and chooses every other action-message (a,q) with probability zero.[17]

Define ρ^* so that R chooses C with probability one if either S chooses (r,q) for any q, or if S chooses (l,q) such that $q(t_1) \in (1/3, 2/3)$; and so that R chooses D and C with probability one-half each if S chooses (l,q) such that $q(t_1) \leq 1/3$; and so

[16] More precisely, $\nu_{l,C}$ puts probability one-half each on (t_1, l, C) and (t_2, l, C), and, $\nu_{r,C}$ puts probability one-half each on (t_1, r, C) and (t_2, r, C).

[17] Thus, on-path, the sender is honest about the strategy that he is using. But the same path can supported with a dishonest statement since, either way, the receiver's on-path beliefs are determined by Bayes' rule.

that R chooses C and U with probability one-half each if S chooses (l, q) such that $q(t_1) \geq 2/3$.

Finally, define β^* so that,

$$\beta^*(t_1|r, q) = \beta^*(t_2|r, q) = 1/2, \text{ for every } q,$$

and so that,

$$\beta^*(t_1|l, q) = \begin{cases} 1/3, & \text{if } q(t_1) < 1/3 \\ q(t_1), & \text{if } q(t_1) \in [1/3, 2/3] \\ 2/3, & \text{if } q(t_1) > 2/3. \end{cases}$$

The assessment $(\sigma^*, \rho^*, \beta^*)$ is sequentially rational for S because both types of S receive an equilibrium payoff of 2, and can obtain a payoff of 2 by choosing (r, q) for any q, and can obtain a payoff of at most 1 by choosing (l, q) for any q. The assessment is sequentially rational for R because ρ^* is defined so that R mixes with equal probability on all actions that are optimal for him given his beliefs β^*. Finally, the beliefs β^* are Bayes consistent because each of the equally-likely types of S chooses (r, p) with probability one and $\beta^*(t_1|r, p) = \beta^*(t_2|r, p) = 1/2$. Hence, $(\sigma^*, \rho^*, \beta^*)$ is a sequential equilibrium of $\Gamma(Q)$ and it clearly induces the type-action distribution $\nu_{r,C}$. It remains to show that the beliefs β^* are straightforward.

We must show that the beliefs β^* are straightforward at every off-path (a, q). There are two kinds of off-path action-messages, namely, (r, q) with $q(t_1) \neq 1/2$, and (l, q) for any q.

Consider first some (r, q) with $q(t_1) \neq 1/2$. Since R responds with C after (r, q), the action-message (r, q) is a best reply for both types against ρ^*. Hence, β^* is trivially straightforward at (r, q).

Consider next any (l, q). There are three cases to consider, namely, $q(t_1) < 1/3$, $q(t_1) \in [1/3, 2/3]$, and $q(t_1) > 2/3$.

In the first case, $q(t_1) < 1/3$. Together with the definition of β^*, this implies,

$$\frac{\beta^*(t_1|l,q)}{\beta^*(t_2|l,q)} = \frac{1/3}{2/3} > \frac{q(t_1)}{q(t_2)}.$$

Therefore, straightforward beliefs given (l, q) requires (l, q) to be a best reply for t_1 against ρ^*. To see that this is the case, observe that according to ρ^*, after t_1 chooses (l, q) with $q(t_1) < 1/3$, R puts equal weight on D and C, giving t_1 an expected payoff of 1, which is his equilibrium payoff.

In the second case, $q(t_1) \in [1/3, 2/3]$ and so $\beta^*(t_1|l,q) = q(t_1)$. Hence, β^* is trivially straightforward at (l, q).

In the third case, $q(t_1) > 2/3$. By switching the roles of types t_1 and t_2 and the roles of actions D and U, this third case is similar to the first and so we may conclude that the beliefs β^* are straightforward. Hence we have shown that $\nu_{r,C}$ is a natural language equilibrium type-action distribution.

We next show that $\nu_{l,C}$, the only other sequential equilibrium type-action distribution of $\Gamma(Q)$, is not a natural language equilibrium type-action distribution.

Let $(\bar{\sigma}, \bar{\rho}, \bar{\beta})$ be any sequential equilibrium of $\Gamma(Q)$ that induces the type-action distribution $\nu_{l,C}$. Note that both types of S receive an equilibrium payoff of 1. To show that $(\bar{\sigma}, \bar{\rho}, \bar{\beta})$ is not a natural language equilibrium, we must show that the beliefs $\bar{\beta}$ are not straightforward. We will do so by showing that $\bar{\beta}$ is not straightforward at some off-path action-message.

Since both types of S choose l with probability one, the action r is chosen with probability zero. So let us consider the off-path action-message (r, p) in which S takes the action r and makes the statement p (whose literal meaning is that S is using a strategy whose posterior conditional on (r, p) puts probability 1/2 on each type).

After observing (r, p), R believes that there is probability $\bar{\beta}(t_1|r,p)$ that S's type is t_1. Notice that $\bar{\beta}(t_1|r,p) \neq 1/2$ since, otherwise, sequential rationality would require R to respond with C after observing (r, p) which would break the equilibrium since both types could then profitably deviate to (r, p).

So either $\bar{\beta}(t_1|r,p) > 1/2$ or $\bar{\beta}(t_2|r,p) > 1/2$. Suppose first that $\bar{\beta}(t_1|r,p) > 1/2$. Then, since $p(t_1) = p(t_2) = 1/2$,

$$\frac{\bar{\beta}(t_1|r,p)}{\bar{\beta}(t_2|r,p)} > \frac{p(t_1)}{p(t_2)}.$$

Consequently, for $\bar{\beta}$ to be straightforward at (r,p), it must be the case that (r,p) is a best-reply for t_1 against $\bar{\rho}$. Thus (r,p) must yield t_1 a payoff of 1, his equilibrium payoff. However (consult Figure 1), this can happen in a sequential equilibrium of $\Gamma(Q)$ in which $\bar{\beta}(t_1|r,p) > 1/2$ only if $\bar{\beta}(t_1|r,p) = 2/3$ and $\bar{\rho}(\cdot|r,p)$ chooses U and C with probabilities 1/3 and 2/3 respectively. But this would break the equilibrium because, in that case, t_2, whose equilibrium payoff is also 1, could profitably deviate to (r,p).

We conclude that if $\bar{\beta}(t_1|r,p) > 1/2$, then $\bar{\beta}$ is not straightforward at (r,p). By a symmetric argument (involving t_2, D and C) the same conclusion follows if $\bar{\beta}(t_2|r,p) > 1/2$. Hence, $\bar{\beta}$ is not straightforward and we are done.

Thus we have shown that, of the two sequential equilibrium type-action distributions, $\nu_{l,C}$ and $\nu_{r,C}$, of $\Gamma(Q)$, only $\nu_{r,C}$ is a natural language equilibrium type-action distribution. Hence, only $\nu_{r,C}$ survives the introduction of natural language, even though the equilibria supporting $v_{r,C}$ and $\nu_{l,C}$ both satisfy Cho and Kreps' (1987) never a weak best reply criterion, the strongest of the criteria distinct from stability that they discuss.

We end this example by showing how Theorem 4.7 can expeditiously draw these same conclusions.

It is not difficult to show that the outcome $\nu_{l,C}$ is not stable in the base game Γ_0 against any small perturbation of the receiver's strategy together with any small perturbation of the sender's strategy in which both sender-types choose action r with the same probability. Therefore, $\nu_{l,C}$ is not a stable equilibrium outcome and so by the second part of Theorem 4.7, $\nu_{l,C}$ is not a natural language equilibrium type-action distribution. Hence, by the first part of Theorem 4.7, the only other sequential equilibrium outcome of Γ_0, namely $\nu_{r,C}$, must be the unique natural language equilibrium type-action distribution for this example.

Example 5.2 Our second example provides a natural language equilibrium type-action distribution that is not induced by any strategy profile for the base game. This happens because, in this very simple example, the language Q permits more separation of the sender's types than is possible in the base game without the language.

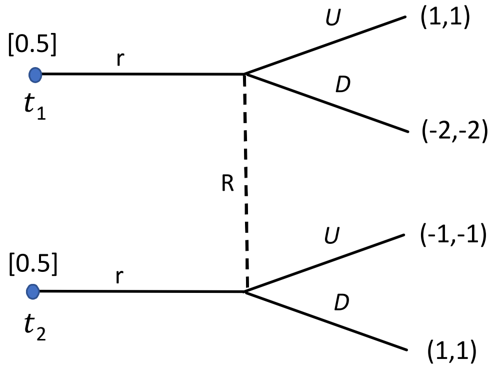


Figure 2.

In the base game shown in Figure 2, the sender has two equally likely types, t_1 and t_2, and has just one action, r. The receiver has two actions, U and D. With just one action, the sender's types must pool. Hence, there is a unique sequential equilibrium of the base game in which both types choose r, after which R believes that both types are equally likely and optimally chooses U. Payoffs in this "pooling" equilibrium are 1 for type t_1, -1 for type t_2, and 0 for R.

Notice that type t_2 would strictly prefer to separate from t_1 since, if R knew that S's type were t_2, then R would respond with D and t_2's payoff would be 1. Moreover, type t_1 would be perfectly content to separate from t_2 since R would still play U if he knew that S's type were t_1. But with only one available action, and without language, the types cannot separate.

Introducing the natural language Q allows the types to separate. Here is one way they can do this. For $i = 1, 2$, let q_i denote the element of Q that puts probability one on type t_i. The literal meaning of statement q_i is, effectively, "my type is t_i."

There is a "separating" natural language equilibrium in which t_1 sends the action-message (r, q_1) after which R puts probability one on t_1 and takes action U, and in which t_2 sends the action-message (r, q_2) after which R puts probability one on t_2 and takes action D, and in which, after observing any other action-message (r, q), R puts probability one on t_1 and takes action U.

It is easy to verify that the assessment just defined is a sequential equilibrium of $\Gamma(Q)$. To establish that it is a natural language equilibrium we must check that the beliefs are straightforward at every off-path action-message. Notice that for any off-path action-message (r, q), R's response, U, makes (r, q) a best reply for t_1. Therefore, since R's beliefs after any off-path (r, q) are concentrated on t_1, R's beliefs are trivially straightforward.

The type-action distribution of this separating equilibrium puts probability one-half on each of the payoff-relevant states (t_1, r, U) and (t_2, r, D). This distribution cannot be induced by any strategy profile for Γ_0 because, in Γ_0, the types must pool on the only available action, r, and $\not R$'s response must then be the same for both types.

The above separating equilibrium is, however, not the only natural language equilibrium here. Indeed, the first part of Theorem 4.7 implies that the pooling equilibrium outcome of Γ_0, being its only equilibrium outcome, must survive the introduction of the language Q.

Consider, for example, the following strategies for $\Gamma(Q)$. Both types of the sender choose the action-message (r, p), where p denotes the uniform prior on the sender's types, after which R places probability one-half on each type and takes the action U, and, after any other action-message (r, q), R places probability one-half on each type and takes the action U. This is clearly a sequential equilibrium of $\Gamma(Q)$. Moreover, because every action-message (r, q) is a best reply for both types, the beliefs are trivially straightforward. Therefore, this strategy profile is a natural language equilibrium whose type-action distribution coincides with the pooling equilibrium outcome of Γ_0.

Why do straightforward beliefs allow pooling to survive? The literal meaning of q_2 is, in effect, "my type is t_2," So one might have thought that type t_2 would be able to credibly convince R of his type by sending the off-path action-message (r, q_2). But in the pooling equilibrium, R's response to this off-path play is to take action U, and so (r, q_2) is actually a best-reply for type t_1. Consequently, R can (and does) place positive probability on the possibility that it is type t_1—in a rational attempt to deceive—who deviated to (r, q_2). See discussion point 9 in Supplemental Appendix A for more.[18]

Example 5.3 This example illustrates the existence problem that can arise with other approaches and shows how natural language equilibrium avoids it. The base game shown in Figure 3 is from Farrell (1993), with slightly modified payoffs.

As in the previous example, the sender here has just one action, r, and so this game has a unique sequential equilibrium in which both types pool by taking action r, and R responds by taking action C since both types are equally likely. In this equilibrium, type t_2 receives his highest possible payoff of 2. In contrast, type t_1, who also receives a payoff of 2 in equilibrium, could obtain a payoff of 3 if he could reveal his type to R, since R would take action U if he knew S's type was t_1. So it seems credible for type t_1 to say "I am type t_1 and you can be certain that I am being truthful because only I stand to gain by revealing my type to you." If it were understood that R would believe this statement, then t_1 would strictly benefit by

[18] One might think that adding small costs to some messages might refine away the pooling equilibrium outcome. But this is incorrect. For example, if S had to incur a small cost of $\varepsilon > 0$ to send the off-path message "my type is t_2," then we could still get close to the pooling equilibrium outcome as follows. Type t_1 sends the (costly) message $q_2 =$ "my type is t_2" with probability 1/3 and sends the (costless) message $q_1 =$"my type is t_1" with probability 2/3. Type t_2 sends the message q_1 with probability 1. Then, after q_2, R knows S's type is t_1 and plays U, netting t_1 a payoff of $1 - \varepsilon$. After q_1, R is indifferent between U and D and plays U with probability $1 - \varepsilon/3$, making t_1 indifferent between messages q_1 and q_2, and making t_2 strictly prefer q_1. Both messages, q_1 and q_2 are on-path. After any off-path (costless) message q, R's beliefs place probability 2/5 on t_1 and R plays U with probability $1 - \varepsilon/3$. It can be checked that this defines an NLE whose outcome converges to the pooling outcome as $\varepsilon \to 0$. Interestingly, the NLE strategies do not converge to the pooling strategies.

making it and this would break the equilibrium.[19]

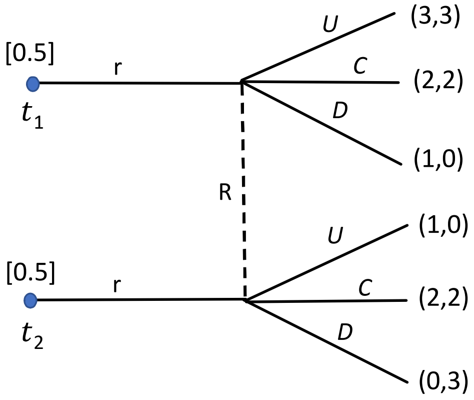


Figure 3.

Farrell (1993) calls statements of this kind "credible neologisms." If a credible neologism exists for some equilibrium, then that equilibrium is not "neologism-proof" and Farrell eliminates it from consideration as a possible solution. Since there is a credible neologism for the unique equilibrium of the game in Figure 3, no neologism-proof equilibrium exists for this game.[20]

Natural language equilibrium avoids this non-existence problem by allowing the receiver to be more skeptical. The following is a natural language equilibrium for the base game in Figure 3. Both sender-types choose the action-message (r, p), where p is the prior that puts probability one-half on each type. The receiver, R, responds to every action-message (r, q) by choosing action C and having uniform beliefs that put probability one-half on each type.

To see how this handles the claim by t_1 that "my type is t_1", let q_1 be the element of Q that puts probability one on t_1. Then, as in the previous example, the effective literal meaning of the statement q_1 is, "my type is t_1." In the natural language equilibrium just described, if, in an attempt to reveal his type, t_1 were to send the off-path

[19] But some (Joe Stiglitz being perhaps the first) have taken arguments such as this a step further and argued that, because t_2 would then inadvertently reveal his type and receive his lowest possible payoff of zero by *not* making the same statement, t_2 would in fact also make the statement and so t_1 cannot separate himself from t_2 after all.

[20] For critiques of neologism-proofness see Rabin (1990, p.21) and Matthews et. al. (1991).

action-message (r, q_1), then, R would believe that it is equally likely that S is telling the truth (and his type really is t_1) as it is that S is being deceptive (and his type his actually t_2). R would then rationally respond by taking action C, making t_1's attempt to separate unsuccessful. Importantly, R's response of C makes (r, q_1) a best reply for type t_2, which is precisely why R's straightforward beliefs can place positive probability on t_2—because (r, q_1) can be seen as a rational attempt by t_2 to deceive. Formally, R's beliefs are trivially straightforward because, after any action-message, (r, q), R's response of C makes (r, q) a best reply for both sender-types.

6 Coarse Natural Language Equilibrium

In this section, we show that even a very coarse language can lead to a powerful refinement of sequential equilibrium.[21]

For each $t \in T$, let q_t denote the element of Q that puts probability one on t. Recall that the literal meaning of the statement q_t is, in effect, "my type is t."

We now coarsen the language by reducing it to the set consisting only of the $|T|$ statements q_t for $t \in T$. For this section only, instead of writing q_t, let us simply write t. Hence, the language here consists of the finite set of messages T, where for each $t \in T$, the literal meaning of the statement t is "my type is t." Specializing Definition 4.1 to this coarser language leads to the following.

Definition 6.1 *Say that* (σ, ρ, β) *is a* coarse natural language equilibrium for the base game Γ_0 *iff it is a sequential equilibrium of* $\Gamma(T)$ *and, for any off-path action-message* (a, t') *and for any* $t \neq t'$, *if* $\beta(t|a, t') > 0$ *then* (a, t') *is a best-reply for* t *against* ρ.

Beliefs that satisfy the condition in Definition 6.1 are called *coarsely-straightforward.* Except for the coarser language, the interpretation of coarsely-straightforward beliefs

[21] For settings in which this coarser language may be appropriate see discussion points 7 and 8 in the Supplemental Appendix.

is the same as before, namely, that when confronted with an off-path action-message (a, t'), R weighs two possibilities. Either the message "my type is t'" is true or it is a rational attempt to deceive. This means that R's posterior can put some or all weight on t' and can give positive weight to other types only if the observed action and message is a best reply for those types against R's equilibrium response.

Natural language equilibrium type-action distributions that use more than $|T|$ messages may not be feasible for any strategy profile of $\Gamma(T)$ and hence cannot be coarse natural language equilibrium type-action distributions. Hence, the existence of a coarse natural language equilibrium does not follow immediately from Theorem 4.3. Nevertheless, we have the following.

Theorem 6.2 *A coarse natural language equilibrium exists. In fact, at least one sequential equilibrium outcome of the base game is a coarse natural language equilibrium type-action distribution.*

We next argue that coarse NLE is a rather strong refinement of sequential equilibrium.

In their influential study of refinements for signaling games, Cho and Kreps (1987) provide a hierarchy of progressively more stringent refinements of sequential equilibrium. The most powerful refinement in this hierarchy, short of stability itself, is a refinement that Cho and Kreps call the "never a weak best reply" criterion which we define next.[22]

Formally, for any set of messages M, a sequential equilibrium outcome ν of $\Gamma(M)$ satisfies the (Cho-Kreps) never a weak best reply (NWBR) criterion iff for any action-message (a, m) that has probability zero under ν, either, there exists a sequential equilibrium (σ, ρ, β) of $\Gamma(M)$ with outcome ν satisfying $\beta(t|a, m) = 0$ for every t such that (a, m) is not a best reply for t in any sequential equilibrium with outcome ν,

[22] This criterion expresses, in the context of signaling games, Kohlberg and Mertens' (1986) "forward-induction" property for stable sets.

or, (a, m) is not a best reply for any t in any sequential equilibrium of $\Gamma(M)$ with outcome ν.

Say that a sequential equilibrium of $\Gamma(M)$ satsifies the NWBR criterion iff its outcome satsifies the NWBR criterion.

Cho and Kreps (1987) show that if a sequential equilibrium outcome satisfies their NWBR criterion, then it also satisfies their "intuitive criterion" and two stronger criteria that they call "D1" and "D2," as well as two refinements called "Divinity" and "Universal Divinity" due to Banks and Sobel (1987). Thus NWBR is more powerful than each of these refinements. Our next result states that coarse NLE is more powerful still.

Theorem 6.3 *Every coarse natural language equilibrium outcome satisfies the Cho-Kreps never a weak best reply criterion, but not conversely.*

Because the proof is simple enough we give it here. Let ν be any coarse NLE outcome and let (σ, ρ, β) be a coarse NLE with outcome ν. Then, (σ, ρ, β) is a sequential equilibrium of $\Gamma(T)$ and the beliefs β are coarsely-straightforward. To prove the first part of the theorem we must show that ν satisfies the NWBR criterion.

So consider any action-message (a_0, t_0) that has probability zero under ν and that is a best reply for some type t_1 in some sequential equilibrium with outcome ν. To prove the first part of the theorem we must show that there is a sequential equilibrium (σ', ρ', β') with outcome ν such that $\beta'(t|a_0, t_0) = 0$ for every t such that (a_0, t_0) is not a best reply for t in any sequential equilibrium with outcome ν.

We now construct the requisite sequential equilibrium (σ', ρ', β'). Define $\sigma' := \sigma$, define $\rho' := \rho$ except that $\rho'(\cdot|a_0, t_0) := \rho(\cdot|a_0, t_1)$, and define $\beta' := \beta$ except that $\beta'(\cdot|a_0, t_0) := \beta(\cdot|a_0, t_1)$. (If $t_1 = t_0$, then $(\sigma', \rho', \beta') = (\sigma, \rho, \beta)$.)

Since (a_0, t_0) has probability zero under σ and under σ', and since the opportunities for profitable deviations for the sender have decreased under (σ', ρ', β') relative to (σ, ρ, β) it is easy to see that (σ', ρ', β') is a sequential equilibrium with the same outcome as (σ, ρ, β), namely ν.

Consider any t such that (a_0, t_0) is not a best reply for t in any sequential equilibrium with outcome ν. To prove the first part of the theorem, it suffices to show that $\beta'(t|a_0, t_0) = 0$. Notice that $t \neq t_1$ by the definitions of t and t_1. Since $\beta'(t|a_0, t_0) := \beta(t|a_0, t_1)$, and since β is coarsely straightforward, if $\beta(t|a_0, t_1)$ were positive, then (because $t \neq t_1$) it would be the case that (a_0, t_1) is a best reply for t in the sequential equilibrium (σ, ρ, β), which would imply that (a_0, t_0) is a best reply for t in the seqential equilibrium (σ', ρ', β'). But this would contradict the definition of t. We conclude that $\beta'(t|a_0, t_0) = 0$, as desired. This proves the first part of the theorem. The second part follows from Example 6.6 below.

Remark 6.4 *It can be shown that if an outcome of the base game is a coarse NLE type-action distribution, then it is a sequential equilibrium outcome of the base game that satisfies the NWBR criterion.*[23]

We close this section with two examples. The first example provides a coarse NLE type-action distribution that is not an NLE type-action distribution and the second example provides a sequential equilibrium of the base game that satisfies the never a weak best reply criterion but whose outcome is not a coarse NLE type-action distribution.

Example 6.5 Consider once again the base game Γ_0 from Example 5.1. Recall that one of its two sequential equilibrium outcomes has both types of S choosing l and has R responding to l by choosing C. We denoted this outcome by $\nu_{l,C}$ and we showed that it is not a natural language equilibrium type-action distribution for Γ_0. We will show that $\nu_{l,C}$ is, however, a coarse natural language equilibrium type-action distribution for Γ_0.

Consider the sequential equilibrium of $\Gamma(\{t_1, t_2\})$ in which, both types of S randomize by choosing (l, t_1) and (l, t_2) with probability one-half each; R chooses C after (l, t_1) and after (l, t_2), R chooses U after (r, t_1), and R chooses D after (r, t_2); and

[23] I thank Songzi Du for asking whether such a result were true.

R's beliefs put probability $1/2$ on each sender-type after any on-path action-message (l, t), and put probability one on t after any off-path action-message (r, t).

It is easy to check that this assessment is a sequential equilibrium of $\Gamma(\{t_1, t_2\})$. To see that the beliefs are coarsely-straightforward, observe that after any off-path message (r, t), R believes that the statement t claiming "my type is t" is true since R's subsequent beliefs place probability one on t. Hence, the condition for coarsely-straightforward beliefs is trivially satisfied (because $\beta(t|a, t') > 0$ fails to hold for any off-path (a, t') and any $t \neq t'$).

Since the constructed sequential equilibrium induces the type-action distribution $\nu_{l,C}$ and its beliefs are coarsely-straightforward, we have shown that $\nu_{l,C}$ is a coarse natural language equilibrium type-action distribution.

Example 6.6 The following example is due to Cho and Kreps (1987). In the base game the sender has three types, t_1, t_2, and t_3, and two actions l and r; and the receiver has three actions, U, C, and D. If any sender-type chooses l, then both players receive zero regardless of R's action, i.e., $u_S(t, l, z) = u_R(t, l, z) = 0$, for all t and z.[24]

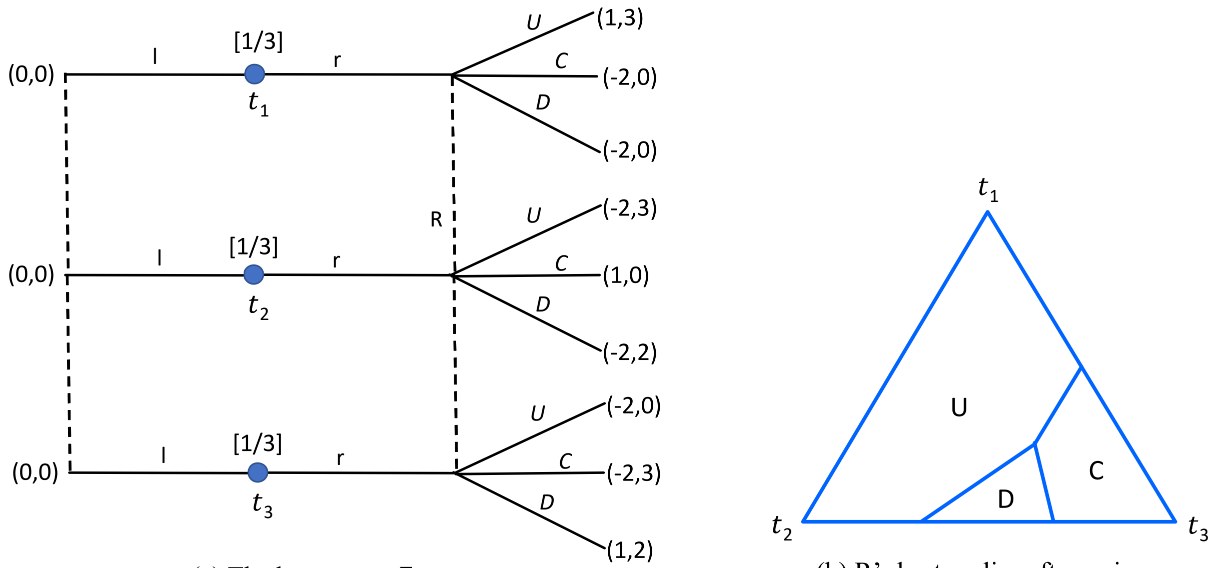
The game-tree in panel (a) of Figure 4 depicts the relevant part of the base-game Γ_0, leaving out the receiver's actions U, C, and D following a choice of l by S because they are redundant—they all lead to the payoff vector $(0, 0)$.

Panel (b) of Figure 4 shows R's best replies as a function of his beliefs over the sender's types after the sender chooses r. For example U is R's unique best reply when his beliefs are in the interior of the region labelled U, while both U and C are optimal for R when his beliefs are on the line segment that is between regions U and C. In particular, there is a unique belief for R that makes all three choices U, C, and D optimal for him and it lies at the intersection of the three interior line segments.

[24] If one worries that these payoffs are not generic, then one can instead assume that after any sender-type chooses l, U is strictly optimal for R and that $u_S(t, l, U) = u_R(t, l, U) = 0$ for all t. Then, in every sequential equilibrium, the choice of l by any sender-type will result in a payoff of zero for both players.

This belief places probabilities 1/4, 1/4, 1/2 on types t_1, t_2, and t_3, respectively.

Consider any outcome ν of Γ_0 in which every sender-type chooses l. Note that ν is a sequential equilibrium outcome of Γ_0 because R's beliefs after the off-path action r can put probability 1/4 each on t_1 and t_2 in which case mixing uniformly over his actions U, D, and C is optimal for R and makes deviating to r unprofitable for each sender-type. Let (σ, ρ, β) denote this sequential equilibrium.



(a) The base-game Γ_0

(b) R's best replies after r given beliefs about S's type.

Figure 4.

Let us show that (σ, ρ, β) satisfies the NWBR criterion. To do so, it suffices to display, for each sender-type t_i, a sequential equilibrium whose outcome is ν and in which type t_i is indifferent between l and r. Let us begin with t_1. As above, we can specify that R's beliefs after the off-path action r put probability 1/4 each on t_1 and t_2. Since this makes R indifferent between all of his actions, it is optimal for him to mix on just U and C, and in a way that gives t_1 an expected payoff of zero (i.e., U is twice as likely as C). Since this makes deviating to r unprofitable for types t_2 and t_3, we are done. A similar construction works also for types t_2 and t_3 (except that R mixes over C and D to make t_3 indifferent). We omit the argument. Hence, the outcome ν satisfies the NWBR criterion.

We next show that ν is not a coarse NLE type-action distribution. Let (σ, ρ, β) is a sequential equilibrium of $\Gamma(T)$ with type-action distribution ν. It suffices to show that the beliefs β are not coarsely straightforward. Suppose, by way of contradiction that β is coarsely straightforward, and consider the off-path action-message (r, t_1). Notice that, after S chooses r, there is no response by R that simultaneously gives any two sender-types an expected payoff of zero and so R's response must make r a best reply for at most one of t_2 and t_3. Hence, because β is coarsely straightforward, either $\beta(t_2|r, t_1) = 0$ or $\beta(t_3|r, t_1) = 0$.

Since either $\beta(t_2|r, t_1)$ or $\beta(t_3|r, t_1)$ is zero, Figure 4 panel (b) indicates that R's beliefs are either on the line joining t_1 and t_3 ($\beta(t_2|r, t_1) = 0$) or on the line joining t_1 and t_2 ($\beta(t_3|r, t_1) = 0$). Either way, R's optimal reply is some mixture—possibly degenerate—over U and C. But then (r, t_1) is not a best reply for t_3 and so, by coarsely-straightforward beliefs, we must have $\beta(t_3|r, t_1) = 0$ implying that R's beliefs are on the line joining t_1 and t_2. Given such beliefs, $R's$ unique optimal response is U, making (r, t_1) a profitable deviation for t_1 and contradicting equilibrium. Hence, β is not coarsely straightforward and we conclude that ν is not a coarse NLE type-action distribution.

It turns out that, for this example, the essentially unique coarse NLE type-action distribution coincides with the essentially unique Kohlberg-Mertens stable outcome of Γ_0 in which t_1 chooses r and R responds with U, and in which both t_2 and t_3 choose l (and R responds with any choice since all are optimal).

7 Related Literature

Farrell (1985, 1993), focusing on pure communication games, initiated the study of language in games, introduced the important concept of a credible neologism, and pointed out the conflict with equilibrium existence. Matthews et. al. (1991) and Matthews and Postlewaite (1994), modify Farrell's ideas to resolve various inconsistencies, and Grossman and Perry (1986) extend Farrell's ideas to signaling games as well as to sequential bargaining games. None of these authors resolve the existence problem.

Myerson (1991) considers a finite extensive form game setting in which, after nature's move, at each of finitely many dates, exactly one player moves and receives some information about the history of play. Thus, sender-receiver games are a special case of Myerson's (1991) model. Myerson (1991) takes an approach that is reminiscent of the classical approaches to defining credible statements. The main distinguishing feature is that Myerson allows only mediated communication rather than direct player to player communication. Mediated communication yields a set of incentive compatible mechanisms for the mediator that is convex, which allows Myerson (1991) to avoid the non-existence problems that plagued the classical studies.[25] Despite this important success, our interest here is in direct player to player communication.

Mailath et. al. (1993), in response to the "Stiglitz critique,"[26] propose an equilibrium refinement called undefeated equilibrium that cannot be upset by any off-path statement that is of the form "my type is this set" and that is not subject to the Stiglitz critique. Undefeated equilibria, which are restricted to pure strategy profiles, do not always exist. However, if the underlying game has ordered action and types sets and satisfies certain conditions that include single-crossing, then undefeated equilibria are shown to exist whenever a pure strategy sequential equilibrium exists.

In a learning model, Clark and Fudenberg (2021) study the equilibrium steady states of a signaling game in which senders can make statements of the form "my type is in this set." If (long-lived) receivers are initially trusting, Clark and Fudenberg identify necessary conditions for a strategy-profile to be a steady state and show

[25] In an incentive compatible mechanism for a signaling game, Myerson's (1991) mediator, after receiving the sender's (truthfully) reported type, sometimes stochastically recommends actions for the receiver that are worse for that sender-type than other receiver actions that, if recommended, would be carried out. This can never happen with direct player to player communication and so expands the set of equilibria.

[26] See footnote 2.

that these conditions are satisfied by an equilibrium concept that they call "justified communication equilibrium." These equilibria bear some connection to classic refinements such as Cho and Kreps' (1987) intuitive criterion and D1 criterion. Natural language equilibrium always (at least weakly) refines Clark and Fudenberg's (2021) justified communication equilibrium.

Clark (2024) introduces two equilibrium refinements for signaling games, "robust neologism-proofness" and the "flexible speech intuitive criterion." Both of these equilibrium refinements can fail to exist for the same reasons as the classical approaches and so there remains a fundamental difficulty. Even so, Clark (2024) shows that both kinds of equilibria do exist in certain classes of signaling games. In one of these classes, called monotonic games, natural language equilibrium, being a refinement of D1, always refines (at least weakly) both of these equilibrium concepts.

We turn next to work that perturbs the sender-receiver game of pure communication so that it becomes a signaling game in which the sender has payoff-relevant actions. In these papers, messages are of the form "my type is t" and the authors impose either a cost of lying about one's type or a benefit to telling the truth about one's type and sometimes include additional modifications.

Kartik et. al. (2007) impose a cost of lying about one's type. They also add a small fraction of naive receivers. A main result is that, when the sender's type-space is unbounded above, there is a separating equilibrium that features exaggeration—each type overstates his true type.

Kartik (2009) also adds lying costs but assumes a bounded type space. He shows that monotonic strategies together with the bound on types implies partial pooling—hence no separating equilibrium exists. He also shows that a particular class of monotonic equilibria involve exaggeration.

Demichelis and Weibull (2008) suppose that the cost of lying is infinitesimal in that players have lexicographic preferences in which truth is secondary to material payoffs. In generic and symmetric n-player coordination games with a unique Pareto efficient outcome, they show that with such lexicographic preferences for truth, a Nash equilibrium component is evolutionarily stable if and only if it yields the Pareto efficient outcome.

Chen (2011), introduces small fractions of behavioral types for both the sender and receiver similar to Crawford (2003). Honest sender-types truthfully reveal their type and naive receivers believe whatever statement the sender makes. In the limit as behavioral types vanish, and under the Crawford and Sobel (1982), henceforth CS, regularity condition, equilibria involve exaggeration and must satisfy the "no incentive to separate" condition of Chen, Kartik, and Sobel (2008), and so only the most efficient equilibrium survives.

Hart et. al. (2017) study evidence games in which all sender-types prefer the same receiver action. The authors perturb the sender's strategy toward truth-telling, with an additional perturbation toward a sender-preference for truth-telling. With these two perturbations, they show that, in the limit as the perturbations vanish, all sequential equilibrium payoffs for the receiver—called the "principal" here—coincide with his best equilibrium under any commitment strategy for him. That is, commitment has no value for the receiver. Interestingly, the equilibria identified by Hart et. al. (2017) must be what we call "coarse" natural language equilibria (Section 6).

Somewhat less directly related is work on pure communication games. Recall that, in this class of games, natural language equilibrium has no refinement power at all. A main difficulty here is refining away the babbling equilibrium without refining away all equilibria.

Rabin (1990, Section III) and Zapater (1997) modify Farrell's (1993) neologism-proofness concept so as to avoid the equilibrium existence problem by weakening the definition of credible statements using Bernheim (1984) and Pearce's (1984) notion of rationalizability.

Blume and Sobel (1995) define credible messages by considering a particular payoff dominance relation and its associated von Neumann and Morgenstern (1944), henceforth vNM, stable sets of sequential equilibria. They show that there always exists at least one vNM stable set and that in common interest games any equilibrium in a vNM stable set must be efficient.[27]

Blume (2023) refines equilibria in pure communication games and avoids the equilibrium existence problem by using iterative best replies instead of equilibrium payoffs to determine whether a message is credible. Blume (2023) also considers a number of other interesting issues including determination of the equilibrium meaning of statements in the language and the effects of language when that language is not common knowledge.

In the canonical CS model, Gordon et. al. (2023) show that, when the sender is restricted to monotonic strategies, and when the CS regularity condition holds, iterative best reply dynamics and a particular order of iterative elimination of weakly dominated strategies both lead to the most efficient equilibrium.

Sémirat and Forges (2023) also consider an iterative best-reply dynamic for the canonical CS model and show that it always converges to an equilibrium that is undefeated in the sense of Mailath et. al. (1993).

A number of the above studies restrict the sender to monotonic strategies. To be sure, when the sender's types and messages can be ordered, monotonicity is a natural assumption that warrants serious study. However, monotonicity is sometimes interpreted as an indirect way to study language in games. The clearest statement of this point of view is in Gordon et. al. (2023) who interpret the restriction to monotonic strategies as (p.2) "...a way to incorporate 'exogenous meaning' into communication: players enter the strategic setting with a shared ordering of messages and it is common knowledge that they will behave in a way that is consistent with this ordering."

We are wary of this perspective. Incorporating exogenous, or literal, meaning into communication is one thing. But assuming common knowledge of some aspect

[27] Von Neumann and Morgenstern (1944) stable sets, a solution concept for cooperative games, is entirely distinct from Kohlberg and Mertens (1986) stable sets, a solution concept for non-cooperative games.

of player behavior—e.g., that higher sender-types always send higher messages—is quite another, going well beyond postulating the existence of a commonly understood language with conventions for its use.

In any event, we seek to understand how language and conventions about language affect strategic behavior, an effort that seems best served by refraining from making assumptions about that behavior a priori.

8 Proofs of Theorems

Proof of Theorem 4.3. By Theorem 4.7, there is a natural language equilibrium whose type-action distribution is the outcome of a sequential equilibrium of the base game. ■

Our first lemma establishes that even if we allowed the message space M to contain a continuum of messages, there exists a fixed finite number of messages that suffices to support any Nash or sequential equilibrium type-action distribution of $\Gamma(M)$. Related results on the number of messages required to support any equilibrium payoff rather than type-action distribution have been obtained by Heumann (2020) as well as by Koessler, Laclau, and Tomala (2024). All of these results, including our own, make use of Caratheodory's theorem. We presume for the purposes of the following lemma and proof that when M has a continuum of elements, it is endowed with a sigma algebra of measurable sets and that strategies are appropriately measurable. These technical details are routine and so are ignored here.

Lemma 8.1 *Let M be any non-empty—perhaps even uncountably infinite—message space. Then any Nash (resp., sequential) equilibrium type-action distribution of $\Gamma(M)$ is a Nash (resp., sequential) equilibrium type-action distribution of $\Gamma(T \times Z)$.*

Proof. Let $M^* := T \times Z$, let $K = |T \times Z|$, and let $m_1^*, ..., m_K^*$ be all of the elements of M^*. Considering first the result for Nash equilibria, let (σ_ν, ρ_ν) be a Nash equilibrium of $\Gamma(M)$ whose type-action distribution is ν. We will construct a Nash equilibrium, (σ^*, ρ^*) of $\Gamma(M^*)$ whose type-action distribution is ν.

Let $A^+ := \{a : \sum_t p(t)\sigma_\nu(a, M|t) > 0\}$ denote the set of on-path actions,[28] and let $(q(t|a,m))_{t \in T}$ be any version of the conditional probability distribution over T given (a,m), for any (a,m). Hence, for every t and a,

$$\int_C q(t|a,m) \sum_{t'} p(t')\sigma_\nu(a, dm|t') = p(t)\sigma_\nu(a, C|t) \text{ for every subset } C \text{ of } M.^{29} \quad (8.1)$$

Since (σ_ν, ρ_ν) is a Nash equilibrium, if for each a we define $C_a := \{m : \rho_\nu(\cdot|a,m)$ is optimal for R when R's beliefs are $q(\cdot|a,m)$ and S's action is $a\}$, then $\sigma_\nu(a, C_a|t) = \sigma_\nu(a, M|t)$ for every t and every a. Because the type-action distribution of (σ_ν, ρ_ν) is ν, the probability of any (t, a, z) is,

$$\begin{aligned} \nu(t,a,z) &= p(t) \int_M \rho_\nu(z|a,m)\sigma_\nu(a, dm|t). \\ &= \int_M \rho_\nu(z|a,m) q(t|a,m) \sum_{t'} p(t')\sigma_\nu(a, dm|t'), \end{aligned} \quad (8.2)$$

where the second equality follows from (8.1). Hence, for any $a \in A^+$,

$$\begin{aligned} \frac{\nu(t,a,z)}{\sum_{t'} p(t')\sigma_\nu(a, M|t')} &= \int_M \rho_\nu(z|a,m) q(t|a,m) \mu(dm|a), \\ &= \int_{C_a} \rho_\nu(z|a,m) q(t|a,m) \mu(dm|a), \end{aligned} \quad (8.3)$$

where $\mu(dm|a) := \sum_{t'} p(t')\sigma_\nu(a, dm|t') / \sum_{t'} p(t')\sigma_\nu(a, M|t')$ and where the second equality follows because $\mu(C_a|a) = \mu(M|a)$. Notice that $\mu(C_a|a) = \mu(M|a) = 1$ for every $a \in A^+$.

[28] Throughout the proof, for any subset C of M, we write $\sigma_\nu(a, C|t)$ instead of the more cumbersome $\sigma_\nu(\{a\} \times C|t)$ and when $C = \{m\}$ is a singleton we write $\sigma_\nu(a, m|t)$.

[29] By "every" subset of M, we will always mean every measurable subset.

For each $a \in A^+$, define

$$\nu_a := \left(\frac{\nu(t,a,z)}{\sum_{t'} p(t')\sigma_\nu(a,M|t')} \right)_{t\in T, z\in Z}.$$

Notice that ν_a is in $\mathbb{R}^{|T||Z|-1}$ since its coordinates sum to unity by (8.2).

Since $\mu(C_a|a) = 1\ \forall a \in A^+$, (8.3) and Lemma 1 in Appendix A of Pearce (1984) imply that $\nu_a \in \text{co}\{(\rho_\nu(z|a,m)q(t|a,m))_{t\in T,z\in Z} : m \in C_a\} \subseteq \mathbb{R}^{|T\times Z|-1} = \mathbb{R}^{K-1}$. Hence, by Caratheodory's theorem, ν_a can be written as a convex combination of no more than K elements of $\{(\rho_\nu(z|a,m)q(t|a,m))_{t\in T,z\in Z} : m \in C_a\}$.

Hence, for every $a \in A^+$, there are K not necessarily distinct elements $m_{a,1}, ..., m_{a,K}$ of C_a, and there are K positive real numbers $\alpha_{a,1}, ..., \alpha_{a,K}$ summing to $\sum_{t'} p(t')\sigma_\nu(a,M|t')$ such that,

$$\nu(t,a,z) = \sum_{k=1}^{K} \alpha_{a,k}\rho_\nu(z|a,m_{a,k})q(t|a,m_{a,k})),\ \forall t, \forall a \in A^+, \forall z. \tag{8.4}$$

Summing over z and noting that $\sum_z \nu(t,a,z) = p(t)\sigma_\nu(a,M|t)$ by the first equality in (8.2) gives,

$$p(t)\sigma_\nu(a,M|t) = \sum_{k=1}^{K} \alpha_{a,k}q(t|a,m_{a,k}),\ \forall t, \forall a \in A^+. \tag{8.5}$$

Recall that $M^* = \{m_1^*, ..., m_K^*\}$. Define the strategy σ^* for S in $\Gamma(M^*)$ so that,

$$\sigma^*(a, m_k^*|t) := \frac{\alpha_{a,k}q(t|a,m_{a,k})}{p(t)},\ \forall t, \forall a \in A^+, \forall k = 1, ..., K, \tag{8.6}$$

and so that $\sigma^*(a, m_k^*|t) = 0$ for any k whenever $a \notin A^+$.

Fixing any $m_0 \in M$, define the strategy ρ^* for R in $\Gamma(M^*)$ as follows. For any $m_k^* \in M^*$,

$$\rho^*(\cdot|a, m_k^*) := \begin{cases} \rho_\nu(\cdot|a, m_{a,k}), & \text{if } a \in A^+ \\ \rho_\nu(\cdot|a, m_0), & \text{if } a \notin A^+. \end{cases}$$

To see that σ^* is a valid strategy for S, notice that $\sigma^*(a, m_k^*|t) \geq 0$ for every a, every k, and every t, and that, for every t,

$$\begin{aligned}
\sigma^*(A \times M^*|t) &= \sum_{a \in A^+} \sum_{k=1}^{K} \sigma^*(a, m_k^*|t) \\
&= \sum_{a \in A^+} \sum_{k=1}^{K} \frac{\alpha_{a,k} q(t|a, m_{a,k})}{p(t)} \\
&= \sum_{a \in A^+} \sigma_\nu(a, M|t) \\
&= 1,
\end{aligned}$$

where the first equality follows because $\sigma^*(a, m_k^*|t) = 0$ whenever $a \notin A^+$, the second equality follows from the definition of σ^*, the third equality follows from (8.5), and the fourth equality follows from the definition of A^+. So to complete the proof we need only show that (σ^*, ρ^*) is a Nash equilibrium with type-action distribution ν. We first show that the type-action distribution is ν.

Consider any (t, a, z). If $a \notin A^+$, then $\nu(t, a, z)$ is zero by the definition of A^+, and the probability of (t, a, z) under (σ^*, ρ^*) is also zero because $\sigma^*(a, M^*|t) = 0$ for every t. If $a \in A^+$, then, because $\sigma^*(A^+ \times M^*|t) = 1$, the probability of (t, a, z) under (σ^*, ρ^*) is,

$$\begin{aligned}
p(t) \sum_{k=1}^{K} \sigma^*(a, m_k^*|t) \rho^*(z|a, m_k^*) &= p(t) \sum_{k=1}^{K} \frac{\alpha_{a,k} q(t|a, m_{a,k})}{p(t)} \rho^*(z|a, m_k^*) \\
&= \sum_{k=1}^{K} \alpha_{a,k} q(t|a, m_{a,k}) \rho_\nu(z|a, m_{a,k}) \\
&= \nu(t, a, z), \text{by (8.4)},
\end{aligned}$$

and so the outcome of (σ^*, ρ^*) is ν. It remains only to show that (σ^*, ρ^*) is a Nash equilibrium.

We first show that σ^* is a best reply for S against ρ^*. Since the type-action distribution is ν under both (σ^*, ρ^*) and (σ_ν, ρ_ν), the sender's payoff under (σ^*, ρ^*)

is the same as under (σ_ν, ρ_ν). For any type t and action a for the sender, consider the set of all distributions over z that S can induce by varying his message $m \in M^*$ given that R uses the strategy ρ^*. By the definition of ρ^*, every distribution in this set is a distribution that was available to S when R used the strategy ρ_ν. Hence, the maximum payoff that is achievable for S against ρ^* is no greater than that against ρ_ν. Since S's payoff has not changed, and since S's payoff is the largest that is achievable against ρ_ν, it is the largest that is achievable against ρ^*. That is, σ^* is a best reply for S against ρ^*. It remains only to show that ρ^* is a best reply for R against σ^*.

Because $\sigma^*(A^+ \times M^*|t) = 1$ for every t, to show that ρ^* is a best reply for R against σ^*, it suffices to show, for each of the finitely many $(a, m_k^*) \in A^+ \times M^*$ given positive probability by σ^*, that $\rho^*(\cdot|a, m_k^*)$ is optimal for R when his beliefs over T are equal to the Bayes posterior over T given (a, m_k^*) and S's action is a. Since for $(a, m_k^*) \in A^+ \times M^*$, $\rho^*(\cdot|a, m_k^*) = \rho_\nu(\cdot|a, m_{a,k})$ and since $m_{a,k} \in C_a$ implies that $\rho_\nu(\cdot|a, m_{a,k})$ is optimal for R given the distribution $q(\cdot|a, m_{a,k})$ over T and given that S's action is a, it suffices to show that $q(\cdot|a, m_{a,k})$ is the Bayes posterior over T given any $(a, m_k^*) \in A^+ \times M^*$.

So consider any $(a, m_k^*) \in A^+ \times M^*$. By (8.6), for every t, $p(t)\sigma^*(a, m_k^*|t) := \alpha_{a,k} q(t|a, m_{a,k})$. Summing over t and noting that $\sum_t q(t|a, m_{a,k}) = 1$ gives $\sum_t p(t)\sigma^*(a, m_k^*|t) = \alpha_{a,k} > 0$. Hence, (a, m_k^*) is given positive probability by σ^* and the Bayes posterior probability of t given (a, m_k^*) is,

$$\frac{p(t)\sigma^*(a, m_k^*|t)}{\sum_{t'} p(t')\sigma^*(a, m_k^*|t')} = \frac{\alpha_{a,k} q(t|a, m_{a,k})}{\alpha_{a,k}} = q(t|a, m_{a,k}), \tag{8.7}$$

as desired. This proves the result for Nash equilibria.

If the type-action distribution ν were instead that of a sequential equilibrium $(\sigma_\nu, \rho_\nu, \beta_\nu)$, then we would proceed exactly as above to obtain (σ^*, ρ^*). It remains to define beliefs β^*. As shown above when establishing (8.7), every $(a, m_k^*) \in A^+ \times M^*$ is on-path for σ^*. Hence, for every $(a, m_k^*) \in A^+ \times M^*$, we can define $\beta^*(\cdot|a, m_k^*)$ to be the Bayes posterior over T given (a, m_k^*). Then, because (σ^*, ρ^*) is a Nash equilibrium,

$\rho^*(\cdot|a,m_k^*)$ is a best reply for R when S chooses a and R's beliefs are $\beta^*(\cdot|a,m_k^*)$. It remains only to define $\beta^*(\cdot|a,m_k^*)$ for any (a,m_k^*) such that a is not in A^+. Since, by our construction of ρ^*, for any a outside A^+, $\rho^*(\cdot|a,m_k^*) = \rho_\nu(\cdot|a,m_0)$, and since $\rho_\nu(\cdot|a,m_0)$ is a best-reply for R when S chooses a and R's beliefs are $\beta_\nu(\cdot|a,m_0)$ (because $(\sigma_\nu,\rho_\nu,\beta_\nu)$ is a sequential equilibrium), we can define $\beta^*(\cdot|a,m_k^*) = \beta_\nu(\cdot|a,m_0)$ for any (a,m_k^*) such that a is outside A^+. Then $(\sigma^*,\rho^*,\beta^*)$ is a sequential equilibrium with type-action distribution ν. This proves the result for sequential equilibrium and completes the proof. ■

Lemma 8.2 *A type-action distribution ν is an NLE type-action distribution iff ν is a Nash equilibrium type-action distribution of $\Gamma(M)$ for some finite set of messages M, and, for every a satisfying $\sum_{t,z}\nu(t,a,z) = 0$ and for every $q \in \Delta(T)$, there exist $\beta(\cdot|a,q) \in \Delta(T)$ and $\rho(\cdot|a,q) \in \Delta(Z)$ such that,*
(a) $\rho(\cdot|a,q)$ is a best reply for R after a when his beliefs are $\beta(\cdot|a,q)$, and
(b) for every t,t', $\pi_S(t) \geq \sum_z \rho(z|a,q)u_S(t,a,z)$ with equality if $\beta(t|a,q)/\beta(t'|a,q) > q(t)\backslash q(t')$,[30]
where $\pi_S(t) := \sum_{t,z}\nu(t,a,z)u_S(t,a,z)/p(t)$ is sender-type t's payoff under ν.

Proof. Let us begin with the "only if" statement. Suppose that ν is an NLE type-action distribution. Then there is a sequential equilibrium $(\sigma^*,\rho^*,\beta^*)$ of $\Gamma(Q)$ with straightforward beliefs whose type-action distribution is ν. By Lemma 8.1, there is a finite message space, namely $M = T \times Z$, such that ν is a Nash equilibrium of $\Gamma(M)$, proving the first part of the "only if" statement. To prove the second part, choose any a such that $\sum_{t,z}\nu(t,a,z) = 0$ and choose any $q \in Q$ (we will allow any $q \in \Delta(T)$ shortly). Since $(\sigma^*,\rho^*,\beta^*)$ is a sequential equilibrium of $\Gamma(Q)$ with straightforward beliefs, $\beta^*(\cdot|a,q)$ and $\rho^*(\cdot|a,q)$ satisfy (a) and (b). Hence we have shown that for every a satisfying $\sum_{t,z}\nu(t,a,z) = 0$ and for every $q \in Q$, there exist $\beta(\cdot|a,q) \in \Delta(T)$ and $\rho(\cdot|a,q) \in \Delta(Z)$ such that (a) and (b) hold. To complete the argument we

[30] See footnote 11.

must extend this result from any $q \in Q$ to any $q \in \Delta(T)$. But, since Q is dense in $\Delta(T)$, this extension follows easily by defining, for any a and for any $q \in \Delta(T)\backslash Q$, $(\rho(\cdot|a,q), \beta(\cdot|a,q))$ to be any accumulation point of $(\rho^*(\cdot|a,q_n), \beta^*(\cdot|a,q_n))$, where q_n is any sequence in Q that converges to q.

To prove the "if" statement, suppose that ν is a Nash equilibrium type-action distribution of $\Gamma(M)$ for some finite set of messages M, and that, for every a satisfying $\sum_{t,z} \nu(t,a,z) = 0$ and for every $q \in \Delta(T)$, there exist $\beta(\cdot|a,q) \in \Delta(T)$ and $\rho(\cdot|a,q) \in \Delta(Z)$ such that (a) and (b) hold. We must show that ν is an NLE. We will do so by constructing an NLE whose type-action distribution is ν.

Let $(\hat{\sigma}, \hat{\rho})$ be a Nash equilibrium of $\Gamma(M)$ whose type-action distribution is ν. To each $m \in M$, associate a distinct $q_m \in Q$. Define an assessment $(\sigma^*, \rho^*, \beta^*)$ for $\Gamma(Q)$ as follows. For every (a,m) that is on-path for $\hat{\sigma}$, define $\sigma^*(a, q_m|\cdot) := \hat{\sigma}(a,m|\cdot)$, define $\rho^*(\cdot|a,q_m) := \hat{\rho}(\cdot|a,m)$, and define $\beta^*(\cdot|a,q_m)$ to be the Bayes' posterior under $\hat{\sigma}$ given (a,m). For any (a,q) for which $\sigma^*(a,q|\cdot)$, $\rho^*(\cdot|a,q)$, and $\beta^*(\cdot|a,q)$ have yet to be defined, define $\sigma^*(a,q|\cdot) := 0$, define $\rho^*(\cdot|a,q) := \rho(\cdot|a,q)$, and define $\beta^*(\cdot|a,q) := \beta(\cdot|a,q)$.

Clearly, the type-action distribution of $(\sigma^*, \rho^*, \beta^*)$ is ν. So it remains only to show that $(\sigma^*, \rho^*, \beta^*)$ is an NLE. That $(\sigma^*, \rho^*, \beta^*)$ is a sequential equilibrium of $\Gamma(Q)$ follows from the fact that the players' payoffs are unchanged (since the type-action distribution is unchanged) and any unilateral deviation by S leads to a continuation in which either R responds with $\hat{\rho}$,and so cannot be profitable since $(\hat{\sigma}, \hat{\rho})$ is a Nash equilibrium, or R responds with ρ and so cannot be profitable by (b). Hence σ^* is optimal against ρ^*. Also, after any (a,q), R's beliefs and response are either those according to the Nash equilibrium $(\hat{\sigma}, \hat{\rho})$—with Bayes posterior beliefs on-path—or are those given by β and ρ satisfying (a). In either case, R's play is sequentially rational and his beliefs are Bayes' consistent on-path. That the beliefs β^* are straightforward follows directly from (b). Hence, $(\sigma^*, \rho^*, \beta^*)$ is an NLE as desired. ■

Lemma 8.3 *If M is any non-empty finite set of messages and ν is a sender-stable* *type-action distribution of* $\Gamma(M)$*, then* ν *is a natural language equilibrium type-action distribution.*

Proof. Since ν is a sender-stable type-action distribution of $\Gamma(M)$, ν is, in particular, a Nash equilibrium type-action distribution of $\Gamma(M)$. Hence, it suffices to establish conditions (a) and (b) of Lemma 8.2. Let $((\pi_S(t))_{t\in T}, \pi_R)$ be the payoff from ν, let a_0 be any action such that $\sum_{t,z} \nu(t, a_0, z) = 0$ and let q_0 be any element of $\Delta(T)$. The remainder of the proof shows how to define the requisite $\rho(\cdot|a_0, q_0)$ and $\beta(\cdot|a_0, q_0)$.

Consider first the case in which q_0 gives strictly positive probability to every type, i.e., $q_0(t) > 0$ for every t. Then we can define a strictly mixed strategy σ_0 for S in the game $\Gamma(M)$ as follows. Recall that p is the prior over S's types. Choose $\delta > 0$ so that $\delta q_0(t)/p(t) < 1$ for every t, and let m_0 be any message in M. Let σ_0 be any strictly mixed strategy for S in $\Gamma(M)$ such that $\sigma_0(a_0, m_0|t) = \delta q_0(t)/p(t)$ for every t.[31] Notice that if S uses σ_0 and R observes (a_0, m_0), then R's Bayes posterior is q_0.

For each $n = 1, 2, ...$, consider the game, let us call it G_n, in which, when S chooses σ and R chooses ρ in $\Gamma(M)$, S's strategy is perturbed slightly to $(1 - \frac{1}{n})\sigma + \frac{1}{n}\sigma_0$, but R's strategy ρ is not perturbed. Since ν is a sender-stable type-action distribution of $\Gamma(M)$ there is, for every $n = 1, 2, ...$, a Nash equilibrium (σ_n, ρ_n) of G_n whose type-action distribution, ν_n say, converges to ν as $n \to \infty$. In particular, because $\sum_{t,z} \nu(t, a_0, z) = 0$, we must have $\sigma_n(a_0, m_0|t) \to_n 0$ for every t.

Since, for each n, $(1 - \frac{1}{n})\sigma_n + \frac{1}{n}\sigma_0$ is completely mixed, we may define $\beta_n(\cdot|a_0, m_0)$ to be the Bayes posterior over T given (a_0, m_0), and we may assume without loss of generality that $\lim_n \beta_n(t|a_0, m_0)$ exists for every t and that $\lim_n \rho_n(z|a_0, m_0)$ exists for every z. Then we can define $\beta^*(t|a_0, q_0) = \lim_n \beta_n(t|a_0, m_0)$ for every t and we can define $\rho^*(z|a_0, q_0) = \lim_n \rho_n(z|a_0, m_0)$ for every z.

Let us show that (a) in Lemma 8.2 holds. Since in the nth perturbed game G_n, the perturbation, $(1 - \frac{1}{n})\sigma_n + \frac{1}{n}\sigma_0$, of S's equilibrium strategy σ_n gives (a_0, m_0)

[31] Such a σ_0 exists because $\sum_{t,z} \nu(t, a_0, z) = 0$ implies that a_0 is not the only action and hence that (a_0, m_0) is not the only action-message (it is possible that m_0 is the only message).

positive probability, and because $\beta_n(\cdot|a_0, m_0)$ is the Bayes posterior given (a_0, m_0), equilibrium implies that R's mixture $\rho_n(\cdot |a_0, m_0)$ is a best reply after a_0 given the beliefs $\beta_n(\cdot|a_0, m_0)$. Hence, by continuity, $\rho^*(\cdot|a_0, q_0)$ is a best reply for R after a_0 given the beliefs $\beta^*(\cdot|a_0, q_0)$.

We next show that (b) in Lemma 8.2 holds. Let t be any type and let $\pi^n_S(t)$ denote t's equilibrium payoff in the nth perturbed game G_n. Then $\pi^n_S(t) \geq \sum_z \rho_n(z|a_0, m_0)u_S(t, a_0, z)$ because deviating to (a_0, m_0) cannot be strictly profitable for t. Since the equilibrium type-action distribution ν_n of G_n converges to ν, $\pi^n_S(t)$ converges to $\pi_S(t)$. Hence, taking limits of both sides of the inequality in the previous sentence yields the inequality part of (b). It remains to show that $\pi_S(t) = \sum_z \rho^*(z|a, q)u_S(t, a, z)$ if $\beta^*(t|a_0, q_0)/\beta^*(t'|a_0, q_0) > q_0(t)/q_0(t')$ for some t'.

By the definition of $\beta_n(t|a_0, m_0)$ and of σ_0, direct computation gives,

$$\beta_n(t|a_0, m_0) = \lambda_n q_0(t) + (1 - \lambda_n)\gamma_n(t), \text{ for every } t, \tag{8.8}$$

where $\lambda_n := \delta/(\delta+(n-1)\sum_{t'} p(t')\sigma_n(a_0, m_0|t'))$, and $\gamma_n(t) := p(t)\sigma_n(a_0, m_0|t)/\sum_{t'} p(t')\sigma_n(a_0, m_0|t')$. Without loss of generality, we may assume that all limits converge (note that $\lambda_n, \gamma_n(t) \in [0, 1]$). Hence, letting $\lambda := \lim_n \lambda_n$ and $\gamma(t) := \lim_n \gamma_n(t)$ for every t, gives

$$\beta^*(t|a_0, q_0) = \lambda q_0(t) + (1 - \lambda)\gamma(t), \text{ for every } t. \tag{8.9}$$

Suppose that for some t and t', $\beta^*(t|a_0, q_0)/\beta^*(t'|a_0, q_0) > q_0(t)/q_0(t')$. We must show that $\pi_S(t) = \sum_z \rho^*(z|a_0, q_0)u_S(t, a_0, z)$. We claim that $\lambda < 1$. Otherwise, $\lambda = 1$ and (8.9) implies that $\beta^*(t|a_0, q_0) = q_0(t)$ for every t contradicting $\beta^*(t|a_0, q_0)/\beta^*(t'|a_0, q_0) > q_0(t)/q_0(t')$ and proving the claim.[32]

Next, we claim that $\gamma(t) > 0$. Otherwise, $\gamma(t) = 0$ and so, by (8.9), $\beta^*(t|a_0, q_0) = \lambda q_0(t)$ and $\beta^*(t|a_0, q_0)/\beta^*(t'|a_0, q_0) = \lambda q_0(t)/(\lambda q_0(t') + (1 - \lambda)\gamma(t')) \leq q_0(t)/q_0(t')$ contradicting $\beta^*(t|a_0, q_0)/\beta^*(t'|a_0, q_0) > q_0(t)/q_0(t')$ and proving the claim. But if

[32] It is possible that $\alpha < 1$ because $\lim_n \sigma_n(a_0, m_0|t) = 0$ for every t.

$\gamma(t) > 0$, then there is $\bar{n}$ such that $\gamma_n(t) > 0$ for all $n > \bar{n}$ which, by definition of $\gamma_n(t)$, implies that $\sigma_n(a_0, m_0|t) > 0$ for all $n > \bar{n}$. But $\sigma_n(a_0, m_0|t) > 0$ implies that (a_0, m_0) must be a best reply for t against ρ_n in the nth perturbed game G_n, which means that (a_0, m_0) yields t a payoff of $\pi_S^n(t)$ against ρ_n. Hence, $\pi_S^n(t) = \sum_z \rho_n(z|a_0, m_0)u_S(t, a_0, z)$, and taking the limit on both sides of the equality yields, $\pi_S(t) = \sum_z \rho^*(z|a_0, q_0)u_S(t, a_0, z)$ as desired. Thus we have shown that (a) and (b) in Lemma 8.2 hold for every (a_0, q_0) such that q_0 is strictly positive.

It remains only to show how to define $\rho^*(\cdot|a_0, q_0)$ and $\beta^*(\cdot|a_0, q_0)$ when q_0 gives some type probability zero. In that case, we may consider a sequence $q_1, q_2, ...$ of elements of $\Delta(T)$ converging to q_0 such that each q_n in the sequence gives every type positive probability. By what we have just shown, we can, for each n define $\rho_n^*(\cdot|a_0, q_n)$ and $\beta_n^*(\cdot|a_0, q_n)$ satisfying (a) and (b) in Lemma 8.2. We can then define $\rho^*(z|a_0, q_0) = \lim_n \rho_n^*(z|a_0, q_0)$ for every z and $\beta^*(t|a_0, q_0) = \lim_n \beta_n^*(t|a_0, q_0)$ for every t, where all limits can be assumed to exist without loss of generality. It is then a simple matter to check that, so defined, $\rho^*(\cdot|a_0, q_0)$ and $\beta^*(\cdot|a_0, q_0)$ satisfy the desired conditions. ■

Proof of Theorem 4.4. To prove the first part, suppose that ν is an NLE type-action distribution. Then, by Lemma 8.1, ν is a Nash equilibrium type-action distribution of $\Gamma(M)$ for $M = T \times Z$ and by the "only if" part of Lemma 8.2 conditions (a) and (b) there are satisfied. But then the proof of the "if" part of Lemma 8.2 shows that ν is the type-action distribution of an NLE that uses no more than $|M| = |T \times Z|$ messages. To prove the second part simply set $M = T \times Z$ in Lemma 8.3. ■

Proof of Theorem 4.7. We begin by showing that at least one sequential equilibrium outcome of the base-game Γ_0 is a natural language equilibrium type-action distribution. By Kohlberg and Mertens (1986, paragraph following Remark 1 on p.1027), stable outcomes exist for generic games. Consequently, we may choose a sequence of games Γ_1, Γ_2, ..., each obtained from the base-game Γ_0 by perturbing the players' utilities at the endpoints of the game-tree, so that the corresponding sequence of utility functions $u_{S,n}$ and $u_{R,n}$ for Γ_n converge to the utility functions u_S and u_R, respectively, for the base game Γ_0 and so that each game Γ_n has at least one stable outcome, ν_n say. Let $((\pi_{S,n}(t))_{t\in T}, \pi_{R,n})$ denote the payoff from ν_n.

Since the support of each ν_n is contained in the set of endpoints $T \times A \times Z$ of Γ_0, we may assume without loss of generality that ν_n converges to a distribution, ν say, over $T \times A \times Z$. Since each ν_n is a Nash equilibrium outcome of Γ_n, ν is a Nash equilibrium outcome of Γ_0. Let $((\pi_S(t))_{t\in T}, \pi_R)$ denote the payoff from ν and note that $\nu_n \to \nu$ implies that $((\pi_{S,n}(t))_{t\in T}, \pi_{R,n}) \to ((\pi_S(t))_{t\in T}, \pi_R)$. We must show that ν is a natural language equilibrium type-action distribution.

Recall that Γ_0 is strategically equivalent to the game $\Gamma(\{m_0\})$ obtained by adding a single cheap-talk message m_0.. The same is true of each Γ_n, i.e., adding a single cheap-talk message to Γ_n makes no strategic difference. Therefore, interpreting Γ_n as having exactly one cheap-talk message, and because each stable ν_n is, a fortiori, sender-stable, Lemma 8.3 implies that, for each n there is a natural language equilibrium $(\sigma_n, \rho_n, \beta_n)$ for Γ_n with type-action distribution ν_n and payoff $((\pi_{S,n}(t))_{t\in T}, \pi_{R,n})$. In particular, each $(\sigma_n, \rho_n, \beta_n)$ is a sequential equilibrium of $\Gamma_n(Q)$ with straightforward beliefs. Hence, for each n, we have that for every a and for every $q \in Q$,

(a) $\rho_n(\cdot|a, q)$ is a best reply for R after a when his beliefs are $\beta_n(\cdot|a, q)$, and

(b) for every t, t', $\pi_{S,n}(t) \geq \sum_z \rho_n(z|a, q)u_{S,n}(t, a, z)$ with equality if $\beta_n(t|a, q)/\beta_n(t'|a, q) > q(t)\backslash q(t')$,

where (b) holds for all (a, q) and not merely for off-path (a, q) by Remark 4.2.

Because Q is countably infinite, we may assume without loss of generality that each of the limits, $\lim_n \rho_n(z|a, q)$ and $\lim_n \beta_n(t|a, q)$ exists for every t, a, q, and z such that $q \in Q$. Therefore we may define $\rho^*(z|a, q) := \lim_n \rho_n(z|a, q)$, and $\beta^*(t|a, q) := \lim_n \beta_n(t|a, q)$ for every t, a, z, and every $q \in Q$. Hence, taking limits in (a) and (b) we have that for every a and for every $q \in Q$,

(a′) $\rho^*(\cdot|a, q)$ is a best reply for R after a when his beliefs are $\beta^*(\cdot|a, q)$, and (b$'$) for every t, t', $\pi_S(t) \geq \sum_z \rho^*(z|a,q)u_S(t,a,z)$ with equality if $\beta^*(t|a,q)/\beta^*(t'|a,q) > q(t)\backslash q(t')$.

Since Q is dense in $\Delta(T)$, we can extend (a$'$) and (b$'$) to every a and every $q \in \Delta(T)$ by defining, for any a and for any $q \in \Delta(T)\backslash Q$, $(\rho^*(\cdot|a,q), \beta^*(\cdot|a,q))$ to be any limit point of $(\rho^*(\cdot|a,q_n), \beta^*(\cdot|a,q_n))$, where q_n is any sequence in Q that converges to q. Hence, the type-action distribution ν whose payoff is $((\pi_S(t))_{t\in T}, \pi_R)$ is a Nash equilibrium type-action distribution of $\Gamma(\{m_0\})$ and conditions (a) and (b) of Lemma 8.2 are satisfied. Consequently, Lemma 8.2 implies that ν is a natural language equilibrium type-action distribution.

To prove the second, "if and only if" part, let us begin with the "if" part by supposing that ν is a stable outcome of the base game. We must show that ν is a natural language equilibrium type-action distribution. Since the base-game Γ_0 is equivalent to the game $\Gamma(\{m_0\})$ obtained by adding a single cheap-talk message m_0, ν is a stable type-action distribution of $\Gamma(\{m_0\})$. Hence, a fortiori, ν is a sender-stable type-action distribution of $\Gamma(\{m_0\})$. Lemma 8.3 then implies that ν is a natural language equilibrium type-action distribution.

Turning now to the "only if" part, we must show that for a generic set of base-game utilities, every outcome of the base game that is a natural language equilibrium type-action distribution is also a stable outcome of the base game. We will apply the following result due independently to Banks and Sobel (1987, Theorem 3) and Cho and Kreps (1987, Proposition 4).

(Base-Game) Stability Characterization Theorem.[33] There is a generic subset $\mathcal{U} \subseteq \mathbb{R}^{T\times A\times Z}$ of base-game utilities with the following property. If u_S and u_R are any utilities in $\mathcal{U}$ and if ν is any outcome of the base-game $\Gamma_0(u_S, u_R)$ and the payoff from ν is $((\pi_S(t))_{t\in T}, \pi_R)$, then ν is a stable outcome of the base game if and only if for every a such that $\sum_{t,z} \nu(t,a,z) = 0$ (i.e. for every off-path a) and for every $q \in \Delta(T)$

[33] Because this theorem includes the hypothesis that utilities are from a generic set, its "if" part does not formally apply to *all* utilities. This accounts for our separate proof of the "if" part of our result which does apply to all utilities.

there exist $\lambda \in [0,1]$, $\gamma \in \Delta(T)$, and $r \in \Delta(Z)$ such that,

(i) r is a best reply for R after action a when R's beliefs are $\lambda q + (1-\lambda)\gamma$, and

(ii) for every t, $\pi_S(t) \geq \sum_{z \in Z} r(z) u_S(t,a,z)$ and equality holds if $\lambda < 1$ and $\gamma(t) > 0$.

To complete our proof, let $\mathcal{U}$ be the generic set identified in the above stability characterization theorem, and let u_S and u_R be any elements of $\mathcal{U}$. Suppose that ν is an outcome of the base-game $\Gamma_0(u_S, u_R)$ and that the payoff from ν is $((\pi_S^*(t))_{t\in T}, \pi_R^*)$. Suppose also that ν is a natural language equilibrium type-action distribution. We must show that ν is a stable outcome of the base game $\Gamma_0(u_S, u_R)$.

Let a be any action such that $\sum_{t,z} \nu(t,a,z) = 0$ and ley q be any element of $\Delta(T)$. By the base-game stability characterization theorem, it suffices to find λ, γ, r satisfying (i) and (ii). By Lemma 8.2 there exist $\beta(\cdot|a,q) \in \Delta(T)$ and $\rho(\cdot|a,q) \in \Delta(Z)$ such that,

(a) $\rho(\cdot|a,q)$ is a best reply for R after a when his beliefs are $\beta(\cdot|a,q)$, and

(b) for every t, t', $\pi_S(t) \geq \sum_z \rho(z|a,q) u_S(t,a,z)$ with equality if $\beta(t|a,q)/\beta(t'|a,q) > q(t)/q(t')$.

There are two cases to consider, namely $\beta(\cdot|a,q) = q$ and $\beta(\cdot|a,q) \neq q$. If $\beta(\cdot|a,q) = q$, then let $\lambda := 1$, let γ be any element of $\Delta(T)$, and let $r := \rho(\cdot|a,q)$. Then (i) holds by (a), because $\beta(\cdot|a,q) = q = \lambda q + (1-\lambda)\gamma$, and (ii) holds by (b) and because $\lambda = 1$.

If $\beta(\cdot|a,q) \neq q$ then there exists (see footnote 12) $\lambda < 1$ and $\gamma \in \Delta(T)$ such that $\beta(\cdot|a,q) = \lambda q + (1-\lambda)\gamma$ and $\gamma(t) > 0$ implies that $\beta(t|a,q)/\beta(t'|a,q) > q(t)/q(t')$ for some t'. Hence, (ii) holds by (b) and, letting $r := \rho(\cdot|a,q)$, (i) holds by (a). ■

Lemma 8.4 *If an NLE type-action distribution (or payoff) is a Nash equilibrium type-action distribution (or payoff) of* $\Gamma(M)$ *and* $|M| \leq |T|$, *then it is a coarse NLE type-action distribution (or payoff).*

Proof. We give the proof for type-action distributions only. The proof for payoffs is similar. Suppose that $(\tilde{\sigma}, \tilde{\rho}, \tilde{\beta})$ is an NLE for Γ_0, that $(\hat{\sigma}, \hat{\rho})$ is Nash equilibrium of $\Gamma(M)$ where $|M| \leq |T|$, and that ν is the common type-action distribution of both equilibria. We will construct a coarse natural language equilibrium $(\sigma^*, \rho^*, \beta^*)$ with type-action distribution ν.

Associate each message $m \in M$ with a distinct message $t_m \in T$. For each a, m, and t, define the strategy σ^* for S in $\Gamma(T)$ so that $\sigma^*(a, t_m|t) := \hat{\sigma}(a, m|t)$. This pins down σ^*.

For each $t \in T$, let q_t be the element of $\Delta(T)$ that puts probability one on t. Define the strategy ρ^* and the beliefs β^* for R in $\Gamma(T)$ as follows. For every action-message (a, m) that is on-path for $\hat{\sigma}$, define $\rho^*(\cdot|a, t_m) := \hat{\rho}(\cdot|a, m)$ and define $\beta^*(\cdot|a, t_m)$ to be the Bayes' posterior over T from $\hat{\sigma}$ conditional on (a, m). This defines $\rho^*(\cdot|a, t)$ and $\beta^*(\cdot|a, t)$ for every (a, t) that is on-path for σ^*. For every (a, t) that is off-path for σ^*, define $\rho^*(\cdot|a, t) := \tilde{\rho}(\cdot|a, q_t)$ and define $\beta^*(\cdot|a, t) := \tilde{\beta}(\cdot|a, q_t)$.

Using the fact that $(\hat{\sigma}, \hat{\rho})$ is a Nash equilibrium of $\Gamma(T)$ with type-action distribution ν and that $(\tilde{\sigma}, \tilde{\rho}, \tilde{\beta})$ is an NLE with type-action distribution ν, it is easy to verify that $(\sigma^*, \rho^*, \beta^*)$ is a coarse NLE with type-action distribution ν. ■

Proof of Theorem 6.2. It suffices to show that there is a sequential equilibrium outcome of the base game that is a coarse NLE type-action distribution. By Theorem 4.7 there is a sequential equilibrium outcome ν of the base game that is an NLE type-action distribution. In particular ν is an NLE type-action distribution as well as a Nash equilibrium type-action distribution of $\Gamma(\{m_0\})$ for any cheap-talk message m_0. Hence, by Lemma 8.4, ν is a coarse NLE type-action distribution. In particular, a coarse NLE exists. ■

References

Banks, J. S., and J. Sobel (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 55, 647-661.

Bernheim, B. D. (1984): “Rationalizable Strategic Behavior,” *Econometrica*, 52, 1007-1028.

Blume, A. (1994): Equilibrium Refinements in Sender-Receiver Games, *Journal of Economic Theory*, 64, 66-77.

Blume, A. (2023): “Meaning in Communication Games,” working paper, Department of Economics, University of Arizona.

Blume, A., and J. Sobel: (1995): “Communication-Proof Equilibria in Cheap-Talk Games,” *Journal of Economic Theory*, 65, 359-382.

Chen, Y. (2011): “Perturbed Communication Games with Honest Senders and Naive Receivers,” *Journal of Economic Theory*, 146, 401-424.

Chen, Y., N. Kartik, and J. Sobel (2008): “Selecting Cheap-Talk Equilibria,” *Econometrica*, 76, 117–136.

Cho, I.-K., D. M. Kreps (1978): “Signaling Games and Stable Equilibria,” *The Quarterly Journal of Economics*, 102, 179-221.

Clark (2024): “Robust Neologism Proofness and the Flexible Speech Intuitive Criterion,” working paper, Department of Economics, MIT.

Clark, D., and D. Fudenberg (2021): “Justified Communication Equilibrium,” *American Economic Review*, 111, 3004–3034.

Crawford, V. P.: (2003): “Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions,” *American Economic Review*, 93, 133-149.

Crawford, V. P., and J. Sobel (1982): “Strategic Information Transmission,” *Econometrica*, 50, 1431-1451.

Demichelis, S., and J. Weibull (2008): "Language, Meaning, and Games: A Model of Communication, Coordination, and Evolution," *American Economic Review*, 98, 1292–1311.

Farrell, J. (1985): "Credible Neologisms in Games of Communication," MIT working paper 386.

Farrell, J. (1993): "Meaning and Credibility in Cheap-Talk Games," *Games and Economic Behavior*, 5, 514-531.

Gordon, S., N. Kartik, M. Pei-yu Lo, W. Olszewski, and J. Sobel (2023): "Effective Communication in Cheap-Talk Games," working paper, LEDA, Université Paris-Dauphine.

Green, J., and N. Stokey (2007): "A Two-Person Game of Information Transmission," *Journal of Economic Theory*, 135, 90-104.

Grossman, S. J., and M. Perry (1986): "Perfect Sequential Equilibrium," *Journal of Economic Theory*, 39, 97-119.

Hart, S., I. Kremer, and M. Perry (2017): "Evidence Games: Truth and Commitment," *American Economic Review*, 107, 690–713.

Heumann, T. (2020): "On the Cardinality of the Message Space in Sender–Receiver Games," *Journal of Mathematical Economics*, 90, 109-118.

Kartik, N. (2009): "Strategic Communication with Lying Costs," *The Review of Economic Studies*, 76, 1359-1395.

Kartik, N., M. Ottaviani, F. Squintani (2007): "Credulity, Lies, and Costly Talk," *Journal of Economic Theory*, 134, 93-116.

Koessler, F., M. Laclau, and T. Tomala (2024): "A Belief-Based Approach to Signaling." HAL working paper halshs-04455227.

Kohlberg, E., and J.-F. Mertens (1986): “On the Strategic Stability of Equilibria,” *Econometrica*, 54, 1003-1037.

Mailath, G. J, M. Okuno-Fujiwara, and A. Postlewaite (1993): “Belief-based refinements in Signalling Games,” *Journal of Economic Theory*, 60, 241–276.

Matthews, S. A., M. Okuno-Fujiwara, and A. Postlewaite (1991): “Refining Cheap-Talk Equilibria,” *Journal of Economic Theory*, 55, 241-273.

Matthews, S. A., and A. Postlewaite (1994): “On Modelling Cheap Talk in Bayesian Games,” in *The Economics of Informational Decentralization: Complexity, Efficiency, and Stability: Essays in Honor of Stanley Reiter*, Edited by John O. Ledyard, Kluwer Academic Publishers.

Myerson, R. B. (1989): “Credible Negotiation Statements and Coherent Plans,” *Journal of Economic Theory*, 48, 264-303.

Park, I.-U. (1997): “Generic Finiteness of Equilibrium Outcome Distributions for Sender Receiver Cheap-Talk Games,” *Journal of Economic Theory*, 76, 431-448.

Pearce, D. G. (1984): “Rationalizable Strategic Behavior and the Problem of Perfection,” *Econometrica*, 52, 1029-1050.

Rabin, M. (1990): “Communication between Rational Agents,” *Journal of Economic Theory*, 51, 144-170.

Reny, P. J. (2024): “Natural Language Equilibrium I: Off-Path Conventions,” Univeristy of Chicago working paper.

Riley, J., (1979): “Informational Equilibrium,” *Econometrica*, 48, 331-59.

Sémirat, S., and F. Forges (2023): “Forward-neologism-proof equilibrium and better response dynamics,” working paper, Université Paris-Dauphine.

Spence, M. (1973): "Job Market Signaling," *The Quarterly Journal of Economics*, 87, 355-374.

Sobel, J. (2020): "Lying and Deception in Games," *Journal of Political Economy*, 128, 907-947.

Von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

Zapater, I. (1997): "Credible Proposals in Communication Games," *Journal of Economic Theory*, 72, 173-197.

Supplemental Appendix: Discussion

1. We have limited our language to a particular set of statements about the sender's strategy represented by the countably infinite set Q. We could instead have expanded the language to include statements q for any $q \in \Delta(T)$. None of our results would change had we done so and the set of NLE type-action distributions would also be unchanged (see Lemmas 8.1 and 8.2 and Remark **??**). However, measurability issues then arise, and defining terms such as "on-path" and "off-path" become more nuanced, etc. None of these issues is serious. They are simply technical nuisances that we are able to avoid with the route we have taken.

2. The language Q excludes many kinds of statements. For example, no statement in the language says "my type is either t_1 or t_2". Likewise, for any action $z \in Z$ of the receiver, no statement in the language says "I (the sender) think that you (the receiver) should take action z."[34] We wish to argue that adding these or any other statements would have no effect on the set of NLE type-action distributions, given the convention that we are studying. Since, as already discussed in point 1 above, the languages $\Delta(T)$ and Q lead to the same NLE type-action distributions, it suffices to argue that adding statements to the language $\Delta(T)$ does not change the set of NLE type-action distributions. So suppose that some NLE (σ, ρ, β) is in effect and consider adding a new statement m^* to the language $\Delta(T)$. Further, suppose that if R were to accept the statement m^* as true according to its literal meaning, then R's posterior would be $q^* \in \Delta(T)$. We can then extend the given NLE to the new larger language $\Delta(T) \cup \{m^*\}$ as follows. For any action a, extend S's strategy so that every sender-type places probability zero on (a, m^*), and extend R's beliefs and behavior after (a, m^*) to be exactly as they are after (a, q^*). This does not affect the equilibrium path

[34] Both kinds of statements are possible in Myerson (1991). But Myerson also assumes other conventions are in place. See discussion point 3 below.

and, because messages are payoff irrelevant, the extended assessment remains a sequential equilibrium. Moreover, this extension respects our language convention because $\beta(\cdot|a,m^*)$, since it is equal to $\beta(\cdot|a,q^*)$, has the property that if it differs from q^*—the belief that R would have if he accepted m^* as true—then R's beliefs can give higher relative probability only to types for whom (a,m^*) is a best reply against ρ. Hence, adding the new message m^* does not refine away any existing NLE type-action distribution. One can also show that adding new messages cannot create new NLE type-action distributions and so the set of NLE type-action distributions is unchanged. Consequently, we never need more than the set $\Delta(T)$ (and in fact never more than Q) to represent whatever language we might have in mind, at least for the convention that we have considered here.

3. Adding statements outside Q can make a difference when other conventions are in place, such as a convention whereby whenever the sender recommends an action that is a best reply for the receiver, the receiver cooperatively carries out that action even if there are many other best replies the receiver could play. Such a "cooperative/bargaining" convention certainly would yield additional results, but it appears to us to go beyond the fundamental connection between a statement's literal meaning and its informational content. Rather, it is closer in spirit to an equilibrium selection criterion. In any case, we do not consider such conventions here, preferring to first develop a pure theory of language in games prior to appending to it any equilibrium selection criteria.

4. If for some NLE type-action distribution, an off-path action, a say, is strictly suboptimal in every sequential equilibrium of $\Gamma(Q)$ with that type-action distribution,[35] then NLE implies that, for any q, R's beliefs after (a,q) must be q. This reflects an Occum's-razor-like effect of our language convention, namely

[35] In the context of the base game Γ_0, Cho and Kreps (1987) call such actions *equilibrium dominated.*

that, absent a rational explanation for an observed off-path action and statement, one defaults to simply accepting the statement as true according to its literal meaning. Note however, that for any such action, one could specify any beliefs and corresponding sequentially rational behavior for R and the assessment would remain a sequential equilibrium. Hence, the particular beliefs that result from accepting the statement as true have no effect on the equilibrium path.

5. We have assumed that all statements in the language are cheap talk. But what if some statements are costly for the sender? That is, how would our theory change if we allowed the sender's payoff $u_S(t, a, m, z)$ to depend on $m \in M$ when the language is M? The answer is that we would continue to suppose that each statement $m \in M$ induces a unique belief $q_m \in \Delta(T)$ for R about S's type when R believes that m is true according to its literal meaning. We would also assume that every belief is induced by some message, i.e., $\{q_m : m \in M\} = \Delta(T)$. However, we would now allow that multiple messages $m, m', m'', ...$ might induce the same beliefs, i.e., $q_m = q_{m'} = q_{m''} = ...$ because different messages $m, m', m''...$ might entail different costs, i.e., $u_S(t, m, a, z)$, $u_S(t, m', a, z)$, $u_S(t, m'', a, z), ...,$ might not all be equal. Even so, the definition of natural language equilibrium—and its interpretation—would remain essentially unchanged, being now as follows: Say that (σ, ρ, β) is a *natural language equilibrium for the base game* Γ_0 iff it is a sequential equilibrium of $\Gamma(M)$ and, for any action-message (a, m) and for any pair of distinct types t and t', if $\beta(t|a, m)/\beta(t'|a, m) > q_m(t)/q_m(t')$ then (a, m) is a best-reply for t against ρ.[36] Our characterization theorems would of course no longer hold as stated and even establishing existence would require some additional assumptions because there are now potentially infinitely many payoff-relevant action-messages.

[36] Here, the base game utilities are those associated with a costless message, at least one of which would be assumed to exist.

6. Expanding on point 5, we can introduce a model in which speakers can affect the cost of making any particular statement (e.g., by shouting, becoming red-faced, perhaps even jumping up and down all the while, or by using any other form of money-burning). A simple model of this kind has the language $M := \Delta(T) \times [0, \infty)$, where the literal meaning of a message $m = (q, c)$ is, briefly, "I am using a strategy that induces the posterior q." Moreover, the message $m = (q, c)$ has utility cost c, i.e., $u_S(t, a, m, z) := u_S(t, a, z) - c$, where $u_S(t, a, z)$ is S's utility in the base game Γ_0.[37] One can then apply the definition of NLE from discussion point 5 above. A model of this kind may help to resolve the puzzling aspects of Example 5.3 due to Farrell (1993) because now type t_1 can, by increasing the cost of his message separate himself from type t_2.[38] Such a "resolution" to examples like these seems at least as convincing—without giving up on equilibrium existence—as the never-converging dynamics discussed in Farrell (1993).

7. Because we assume that S knows his own strategy, any off-path statement by S about the strategy he is using is either true or is intentionally false (in which case we have called it an "attempt to deceive"). But the assumption that S knows his strategy entails counterfactuals, and so might give one pause. When S takes his action, a, and sends a message q, he certainly knows his type t and how he randomized over his actions. However, he may not know how he would have randomized over his actions in the counterfactual world in which his type had been any of his other possible types, and he would need to know this in order to know the strategy that he is using.[39] If S does not know his

[37] See Clark (2024) for a similar model.

[38] The example is puzzling only when one assumes, as in Farrell (1993), that among the infinite number of possible cheap-talk statements, only finitely many can have positive probability in equilibrium. The puzzle disappears either when one allows all cheap-talk statements to have positive probability in equilibrium (as in the main text), or when one assumes that of the infinitely many possible statements only finitely many cost less than c, for any $c > 0$.

[39] This is so even if there is an ex-ante stage before S learns his type and, at that stage, he planned to use the strategy σ. Even then, after his type t is realized and even after he chooses his action a by

strategy, then when S makes an off-path statement about the strategy that he is using, there are no longer just the two possibilities that either his statement is true or he is attempting to deceive. There is now a third possibility, namely, that he is being truthful but is *mistaken* about the strategy that he is using, in which case the restrictions on R's beliefs loosen considerably. Interestingly, this potential criticism of NLE does not arise for coarse NLE because the only statements allowed for S in a coarse NLE are of the form "my type is t" and when S makes any such statement, either that statement is true or he is lying because he definitely knows his type—no counterfactuals are involved.

8. Continuing discussion point 7, if each type t were a different player, then player t does not know any more than R about the randomization over actions used by any other player t'. Hence, when the types are distinct players, coarse NLE would appear to be the more appropriate concept.

9. The "off-path" convention we have studied here, namely, that any off-path statement in the language should be interpreted as true unless it may be a rational attempt to deceive, yields powerful results. Nevertheless, this convention can and should be strengthened even further to ensure that some statements in the language are *always* interpreted as true, *even if they could, in principle, be seen as a rational attempt to deceive.* Take, for instance, Example 5.2. In this example, there is no conflict between S and R. They both do best when R knows S's type. Hence, for this game, the obvious convention is that each of the statements "my type is t_1" and "my type is t_2" should be interpreted as true *whenever made.* Conventions such as this might be termed "on-path conventions" because they can directly influence the equilibrium path (whereas the off-path convention considered here influences the equilibrium path only indirectly). A complete theory of language must include both on-path and off-path

following through with his plan to randomize according to $\sigma(\cdot|t)$, he may not know with certainty that he would have randomized according to $\sigma(\cdot|t')$ if his type had turned out to be t' instead of t.

conventions for language. For the game in Example 5.2, the on-path convention that the statements, "my type is t_i" $i = 1, 2$, are always interpreted as true, yields perfect coordination and no gain can ever come from using either statement deceptively. But our off-path convention allows R to have doubts about the truthfulness of these statements so long as, in equilibrium, each statement is a best reply for both types, a state of affairs that exists in the pooling equilibrium described in the main text. For the same reasons, natural language equilibrium as formulated here has no refinement power in the pure communication games of Crawford and Sobel (1982) and Green and Stokey (2007). See Blume (2023) for one route to using language to refine sequential equilibria in pure communication games.