Speech and social network dynamics in a constrained vocabulary game: design and hypotheses

Sam Tilsen*

1. Introduction

This paper describes the design of a longitudinal speech study and the general hypotheses associated with the study. Simply put, the general hypotheses are that some variation over time in how people speak is caused by their interactions with other people, that the nature of this variation is related to the social attitudes people have toward each other, and that social dynamics are the main source of instability in speech patterns. While these hypotheses may seem straightforward, and perhaps obviously true, testing them rigorously is complicated. I should confess that in testing these hypotheses I have a parallel agenda, which is to promote a research paradigm in which we attend more carefully to the physical grounding of our analytical categories. Because the hypotheses refer to variation over time, I start by considering possible patterns and causes of temporal variation:

1.1 What are the possible forms of temporal variation in speech observables?

When we observe some aspect of speech, what are the possibilities for how that observation can vary over time? The figure below shows some possibilities. The ones in the top row are the sorts of patterns we would expect from linear systems, and we can do a pretty good job of predicting the values of future observations in these cases. The ones in the bottom row are more interesting and arise from more complicated systems.

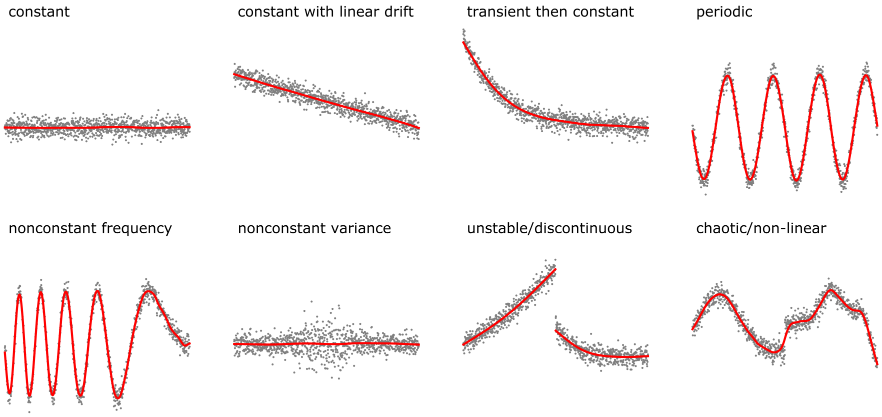


Fig. 1. Some possible patterns of variation over time in speech observables.

* **Acknowledgements**
The experiment described in this report was a team effort. Bruce McKee, Chelsea Sanker, and Stacy Dickerman were instrumental in setting up and testing the game. Additional game testers were: Miloje Despic, John Hale, Amin Nikbin, Todd Snider, and participants in the social network map game seminar: Dan Burgdorf, Amui Chong, Sarah D'Antonio, Naomi Enzinna, Andrea Hummel, Jixing Li, Yanyu Long, and Hao Yi, all of whom served as assistant experimenters and contributed to the ASR procedures by manually labelling data for HMM training.

Which patterns, or which combinations of patterns, occur in which observations of which speech behaviors? The answer should depend on what systems are responsible for the behaviors we observe, but it may also depend on how we make our observations, or how we transform them for analytical purposes. So, to even begin to address the question, we need to consider what we mean by “speech system” and what our observations represent. These considerations are important because the speech and social network dynamics experiment described in this paper is an attempt to better understand the dynamics of speech systems. Since our understanding is necessarily shaped by our observations, we should be aware of how our analytic decisions influence our results.

1.2 Physical grounding of speech systems

The word “system” can be used in many ways, but here I emphasize a physical interpretation. The systems whose dynamics we want to understand in speech are large groups of neurons that serve a common function, or a combination of such groups. Assemblies of neurons have been conceptualized by many as fundamental units of behavior (cf. Hebb, 1949; Kelso, 2009; Nicolelis, 2009).

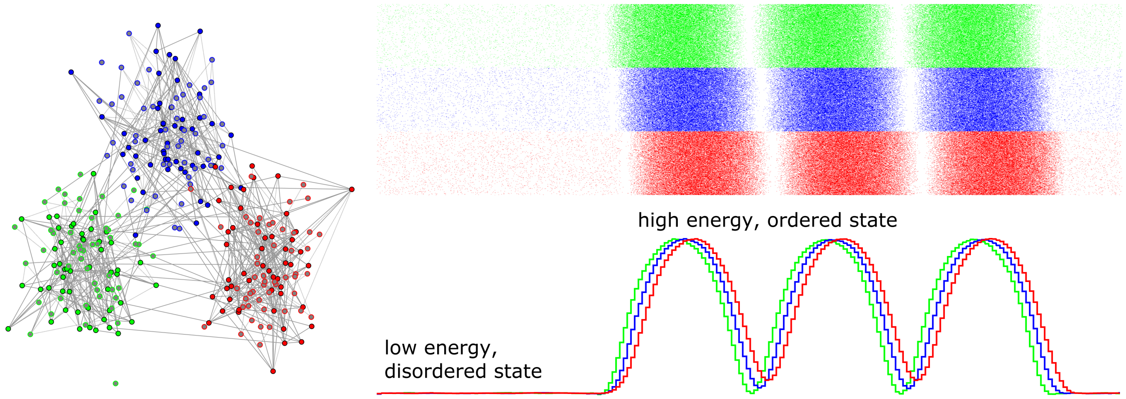


Fig. 2. A physical interpretation of systems as neural assemblies or sets of neural assemblies. These assemblies have a low energy, high disorder state, and a high energy, ordered state.

A key thing about neural assemblies is that they use energy to concentrate order/generate information. These systems are embedded in an energy substrate which provides them with an energy source. A neural assembly has two phases, one in which it draws relatively little energy from its surroundings and its components are uncorrelated, the other in which it draws relatively more energy from its surroundings and its components exhibit highly correlated, collective oscillation. The collective oscillation of the neurons in an assembly results in a drastic decrease in the degrees of freedom of the system, and if we define the system appropriately, we can see that order is created/entropy is reduced in the system. An important choice we always make is where to draw the line between the system we are interested in and its surroundings. Our analysis of a system always depends on how we have conceptualized its boundaries and its interactions with other systems.

1.3 Speech observables are indirect outputs of system dynamics

What are we observing when we study speech? Our observations are usually measures of physical changes that result from energy transfers. For example, we use a microphone to measure fluctuations in acoustic pressure, which are mechanical waves. The energy to produce those waves comes from energy stored in our bodies. Our metabolic systems transform energy stored in chemical bonds into mechanical energy that stretches muscles to increase lung volume; as those muscles are relaxed and lung volume decreases, the potential energy of the resulting pressure gradient is converted to kinetic energy of airflow. Some of this energy becomes acoustic energy. Along the way, something else very important is happening: order/information is being created. We move vocal organs (also by converting chemical to mechanical energy) to adjust the geometry of our vocal tracts. These adjustments of vocal tract geometry influence how acoustic energy is distributed across the frequency spectrum. The result of this influence is always a local concentration of order/reduction of entropy. The most uniform, disordered distribution of acoustic energy is white noise, i.e. the same amount of energy at all frequencies. The spectral distributions of speech sounds are more ordered than white noise:

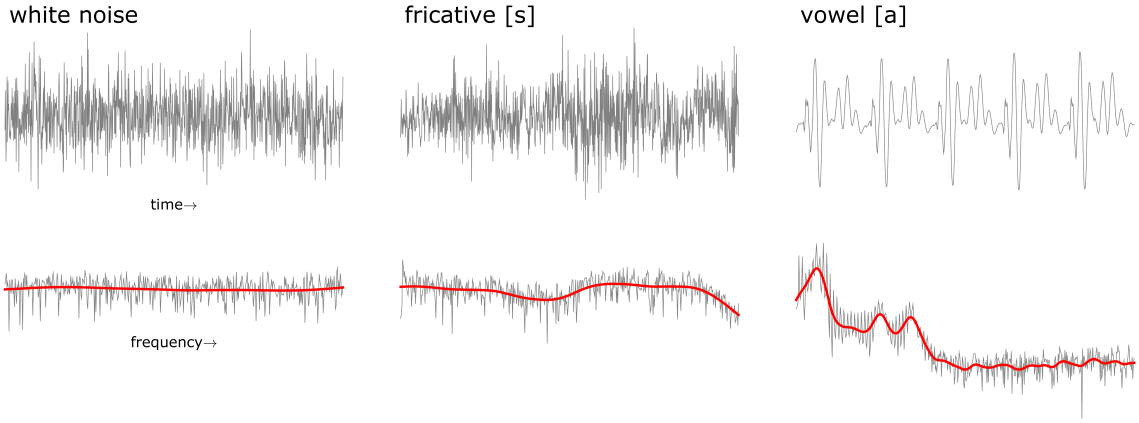


Fig. 3. Speech waveforms and spectral order. The distribution of energy is less uniform in the fricative [s] compared to white noise, and even less uniform in the vowel [a].

Information is a measure of reduction of uncertainty/disorder/entropy (Shannon, 1948). Our alterations of the spectrum of acoustic energy in speech reduce uncertainty and therefore result in a gain of information/order. This evokes a thermodynamic conception of life as concentrating order (Schrödinger, 1944) and speech systems as dissipative structures (Kondepudi & Prigogine, 1998; Nicolis & Prigogine, 1977) which use energy to create order. The creation of order in speech is not just in the acoustic signal. Neural assemblies in the perceptual systems of listeners enter ordered phases when listening to speech, and this can have consequences that result in further creation of order, through behaviors (speech or other) which reduce entropy. Of course, the universe always experiences a net increase of disorder (e.g. heat energy from the movements we make), but when we define the system and surroundings appropriately, order is increased in the relevant part of the universe. The order creation in acoustic energy of speech is caused, indirectly through a chain of processes, by order creation in neural assemblies of speakers. Hence speech observables can be understood as the indirect consequences of order creation in neural assemblies.

1.4 The importance of observation scale

When we observe linguistic behaviors we decide which scales of space and time to observe them on. This decision is often made without much deliberation, and is often determined by logistical constraints on our procedures for making observations. From a physical perspective, we should give careful attention to the spatial and temporal scales of our observations. Studying how our observations change as a function of observation scale can deepen our understanding of speech systems. Furthermore, the procedures we use to make and interpret observations are always structured by conceptual metaphors. These metaphors construct categories which allow us to reason about and make predictions from our observations; the categories themselves can be understood as simplifications, projections from higher-dimensional spaces to lower-dimensional ones. Crucially, our choices of observation scale influence and are influenced by our metaphors. Hence we should always remain aware of (1) how our observations are situated in physical space and time and (2) the analytic categories which enable our interpretations.

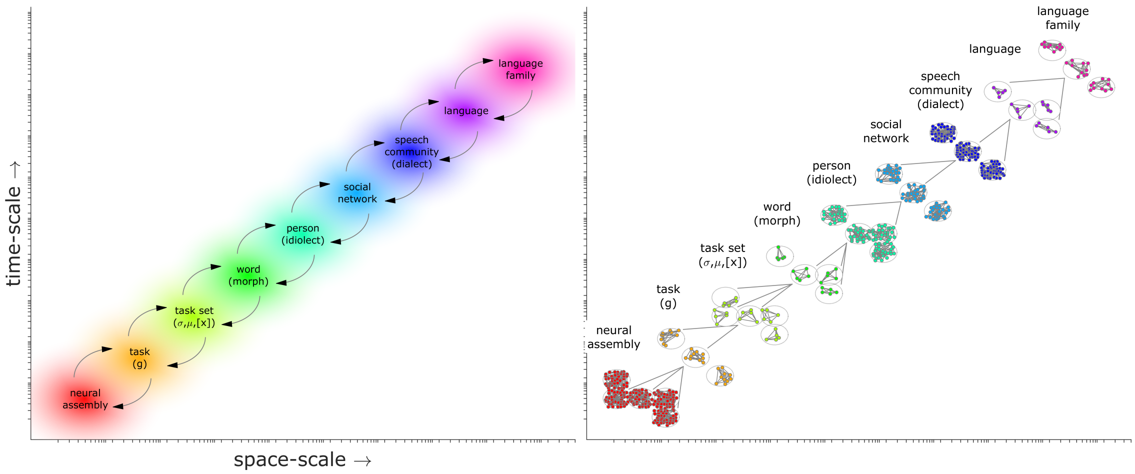


Fig. 4. Analytical categories for describing patterns in a hierarchy of spatial and temporal scales. Left: categories are associated with regions of varying scale in space and time. Right: categories as physical networks in a hierarchy of embedded networks.

Our analytical categories for studying speech can be contextualized in a hierarchy of scales or embedded networks, as illustrated in Fig. 4. Although energy and order concentration may vary with scale, no scale is more important than any other. Each scale is associated with one or more analytical categories, and these categories are associated with systems that tend to interact strongly. The partitioning of scales and labeling of categories is not so important here. The important thing is to keep in mind what each category refers to *in a physical sense*. Here is a list of some analytical categories:

- ***Neural assemblies***. Speech is grounded physically in the collective oscillatory depolarizations of large but localized populations of neurons (termed assemblies, ensembles, groups; Hebb, 1949; Kelso, 2009). Neurons in an assembly extract energy from their surroundings (the metabolic substrate of the brain) to concentrate order, and this process is modulated by interactions with other assemblies.
- ***Tasks***. Tasks are systems of motor and sensory neural assemblies which emerge in speech development because of nonlinearities in interactions between sensory and motor systems (i.e. asymmetries associated with the embodiment of cognition). The concept of a task was developed in the theory of task dynamics (Kelso, Saltzman, & Tuller, 1986), and gestures—the fundamental units of speech in articulatory phonology (Browman & Goldstein, 1986)—are tasks. The mathematical description of gestures implicitly presupposes perfect, instantaneous interactions between sensory and motor assemblies, which is an idealization of coordinative control using forward models. Because tasks involve activation of and feedback between both motor and sensory assemblies, tasks are associated with larger scales of space and time than individual assemblies.
- ***Task sets***. Phonological categories such as segments [*x*], moras μ, and syllables σ are sets of tasks. These sets appear to emerge in speech development from the internalization of feedback control (Tilsen, 2014, 2016), but their role in adult speech is unclear.

- ***Words (morphs).*** Words/morphs are combinations of sensory and motor assemblies associated with task sets and with assemblies which represent concepts or meanings through interactions with various cognitive subsystems. Our analyses of speech usually conceptualize words as abstract objects (writing/literacy perpetuate this), and as independent of a particular physical instantiation. But it should always be remembered that "words" only exist because there are physical systems, i.e. task sets and concept-related neural assemblies, which create ordered states that we associate with words.
- ***People (idiolects)***. Patterns of order creation in assemblies occur in individuals. The current experiment is particularly concerned with how these patterns change over time, and how our analysis of the patterns may change as a function of the analysis scale. The scale of idiolects, or individual speech, is a useful scale of analysis because there are clear physical boundaries between neural assemblies associated with one person and another.
- ***Social networks***. Our interactions with other people often occur in small groups which are determined by social organizations. These groups are dynamic in size and structure. There is no clear division between social networks and speech communities. The concept of a social network is an analytical construct, grounded in our concept of an individual. When we associate speech patterns with a social network, we are constructing categories which generalize over physical states of neural assemblies in a set of individuals.
- ***Speech communities (dialects).*** Large networks of people, often distributed in a bounded region of geographical space, are sometimes called dialects. The boundary in scale between social networks and speech communities is arbitrary. The category of dialect is used for generalizations of speech patterns that correlate with spatial or social distributions of individuals.
- ***Languages.*** Languages are sets of speech patterns which are shared by a network of people. The distinction between language and dialect is often constructed with the concept of mutual intelligibility: if two speakers can understand one another, then differences in their speech patterns are dialectal; if two speakers cannot understand one another, they speak different languages. Of course, mutual intelligibility is a matter of degree, and hence the dialect-language division is not clear cut. More to the point, in a physical interpretation, intelligibility implies that the patterns of order created by speech systems in one individual can induce similar patterns in another.
- ***Language families.*** Groups of languages—i.e. language families—are sets of speech systems which are similar on large spatial and temporal scales. The similarity is a product of the structure and dynamics of the networks of individuals that comprise them (and their neural assemblies). Analyses of language families are the most coarse-grain analyses of speech system networks that we can do. Proceeding from the relatively fine-grain scale associated with an individual, through progressively more coarse scales—social networks, speech communities, languages—we eventually reach a time-scale in which we can no longer make observations. The order created by speech systems is not accessible more than several thousand years ago, and our ability to extrapolate back in time has limits.

1.5 Analytical categories as projections

How would we observe a neural assembly, task, word, language, etc.? From a physical perspective, these categories should be associated with physical states of nervous systems. Imagine that we could observe such states in people. These states would be extremely high-dimensional and would include, for instance, electrical and chemical potentials of all cells in the body along with relevant metabolic variables (even this would be a drastic oversimplification, of course). This state space is far too high-dimensional to be of much practical use, so to conduct analyses, we must derive simplified representations of the high-dimensional state. A useful and general way of thinking about how we make these simplified representations is to view them as projections of objects from a high-dimensional space to a lower-dimensional one. The basic concept of a projection is illustrated below:

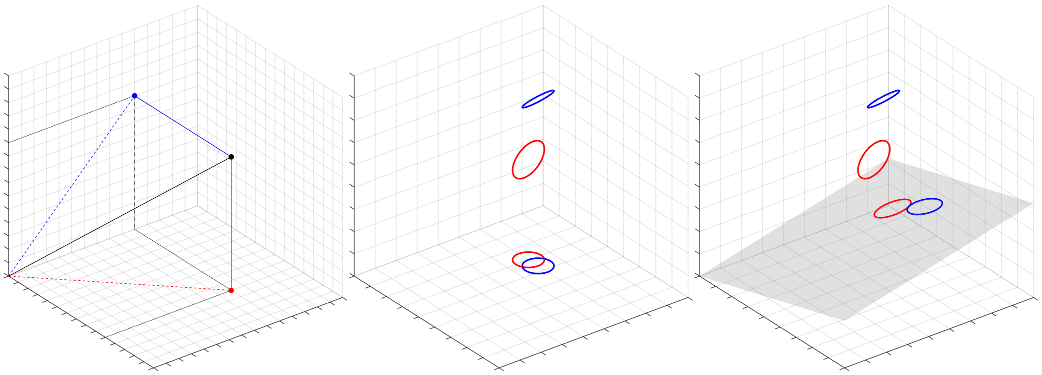


Fig. 5. Projections from higher- to lower-dimensional representation. Left: a vector in 3-dimensional space (black) is projected to 2-dimensional spaces (blue, red). Center and right: trajectories in 3-dimensional space are projected onto two different planes.

The projections we use to construct analytical categories should be motivated by the statistical distributions of our observations. Consider the category of a neural assembly. We might try to inductively identify an assembly by measuring the state of the nervous system over some brief period of time in some localized spatial area of cortex. This measurement gives us a trajectory in a high-dimensional space. But any one observation of this sort would not be enough to identify an assembly. Instead we have to collect a large ensemble of observations. If all trajectories in our ensemble are equally likely, we would have no motivation to construct a simplified representation. But if some trajectories are much more likely than others, then we might associate those with states of an analytical category, i.e. a neural assembly. Motivated in this way, the neural assembly is an analytical category derived from projecting a higher-dimensional representation to a lower-dimensional one, *when the statistics of our observations motivate it.* The projection lets us ignore a vast amount of (presumably irrelevant) variation in space and time.

One thing to note about projections is this: the surface an object is projected onto matters. As shown above, volumes may or may not overlap when projected, and trajectories may or may not intersect, depending on the surface they are projected onto. Likewise, our interpretations of analytical categories can differ according to the surfaces they are projected onto. The categories are objects that are projected, and differences in spatial and temporal scale correspond to different surfaces. To be sure, the idea that analytical categories are projections is a *metaphor*. One could pursue numerous extensions of this metaphor: null spaces, nonlinear projections, mappings, decompositions of projections, etc. That is not the point here. The point is that we should be aware that our analyses are built upon systems of categories which are simplifications of more complicated things, i.e. lower-dimensional representations of objects in higher-dimensional spaces.

The usefulness of reducing dimensionality by using analytical categories is to make it easier to think about processes on different spatial and temporal scales. The detail in high-dimensional systems is too distracting otherwise. Now we return to the analysis scale hierarchy from this perspective. Tasks can be viewed physically as state trajectories which involve various motor and sensory assemblies. In analyzing tasks, we might want to ignore some spatial and temporal variation that occurs on scale of assemblies. We accomplish this by projecting from a representation in assembly space to a new analytical space for tasks. Tasks are thus physically grounded, but somewhat less directly: we project high-dimensional objects to our lower-dimensional assembly space; then we project from assembly space to task space. The projections should be motivated by statistical patterns in the distributions of assembly states. The familiar analytical constructs of phonological theories—segments, moras, syllables, etc. (i.e. task sets)—these categories are useful to the extent that on some spatial and temporal scales they can be motivated by statistical patterns in distributions of tasks.

1.6 Analytical categories as saddle points

It is helpful to think of high-dimensional state space trajectories as guided by forces (Gibson, 1979; Kelso, 1982; Spivey, 2007). These forces can be visualized as the negative gradients of energy potentials, which associate an energy-like scalar quantity to each point in the space. The motions of state trajectories can then be understood to minimize potential energy. The state of the nervous system is continuously guided by a force field, and force field itself varies over time. Since analytical categories correspond to statistical non-uniformity in ensembles of trajectories, it makes sense that state space trajectories associated with an analytical category spend a longer period of time in some regions of state space than others. For this reason, analytical categories can be associated with saddle-point equilibria, where there are both stable directions and unstable directions relative to an equilibrium point. In the vicinity of the saddle-node, the potential gradient is relatively small, so forces on the state variables are small, and the system will spend relatively more time near the equilibrium than moving between equilibria. Examples of saddle-points in two dimensions are shown below, but the reader should keep in mind that this picture must be extended to a very high-dimensional space in understanding speech dynamics.

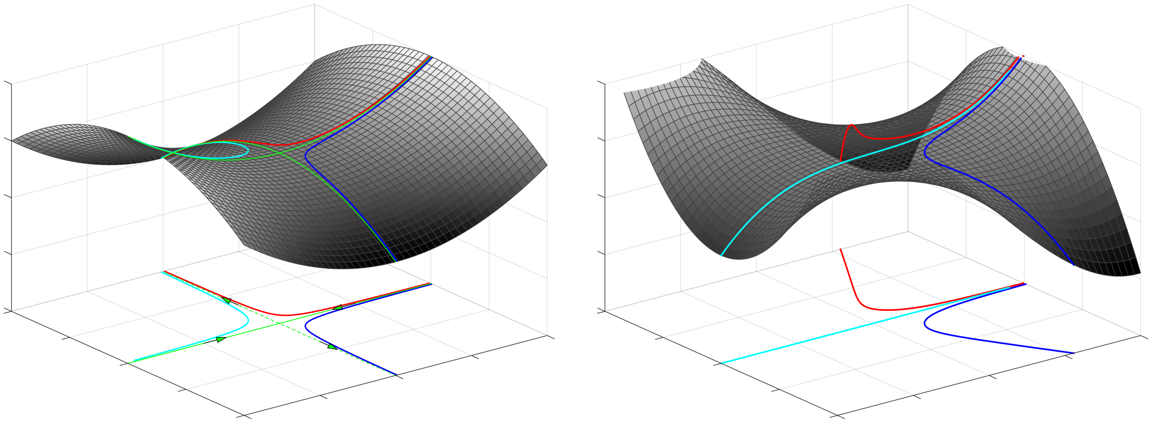


Fig. 6. Examples of saddle points and potential energy. The height of the surfaces represents the potential energy. State space trajectories move in the direction of maximal decrease in energy.

The "context" of an analytical category is also represented in a potential. In the examples shown, notice that there are multiple ways to approach the saddle-point equilibrium, and multiple ways to depart to a new configuration. Different approaches correspond to different contexts and different departures correspond to transitions to alternative analytical categories. Also, the potential landscape itself must be dynamic. The system state and potential landscape must interact in a mutual feedback loop: the system is perpetually adopting new potential surfaces which drive the evolution of the state, and the state evolution in combination with the environment in turn alter the potential. This is useful in understanding feedback interactions between analytical scales: the effects of larger-scale systems on smaller-scale ones can be viewed as forces. All of this matters for speech because if we want to understand the dynamics of speech systems, we need a coherent way of talking about how patterns on different scales interact.

1.7 Analysis across scales

Our analytical categories are constructed with projections from higher-dimensional representations to lower-dimensional representations defined on a range of spatial and temporal scales. One important consequence of this is that *we should investigate how our observations change as a function of analysis scale*. The changes that we might investigate include the sizes of fluctuations; spatial and temporal correlations between observations; changes in the entropy of observation distributions, etc.

Take vowels as an example. The category of *vowels*, any specific vowel category such as [a], more generally the category of *segments*—all of these can be derived from a series of projections from a higher-dimensional space to a much lower-dimensional one. We can analyze some observable parameter of a vowel (e.g. a formant frequency) on the spatial scale of one speaker on the timescale of one day, or on the spatial scale of multiple speakers of a speech community on the timescale of lifetimes, or of a multitude of speakers of a language over centuries (if we had acoustic recordings). In all cases, anything we might say about the observable parameter (e.g. its central tendency in different subpopulations, correlations across space, correlations across time, etc.) implicitly projects our observations to ignore dimensions of variation we are not interested in. Because there is no privileged scale, how should we decide which projections to make and which projections not to make in conducting our analyses? One approach is not to choose any particular scale, but rather to conduct an analysis across all scales and characterize how our estimated quantities vary with scale.

Imagine that in the context of an experiment we made repeated observations of an acoustic parameter associated with the production of a vowel. The observations were made in two different rooms, i.e. two different spatial positions, as shown in the figure below (top, left). Since we are not interested in this spatial dimension (we believe it is irrelevant), we project the observations to subspace which ignores the dimension of variation associated with room (top, middle). Furthermore, the observations were made on different days, but within each day the observations were made in association with a sequence of experimental rounds (e.g. top, right). If we have reason to suspect that the important dynamics occur from round to round, rather than from day to day, we can project to a new time scale which ignores time between days and emphasizes time in rounds (bottom). Implicitly we have assumed that our analytical category of vowel still makes sense and means the same thing regardless of how we project the observations.

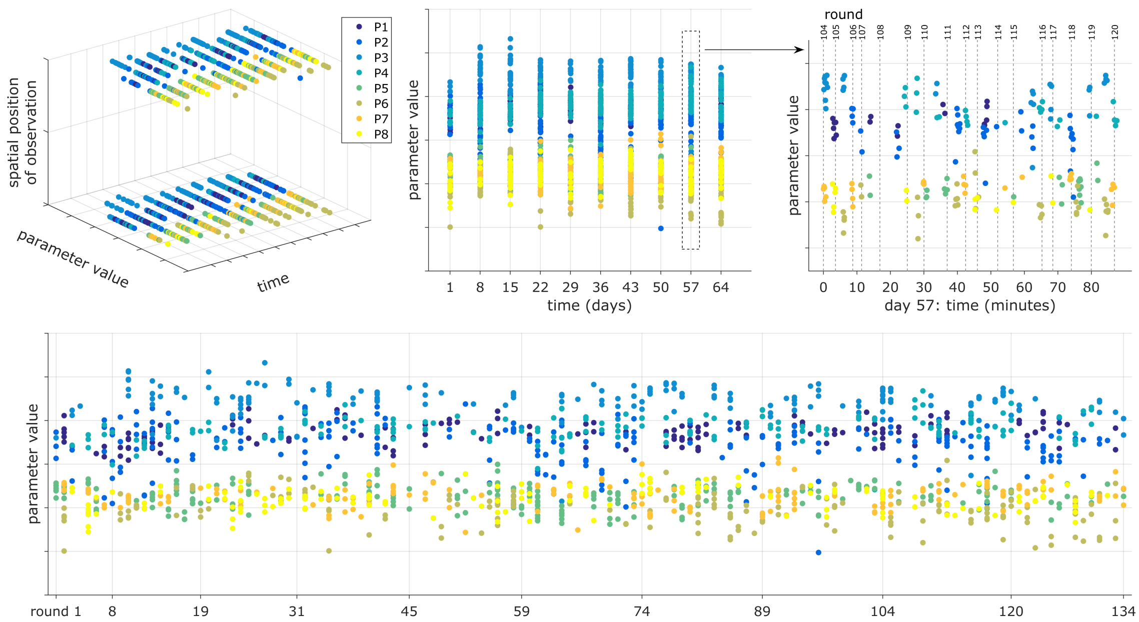


Fig. 7. Projections implicit in a hypothetical analysis of a vowel system. Top left: observations made in different positions in space on different days. Top middle: projection over spatial position of observation. Bottom: projection from days to experimentally relevant time-scale (rounds), which expresses time in rounds (top right) and ignores time between days.

Note that there are also projections we do not implement for our analysis: since we are interested in the spatial variation associated with individuals, variation associated with speakers remains a dimension in our analysis (encoded by colors in the figure). And, note that there are projections we may have assumed without much conscious deliberation: we have represented the observed parameter without consideration of speaker gender, in effect projecting over dimensions of gender-related variation.

The range of time-scales we can analyze depends on the frequency of our observations and the period of time over which we make observations, much like the frequency range of a spectral analysis depends on the sampling period and analysis window. The figure below shows how vowel parameters vary over a range of timescales for two speakers. The parameters were estimated by averaging observations over window sizes (i.e. analysis scales) ranging from 1 round to 134 rounds. (To be more precise, a Gaussian window was used to calculate a weighted average). As we increase the size of the windows, we project over more and more temporal variation that occurs within the windows.

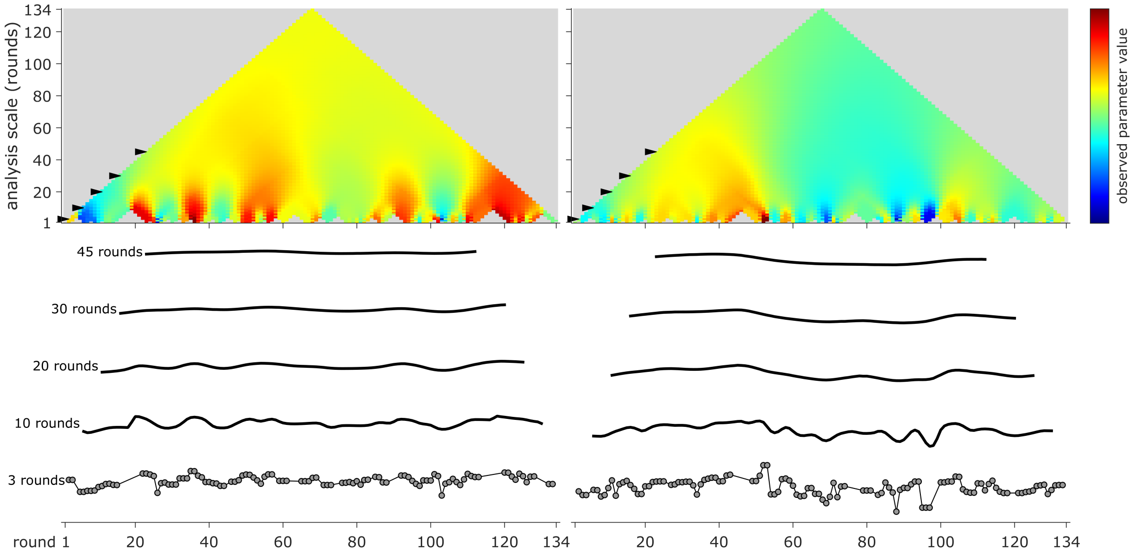


Fig. 8. Fluctuations in vowel parameters for two speakers, measured on a range of time scales. The time series show vowel parameter dynamics on selected scales, which are indicated by arrows in the top panels.

The scaling analysis is useful not only because we can calculate various quantities at each scale (means, variances, correlations, etc.), but because we can study how these quantities change as a function of scale. A scaling analysis of this sort is presented later on in this paper (Section 3). The usefulness of scaling analyses depends on collecting data frequently enough and for a long enough time. Logistically this can be expensive and time-consuming; hence we rarely collect data suitable for analyses of this sort.

1.8 Speech systems: stable or unstable?

Why do linguistic behaviors inevitably change over time? Recall the possible forms of temporal variation for speech observables that were shown at the start of this paper. These possibilities can be viewed as predictions, derived from hypotheses connected with models of speech systems. In general the patterns of temporal variation in our observations are combinations of some of those possibilities (and other ones we failed to include), because in general our observations result from interactions of systems. As analysts, our job is to decompose those combinations, in order to better understand the systems which give rise to our observations. Here are several general hypotheses about the underlying nature of those systems:

1. *No stable systems.* Linguistic behaviors change over time because no linguistic systems are stable. All systems exhibit inherently non-stationary behavior, sensitive dependence on initial conditions, internal instabilities, etc. Under this hypothesis, if we observe stability it is only a consequence of our analysis scale: on a sufficiently large or small scale, all speech systems are unstable.

2. *Stable systems in unstable surroundings*. Linguistic systems tend to stabilize over time, but interactions between systems typically obscure stability. Under this hypothesis, there are some systems which, in the absence of interactions with other systems, will evolve toward a stable equilibrium. Departures from stationarity are often a consequence of larger scale systems exerting strong, sometimes catastrophic effects on the dynamics of smaller scale ones. This hypothesis predicts that if we can separate the effects of external systems on a given system we are interested in, then we will observe stationary distributions of observables associated with the given system. In other words, if define the system-surroundings boundary in the right way, then we can identify

stability by removing the effects of the surroundings from the system. Example of this are shown below, where external factors responsible for non-stationarity of observations are factored out, revealing underlying stability.

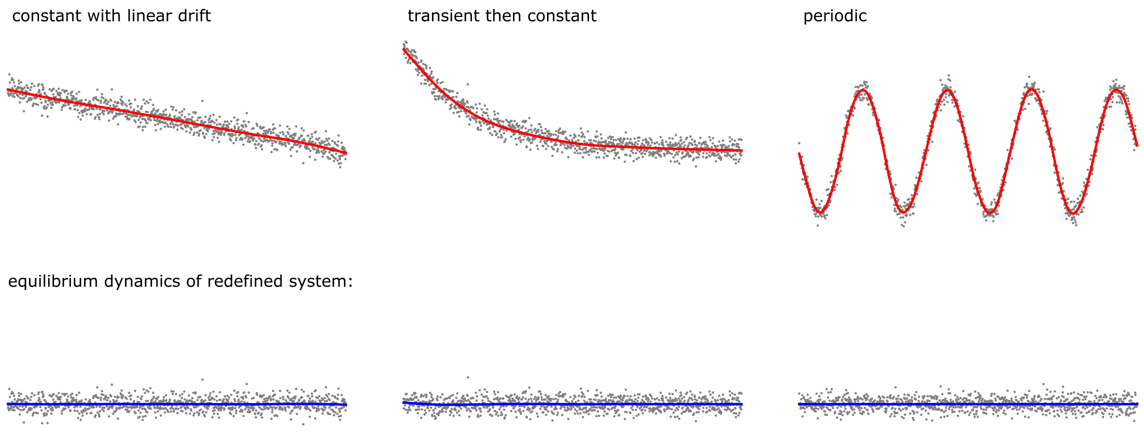


Fig. 9. Stability revealed after appropriate definition of system and surroundings.

3. *Metastable systems*. This is not really an alternative to the *stable systems in unstable surroundings* hypothesis, but an elaboration. If there are indeed some stable speech systems, which interact with other stable systems, then the interactions can prevent some systems from reaching their global equilibria. This has the effect of creating multiple local stable equilibria—pockets of stability in a high-dimensional space. Observations can be stationary under these circumstances, but their values do not reflect a global energy minimum. Interactions with the surroundings may cause transitions from one metastable state to another, and may push some observables away from their equilibria while allowing others to approach equilibria.

1.9 Covariation of social and linguistic observables

People experience both conscious and unconscious changes in cognitive states as a result of interacting with other people. I use the term “social” to describe in a physical sense systems which instantiate these state changes. Speech always occurs in a “social” context because the neural assemblies which constitute our awareness of social context interact with assemblies which are more directly related to speech. If we have some way of measuring indirect observables associated with social systems, then we can analyze whether changes in social systems correlate with changes in linguistic observables. This sort of interaction is highly relevant to question of why linguistic behaviors inevitably change over time: social systems are good candidates for external systems whose interactions with speech systems obscure stability. In relation to the alternatives mentioned above, the following predictions can be made:

1. *No stable systems.* Prediction: social system dynamics may explain some variation in speech observables, but speech observables will be non-stationary even when social system effects are factored out.

2. *Stable systems, unstable surroundings*. Prediction: when the effects of social system dynamics on speech observables are factored out, some speech systems will exhibit stationary dynamics, indicative of equilibrium behavior.

3. *Metastable systems*. Prediction: relatively abrupt changes in speech observables will correlate with abrupt changes in social system dynamics.

Although formulated quite generally, and with ambiguity in how to factor out social dynamics from non-social dynamics, these hypotheses are a useful starting point for working toward a deeper understanding of speech. In all cases consideration of scale is important: stability on one scale may belie instability on another, and vice versa. The hypotheses leave open the question of which observables are appropriate to associate with which speech systems, and indeed the experimental data offer a useful testing ground for numerous linking hypotheses between system models and speech observables.

2. Method

A map game (Anderson et al., 1991; Pardo, 2006) and a framework for organizing gameplay were developed to test the above hypotheses. The main design goal was to obtain sufficient statistical power for hypothesis testing, while minimizing the extent to which experimenter decisions might bias behavioral dynamics. In other words, the aims were to make the experiment as simple as possible, and to collect enough data to test the hypotheses. Another design goal was to make the game engaging and sufficiently challenging but not too hard, since participant drop-out over the 10 weeks of the study was a concern. The complexity of the experimental procedures, logistical considerations, and technological limitations constrained many design decisions. The participants were 4 male gender and 4 female gender undergraduates, self-identified as native speakers of English with no speech or hearing disorders, ages 18-20.

2.1 Gameplay

In each game there are two players seated facing one another in front of laptops. Each player sees one of two almost identical maps. The maps have many locations labeled by name and differing by size, color, shape, and fill (see Fig. 10). One player is the *giver*, the other is the *receiver*. The giver has a 20-location route drawn on their map, but the receiver does not have the route on their map. The goal of the game is for the receiver to draw the route on their map by clicking on the route locations in the correct order. Within each team there is an information asymmetry: the giver has most of the information needed to finish the game. The starting location is known by both players and so the giver must communicate 19 locations to the receiver. The receiver and giver may speak to one another, but in this experiment they were allowed to say only (1) the names of locations on the map and (2) a small set of adjectives and function words (see *game lexicon* below). An experimenter is present in the room and records violations of the allowed words.

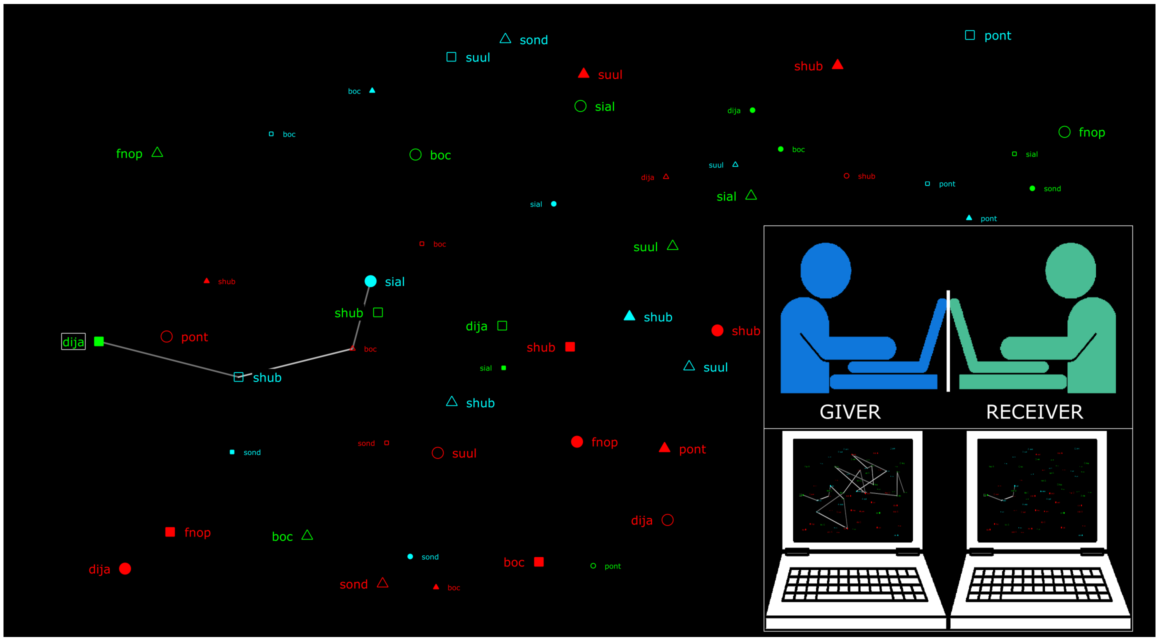


Fig. 10. Illustration of gameplay. The giver and receiver have identical maps. The giver has a route on their map, and the goal of the game is for the the giver to communicate the route to the receiver. The receiver clicks on location to draw the route on their map.

When the receiver clicks on the correct next location on the route, a line from the preceding location is drawn on their map. The receiver typically acknowledges that this has happened by saying “okay”, and the giver then communicates the next location. The process iterates until all 19 segments of the route have been drawn by the receiver. For a route segment to be drawn, the receiver must click on the correct location symbol, not the name of the location. If the receiver clicks anywhere other than on the correct location symbol, a penalty message appears in the upper left of the screen for 5 seconds. During this 5 second period no additional incorrect click penalties are registered.

Players were seated approximately four feet apart, facing one another, in front of identical laptops. They wore headsets with noise cancelling unidirectional microphones. On the table to the left of each laptop was a list of the allowed words. To the right of each laptop was a mouse pad and a specially-designed silent-button mouse, which prevented mouse clicks from being picked up by the microphones. A cardboard screen was positioned between the laptops. The screen was about 0.5 m high and prevented players from seeing the hand movements of their teammates, while still allowing face-to-face visual contact. MATLAB was used to control all aspects of the games, including drawing the maps, recording audio (22050 Hz), recording mouse clicks and mouse position (≈50 Hz), providing feedback on incorrect clicks, and updating the receiver map with each correct click. MATLAB was also used to administer pre- and post-game surveys, to elicit player rankings, and to conduct a recording level test before each game. The two-laptop setup for each game was replicated in two rooms in the Cornell Phonetics Lab. Laptops communicated with each other and with a control PC by reading and writing to a server. One laptop in each room was always used by the giver and the other by the receiver. Substantial efforts were made to ensure that the setups in the two rooms were as identical as possible.

2.2 Experiment and session structure

There were 10 sessions of the experiment, each of which lasted about 90 minutes. The sessions began at the same time on Wednesday evenings, for 10 consecutive weeks. Typically 14-15 rounds were played in each session (see Fig. 11), but in the first 3 sessions fewer rounds were played because participants took longer to finish games. Also, 15 minutes of the first session were devoted to giving instructions. A total of 134 rounds (535 games) were played over the course of the experiment. In all but one round there were 4 games (8 players, 2 players on each team); one game in round 8 was not played because a participant arrived late.

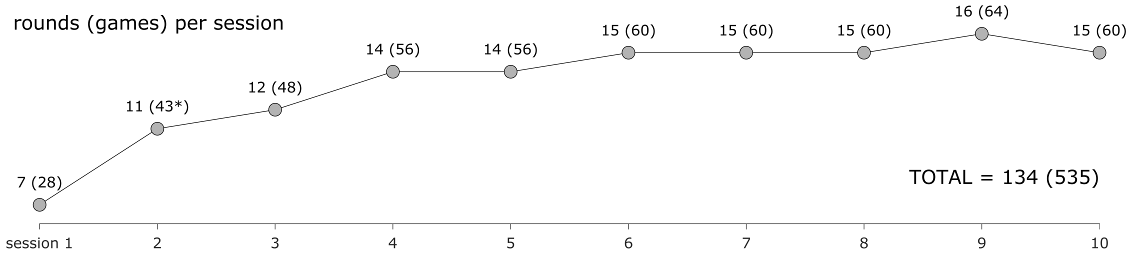
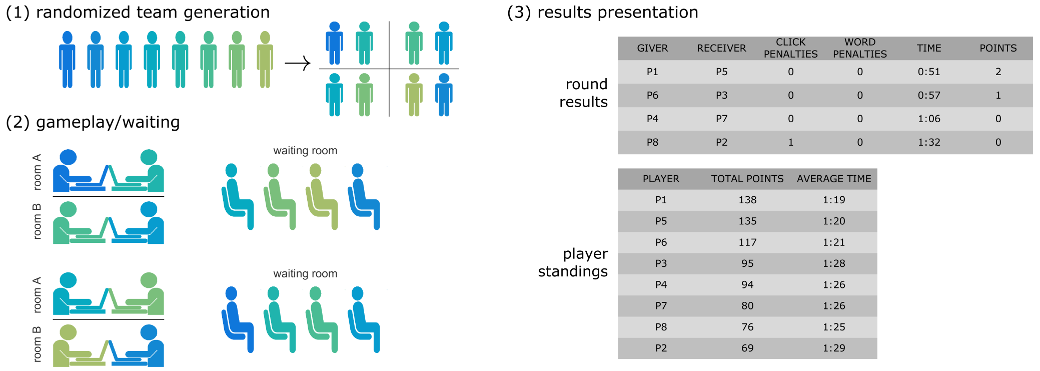


Fig. 11. Rounds (and games) played per session.

The overall structure of each session consisted of iterated rounds, each of which involved three phases: *team generation*; *gameplay/waiting*; and *results presentation* (see Fig. 12). At the start of each session, participants put on nametags on which their names had been printed. Each round began with the random generation of teams, random assignment of two teams to each of the two rooms, and random ordering of teams to determine which two teams play first. This procedure was implemented in MATLAB on an experiment control PC, which was located in a lounge between the two rooms in which games were played. The team generation algorithm (discussed in more detail below) was random but biased by rankings that each participant provided after every game.



GIVER	RECEIVER	CLICK PENALTIES	WORD PENALTIES	TIME	POINTS
P1	P5	0	0	0:51	2
P6	P3	0	0	0:57	1
P4	P7	0	0	1:06	0
P8	P2	1	0	1:32	0

PLAYER	TOTAL POINTS	AVERAGE TIME
P1	138	1:19
P5	135	1:20
P6	117	1:21
P3	95	1:28
P4	94	1:26
P7	80	1:26
P8	76	1:25
P2	69	1:29

Fig. 12. The three phases of each round: team generation, gameplay/waiting, and results presentation.

After team generation, the experimenter verbally instructed the players on the first set of teams using the following phrase (with minor variations): "[*name-giver*] and [*name-receiver*] in [*room*]". The name of the giver was always announced first, and the team assigned to room A was announced before the team assigned to room B. Players were able to see their assignments on the control-PC screen as well. Players entered their assigned rooms, along with the assistant experimenters. The assistant experimenters varied from week to week; all were graduate students in the map game social network speech dynamics experiment seminar. The assistant experimenters ensured that the doors to their respective rooms were closed, to prevent the other participants from overhearing the games.

When one of the first two teams finished the game, that team went to the waiting room and the experimenter directed the appropriate team from the waiting room using the phrase above. In each round participants spent about half of their time in the waiting room. In the first 3 sessions of the study there was substantial variation in game completion times; thus durations of waiting times were variable. In later sessions game completion times were less variable; both of the teams in the first set of games would typically finish within about 10-30 seconds of one another, and so both of the teams in the second set of games would also finish around the same time. By the fourth session, the average raw game time was around 75 seconds (see Fig. 13). For every game, players completed a 4-question pre-game survey, a 4-question post-game survey, and ranked the other players (survey and ranking procedures are described below). The surveys and rankings added about 20 s to the time between entering and exiting the gameplay rooms, and another 15-30 s were needed for participants to transition from the game rooms to the waiting rooms.

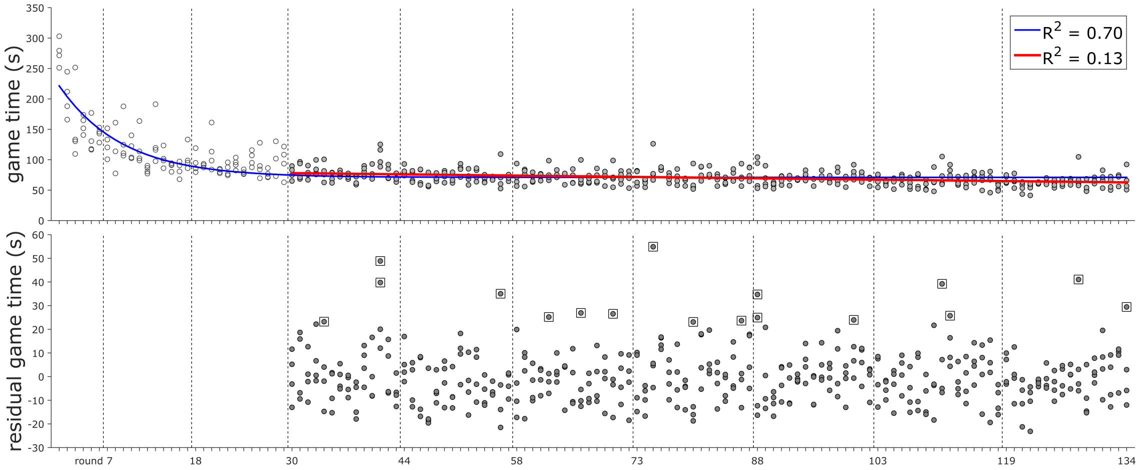


Fig. 13. Game completion times. Top: exponential fit of game completion times across the experiment (blue) and linear fit of game times from sessions 4-10. Bottom: residual game times from linear regression; atypically long games are show with squares.

After the last game in a round was completed, all participants were told to come out of the waiting room to the lounge, where they would view the presentation of results on the control-PC. The experimenter waited until all participants were in the lounge and then queried the assistant experimenters for word violations that had occurred in either set of games. For each observed violation, the assistant experimenters announced the disallowed word/phrase and the name of the player who produced it; then the experimenter added the penalties through the control-PC graphical interface (penalty procedures are discussed in more detail below).

Next the experimenter announced that they would finalize the results and did so using the control-PC interface. Game times from the previous round and points awarded then appeared on the screen (see Fig. 12). The experimenter read the names of the players on the teams with the two fastest times. Both of the players on the 1st place team were awarded 2 points and both of the players on the 2nd place team were awarded 1 point. If the time of 3rd place team was close to those of the 2nd place team, the experimenter read the names of those players as well. Hence the verbal presentation of results conformed to the following pattern (with some minor variations): "[*giver*] and [*receiver*] in 1st place with a time of [*time*], [*giver*] and [*receiver*] in 2nd place with a time of [*time*], and [*giver*] and [*receiver*] just [*n*] seconds behind". The experimenter tried to announce the results with some degree of enthusiasm, but avoided prosodic focus on the names of the players.

After the results had been announced, the experimenter used the control-PC interface to begin a new round, which initiated the randomized team generation and room assignment procedures. The round structure (i.e. team generation, gameplay/waiting, results presentation) was repeated for the duration of the session, and was the same across sessions.

2.3 Instructions

In the first 15 minutes of the first session, a PowerPoint slideshow was used to provide participants instructions on gameplay and other procedures. The experimenter read the text on the slides to the participants. The following gameplay instructions were given (distributed across a number of slides):

- In this study you will be playing a map game with a teammate.
- One of you will be the "directions-giver" and the other will be a "directions-receiver"
- During the game, you and your teammate have identical maps on your screens. Both maps are labeled with many locations.
- The giver has a route on their map. *[An example map screenshot was provided on this slide, see Fig. 10.]*
- The receiver has the same map, with no route. *[The example map from the preceding slide was provided with no route.]*
- The goal of the game is for the receiver to follow the route by **clicking on the correct locations** in the **correct order**.
- To accomplish this, the giver will communicate the route to the receiver.
- The location names are made-up. They are not real places.
- The names appear many times on each map…
- …but the symbols for the locations differ from each other by size, color, shape, and whether the shape is filled with color or unfilled.
- The starting location is always on the **FAR LEFT SIDE** of the map. There is a **WHITE BOX** around the starting location on **BOTH MAPS**. *[The white box was indicated with an arrow.]*
- The receiver should NOT click on the starting location.
- When the receiver clicks on the next correct location on a route, that part of the route will be drawn on their map. *[A series of 4 slides illustrating the addition of 4 consecutive route segments on a map was shown here.]*
- The receiver must click on the location symbol, NOT the location name: *[Two subparts of maps with mouse pointers on the symbol/name were shown here.]*
- Clicking on a location name, incorrect location, or empty space, will result in a 5 second penalty.

Next, more general information regarding the game/round/session structure was provided:

- You will be competing against other teams to see which team can complete the route in the shortest time.
- Each game takes about 2-5 minutes.
- The time limit for the game is 5 minutes. After 5 minutes the game will automatically stop.
- The game may seem hard at first, but most people quickly get better.
- In your first few games you may reach the 5 minute time limit.
- In each session of the 10-week study, you will play many rounds of the game.
- In each round, every player is paired with a teammate.
- If there is an odd number of players, one player sits out for the round.
- For each round, teams are created partly by random, but partly influenced by *teammate preference rankings*.
- After each game, you rank all of the other players, according to which players you most/least prefer to be on a team with.
- You also answer several questions before and after each game.
- All of the rankings and question responses are **STRICTLY CONFIDENTIAL**.
- Your rankings and responses to questions will never be shared with any other participants in the study.
- It is important to understand that the rankings influence who you will be paired with in the next round.
- However, the pairings are also partly random.
- Because of the randomness in pairings, you cannot be sure whether you were paired with someone by chance or because one or both of you ranked the other highly.
- At the end of each round, when all teams have played, their times are compared and points are awarded as follows: fastest team: 2 points for each player; 2nd-fastest team: 1 point for each player; other teams: no points
- You will be shown the results once all teams have finished.

Additional rules for participant conduct outside of games were presented next. Extra emphasis was placed on the rule that players are not allowed to talk to one another except during the games. No violations of this rule were observed.

- Between games, you will wait in a quiet room while the other players play the game.
- While you are waiting, you are not allowed to speak
- **YOU ARE NOT ALLOWED TO TALK TO ANY OTHER PLAYERS AT ANY TIME** during the experimental sessions, except while playing the

to other players.

game.

Special gameplay instructions were then presented:

- There is an important twist to this game. Both players are allowed to say only the following:
  location names labeled on the map
  words on the **allowed words list**
- A list of allowed words is on the table by each of the game stations.
- In your first few games, you may find it difficult to be restricted to the allowed words.
- The experimenter monitoring your game will clap or knock on a table three times if you say a word not on the list. Continue playing if this happens.
- Your team will be **penalized 10 seconds** each time that either player says any words that are not on the allowed words list.
- Hesitations like "uh" and "um" are okay, but please keep the following rules in mind:
  avoid making funny noises
  avoid interjections like "oh!"
  do not use profanity
  do not use non-verbal signals, such as hand movements, to communicate
  do not tap your feet or hands on the table
- When you violate the above rules:
  1st violation: warning
  2nd violation: 10 second penalty

After the special gameplay instructions, participants were shown how to wear the headset microphones. They were told not to adjust the microphones and never to press any keys on the laptops. They were also instructed as follows regarding a volume test before each game: "When the screen says 'Recording', say your name **at a normal volume** (the volume you will use during the game). If the volume test fails, try again, saying your name **again at a normal volume**. If the test fails three times in a row, the experimenter will check the audio setup."

Pilot testing showed that when first learning to play, some players got frustrated because of discoordination between giver and receiver. Hence players were advised as follows:

- Sometimes a team gets stuck on a map, either because the giver has made a mistake, or the receiver is lost.
- In these cases, you should consider using the words on the allowed word list to return to a previous location on the route.

Finally, the players were shown the following list of allowed words for 30 seconds (a copy of this list was always present beside each laptop):

Allowed words
big, large
small, little
red, green, blue
circle, square, triangle
filled, unfilled

up, down
left, right
by, near, to, from
and, or, not
yes, no
okay, wait, back,
repeat, sorry

2.4 Game lexicon

One of the most important design decisions was to restrict the lexicon of the map game. The restriction was communicated to players through the presentation of an "allowed words list" in the instructions (see above), and was enforced by monitoring and penalization. The decision to restrict the game lexicon was motivated by the desire to increase statistical power: in order to observe non-stationarity in linguistic behaviors, those behaviors must be sampled sufficiently often relative to the timescales of interest for analyses. By constraining the game vocabulary, the rates of occurrence of linguistic behaviors were increased. Notably, non-stationarity in a behavior may result from speakers experiencing examples of the behavior from other speakers; the lexicon constraints increase the frequency with which speakers experience exemplars of a given linguistic behavior from other speakers, thereby increasing the potential for observing social effects on behavior. The words in the game lexicon are categorized below, along with estimated category frequencies. The frequencies presented here are estimates, derived from automatic speech recognition (ASR, described below), which is accurate in general but not perfect.

Table 1. Word category frequencies

sizes 10.8% big, large, small, little	**colors** 12.8% red, green, blue	**shapes** 11.7% circle, square, triangle	**fills** 9.7% filled, unfilled
horizontals 9.0% left, right	**verticals** 12.4% up, down	**locatives/directionals** 1.9% by, near, to, from	**conjunction** 1.6% and
confirmations 12.8% okay, yes	**discourse/other** <1% no, not, or, wait, back, repeat, sorry	**location names** 13.3% boc, dija, fnop, pont, shub, sial, sond, suul	**hesitations** 1.9% uh, um

The choices of size, color, shape, and fill words were guided by the choice of location properties (discussed in next section). Two variants of each size term were allowed (*big*, *large*; *small*, *little*). Horizontal and vertical direction words (*left*, *right*, *up*, *down*) were included to facilitate gameplay. The locative preposition words (*by*, *near*) were included to allow players some additional options for communicating the context of route locations. The directional prepositions (*to*, *from*) were included mainly to create the potential for stylistic variation in giver instructions. These words were not used very frequently (1.9%), and most of these were the directional preposition *to* (1.7%). Pilot testing showed that confirmations (*okay*, *yes*) were helpful to coordinate the communication of route locations. Although *yes* was infrequent (< 0.1%), the confirmation *okay* was used by receivers very frequently to indicate they had found a correct location on a route. A handful of additional discourse-facilitating words were included because pilot testing judged them as helpful, but these were rarely observed in the experimental gameplay.

Because of the giver/receiver asymmetry in the game, frequencies vary greatly as function of player role. Fig. 14 illustrates the relative frequency distributions of words/word categories for givers and receivers. While givers used a large portion of the lexicon, about 80% of receiver words were *okay*. Receivers did use location properties to elicit further information from givers or sometimes to question the accuracy of giver instructions.

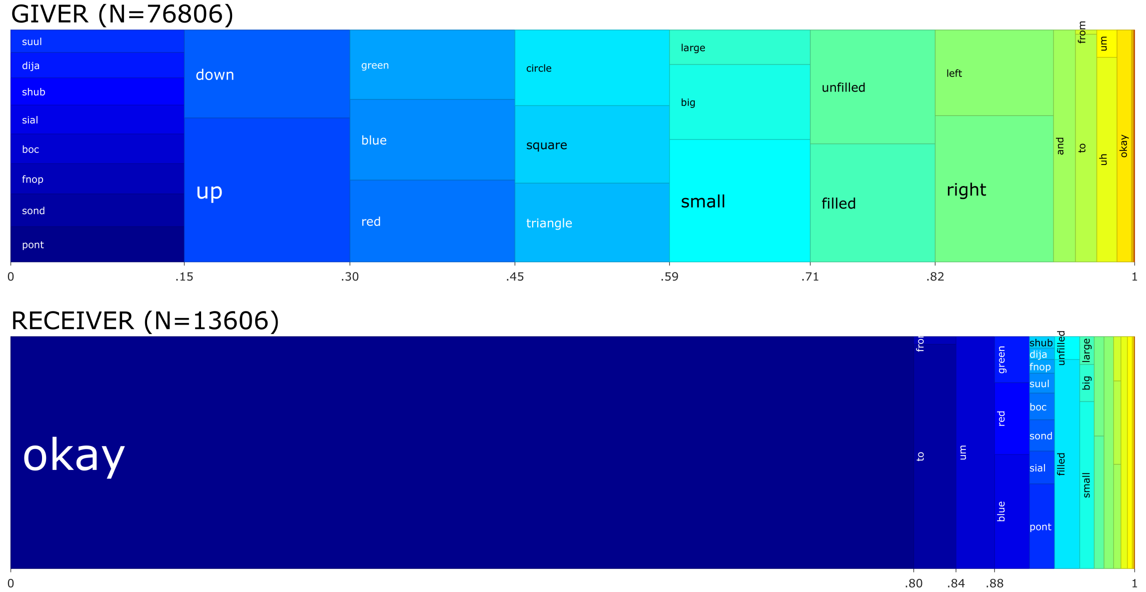


Fig. 14. Relative frequency distributions of words for givers and receivers. Words within categories having more than 1% of the frequency are arranged vertically.

2.5 Disallowed word penalties

Any word produced during a game that was not on the allowed words list was considered a violation. Filled pauses (*uh*, *um*) were not penalized. Interjections (e.g. *oh*) were also not penalized, but players were given a warning upon the second occurrence of an interjection in a game. Disallowed words which formed a phrase (*got it*, *found it*) were considered to be a single violation.

Prior to beginning the experiment, the planned protocol for notifying players of violations was for the monitoring experimenter to knock on a table or clap 3three times when a violation was detected. However, in the first session, the monitoring experimenters found it difficult to respond quickly enough. Hence we adapted the protocol in subsequent sessions so that notifications of violations were given only upon a repeated violation within a game. In the results presentation phase, the control experimenter asked the monitoring experimenters if any violations had been observed in the most recent round. If so, the monitoring experimenters stated the name of the player who produced the violation and the violating word/phrase. The tally of violations for the corresponding team was then incremented in the control gui GUI. When the finalized times and points were calculated, each violation resulted in 5 seconds being added to the game completion time.

Only 29 allowed word violations were observed across the experiment, and these are listed in the table below. In the first session there were 17 violations. Only 2 violations occurred in sessions 5-7, and no violations thereafter. The relative scarcity of violations indicates that participants did not have very much difficulty adapting to the lexicon constraints.

Table 2. Word penalties

session	round	player	role	violation
1	1	P2	g	got it
1	1	P2	g	found it
1	1	P4	r	next to
1	1	P4	r	I mean
1	1	P4	r	of
1	1	P3	g	way down
1	1	P3	g	to the
1	1	P5	g	so
1	1	P5	g	the first
1	1	P5	g	it's
1	3	P3	g	to the
1	3	P2	g	bottom
1	3	P1	r	bottom
1	5	P6	r	got it
1	6	P8	g	to the (right)
1	7	P2	g	bottom
1	7	P2	g	bottom
2	9	P2	g	I mean
2	11	P6	g	gotcha
2	11	P7	r	I mean
2	14	P5	r	above
2	17	P2	r	after
3	25	P3	g	(and) then
3	26	P4	r	alright
4	32	P3	r	aha
4	37	P6	r	alright
4	40	P3	r	two seconds
6	61	P2	g	after
7	81	P7	r	next

2.6 Location names and symbol properties

The number of properties used for location symbols, and number of location names, were important design decisions. These decisions can be contextualized relative to two extremes of variation in location properties. On one extreme, the minimal amount of variation could be accomplished by utilizing no location symbol properties whatsoever: all locations would be represented by, for example, a white dot on the screen. On the other extreme, variation could be greatly increased by using a very large set of location symbol properties and values (e.g. many colors, many different shapes, sizes, textures, etc.). The minimal-variation extreme would force communication to rely on horizontal/vertical direction words and location names. Although this would have the benefit of increasing statistical power in analyses, there would be less space for variation in the choice and ordering of words in giver instructions. Gameplay might also be overly difficult or boring in such circumstances, which could increase the likelihood of participant drop-outs. In contrast, the high-variation extreme would drastically reduce statistical power by decreasing the number of tokens produced for any given category.

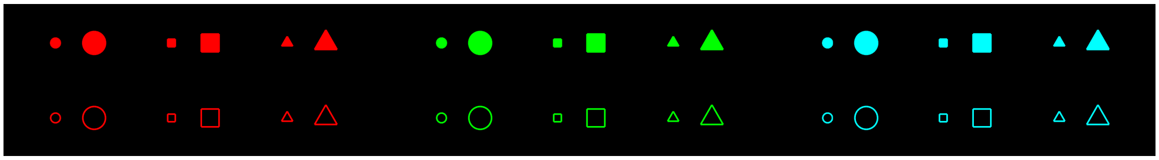


Fig. 15. The full set of 36 location symbols. Symbols varied by color, shape, fill, and size.

Based on the design aim to achieve high statistical power, intuitions regarding game difficulty, and extensive gameplay testing, the following location properties (and values) were used: 3 colors (red, green, blue), 3 shapes (circle, square, triangle), 2 sizes (large, small), and 2 fills (filled, unfilled). Thus 36 unique symbols were used. The three colors were approximately as shown in the figure, with normalized RGB values [1 0 0] (red), [0 1 0] (green), and [0 1 1] (cyan). The blue was actually cyan (green+blue) rather than pure blue, because pure blue is too dark to be easily seen on a black map background. The small/large symbols were approximately 0.65/1.3% of screen width (1366 pixels) and 1.1/2.2% of screen height, with small variations in size across shapes due to software idiosyncrasies.

Another major group of decisions was the number of location names to use, along with the phonological and orthographic features of those names. With fewer names statistical power is increased but the space in which variation occurs is more limited; gameplay is also more difficult with fewer names because each name becomes less informative. On the other hand, with too many names statistical power is decreased. Based on these considerations, eight unique location names were chosen. Because each map had 65 locations, this implies that on average each name occurs about eight times on each map. Moreover, assuming each location on a route is produced once in each game, givers will produce each name an average of 2.375 (= 19/8) times per game.

The location names were designed so as to maximize the potential for observing experiment-timescale dynamics in categorical and sub-categorical properties. To this end, we constructed non-word, unfamiliar names, because exemplar models predict that pronunciations of unfamiliar/novel items are more likely to be affected by conversational interactions than familiar ones (Johnson, 1997; Pierrehumbert, 2002). The location names are listed in the table below. The orthographic compositions of the names were very important because participants never heard them until the first round of gameplay. In some cases the stimuli were deliberately designed to have ambiguous mappings between orthography and phonological categories, in order to create the potential for dynamics at the level of phonological categories. For example, the intervocalic grapheme in *dija* might be interpreted as a voiced post-alveolar affricate or fricative. For another example, the vowel in *suul* might be interpreted as a tense or lax high back vowel.

Table 3. Location names, anticipated categorical variation, and contrasts

name	categorical variation		contrasts
boc			
pont			pont, sond
sond			sond, suul, shub
suul	[uː]	[ʊ]	
shub	[uː]	[ʌ]	
sial	[ai]	[i.a]	
dija	[dʒ]	[ʒ]	
	'σ.σ	σ.'σ	

The majority of the names were designed to be unambiguously monosyllabic, to avoid interactions that might arise between variation in metrical structure and segmental acoustics. However, the name *dija* was included specifically to allow for metrical variation, and the names *suul* and *sial* were included because previous studies have shown that some liquid rimes exhibit interspeaker variation with regard to syllable count judgments (Tilsen, Cohn, & Ricciardi, 2014). Some aspects of the names were designed to provide specific contrasts or environments. The codas of *pont* and *sond* were chosen to allow for analyses of reduction in nasal + voiceless/voiced stop combinations. Although a number of conscious decisions were made in designing the location names, in order to elicit specific contrasts and anticipated forms of categorical variation, unanticipated forms of categorical variation were also expected to arise. Moreover, in addition to categorical variation, the names were expected to provide numerous forms of sub-categorical phonetic variation. Four of the names were designed as sibilant initial because sociolinguistic variation in the distribution of sibilant spectral energy has been extensively studied (Munson, 2007; Stuart-Smith, 2007). The name *pont* allows for measurement of VOT, which has been studied previously in a longitudinal manner (Sonderegger, 2012). All vowel qualities were considered candidates for analysis.

2.7 Map design and generation algorithm

Before the experiment began, 200 unique maps were randomly generated. Each map had 65 locations. All location symbol properties and names were randomly selected without replacement from the set of all possible combinations of names and symbol property values (N = 8 names x 3 colors x 3 shapes x 2 sizes x 2 fills = 288 possible location name-symbol sets). Random selection without replacement was used in order to prevent identical locations from occurring in the same map. The number of locations on each map (65) was chosen with several considerations in mind: too few locations makes the game too easy, too many makes it impractical to randomly position the locations and location names without overlapping names with other names or symbols. The names and symbols must be large enough so that players can read and see them without difficulty, but the maps should not be visually cluttered.

In generating the maps, coordinate axes with normalized units from 0-1 in horizontal and vertical directions were defined. Positions of the 65 locations were randomly determined with an iterative procedure, discarding candidates that were less than 0.05 normalized distance units from any previously generated location. Next a 20 location (19 segment) route was randomly defined as follows. The leftmost and rightmost locations on the map were chosen as the starting and ending locations, respectively. Then all possible line segments within a range of 0.05-0.5 normalized units from the starting location were identified. Any candidate segments which crossed but did not terminate on a location symbol were excluded. One of the candidate segments was randomly selected and the procedure was iterated relative to the previously selection location. For all segments but the first one, candidates that made an angle with the preceding segment between [-7.5°, 7.5°] or [-172.5°, 172.5°] were excluded, as pilot tests showed adjacent segments making small angles could be mistaken for one segment. After selecting the 19th location (18th segment), the final segment was chosen to end on the rightmost location on the route. If this final segment exceeded the maximum length (0.5 normalized units), the route was discarded and a new one was randomly generated until all the above criteria were met.

Location name positioning was accomplished by allowing for six possible alignments/orientations of location names relative to location symbols. These were right/horizontal, left/horizontal, left/angled-up, left/angled-down, right/angled-up, and right/angled-down (angles were 45° and -45°). For each location on the map (selected in random order), a random preference order of the horizontal orientations was chosen, followed by a random preference order of angled orientations. Then, the randomized alignment-orientation sets were tested until one was found which did not result in a location name overlapping with any location symbols or previously positioned names. In most cases (98.8%), either a left- or right-aligned horizontal orientation was selected. In some cases, no alignment satisfied the overlap criteria and the entire map was discarded. The generation algorithm failed to find a solution for positioning location names about 80% of the time when 65 locations were used (the rate fell below 50% with 80 locations). The procedures were iterated to generate 200 maps, and the first 134 of these were used in the 134 rounds of the experiment. When maps were displayed during games, 5% margins were added so that no names were off of the screen. A Java function was used to clip the MATLAB figure window title bar and remove the figure window border.

2.8 Surveys and rankings

Before and after each game, players completed pre-game and post-game surveys. The surveys consisted of four questions, each of which was answered with a numerical value on a seven-point scale. The questions are shown below. The questions in the pre- and post-game surveys paralleled each other and can be grouped into four categories: *enthusiasm-player*, *enthusiasm-teammate*, *performance-team*, and *performance-teammate*. The first three questions were posed with respect to the current game and future games in the pre- and post-game surveys, respectively. For example, the pre-game player-enthusiasm question asks how enthusiastic the respondent is to play with their assigned teammate in the current game, and the post-game question asks essentially the same thing regarding future games. Differences between pre- and post-game responses thus may reflect changes in interplayer social systems. The same holds for the teammate-enthusiasm, team performance, and teammate assessment questions. The teammate assessment question differs slightly in that the post-game question refers to performance in the just-completed game, rather than future performance. The extreme values of the response scale (1 and 7) were labelled as shown in the tables below (e.g. “not very enthusiastic” vs. “very enthusiastic”). In all cases the upper end of the scale connoted a positive assessment.

Pre-game survey

		not very enthusiastic				very enthusiastic	
How enthusiastic are you to play the game with [teammate]:	1	2	3	4	5	6	7
How enthusiastic do you think [teammate] is to be playing the game with you:	1	2	3	4	5	6	7
		not very likely				very likely	
How likely do you think you are to win this round:	1	2	3	4	5	6	7
		not very good				very good	
How good is [teammate] at [giving/receiving] directions:	1	2	3	4	5	6	7

Post-game survey

		not very enthusiastic				very enthusiastic	
How enthusiastic would you be to play the game with [teammate] in the future:	1	2	3	4	5	6	7
How enthusiastic do you think [teammate] will be to play the game with you in the future:	1	2	3	4	5	6	7
		not very likely				very likely	
How likely do you think you and [teammate] would be to win in the future:	1	2	3	4	5	6	7
		not very well				very well	
How well did [teammate] [give/receive] directions:	1	2	3	4	5	6	7

All players were required to select a value for each survey question. The questions were presented one at a time, and the response options were selected with radio buttons. No option was initially selected. Players clicked a *done* button at the bottom of the screen to proceed to the next question. In the first two sessions it was anticipated that participants might not be familiar enough with one another to answer some of the questions, and so the middle response on each scale (value 4) was labelled as “not sure”. This label was removed in the third session.

Survey response dynamics and distributions are shown in Fig. 16. Pre- and post-game survey responses are in general not the same for a given question in a given round. This is also evident from inspection of the experiment-wide distributions of responses for each participant in each question. The variation shows that participants were not simply picking one value out of expedience. One exception seems to be participant P5, whose responses to the enthusiasm-player and team-assessment questions were relatively unchanging from the 4th and 5th rounds onward. Note that the survey responses provide a somewhat sparse matrix of respondent-teammate observations: because the survey responses are specific to the randomly assigned teammate in each round, any given respondent-teammate observation only occurs when those players are paired together.

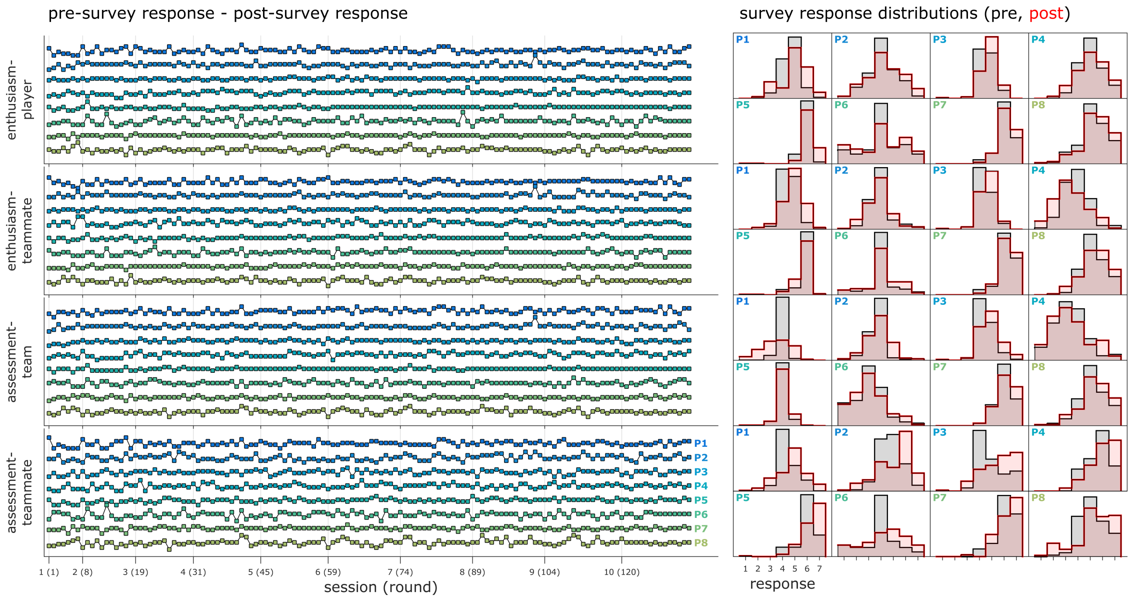


Fig. 16. Summary of survey response variation. Left: variation over time in responses to pre- and post-game survey questions. Right: histograms of pre- and post-game survey responses for each question and player.

Rankings elicited after each post-game survey provide a full density matrix of player-player social measures. Players were required to rank in order all other players with regard to whom they most/least preferred to play within the next round. They did this by selecting names from seven vertically arranged drop-down lists, each of which contained the names of all seven of the other players. In all of the drop down lists, no player names were initially selected, and thus each ranker had to intentionally choose a name in each drop-down list. Within each drop-down list, player names were ordered according to the current standings, i.e. the total number of points the players had accumulated in preceding rounds. The player with the highest standing appeared at the top of each list. Hence when a ranker ranked a player in a position above their standing, this meant the ranker purposefully decided to make that player a more likely teammate in the next round. Conversely, when a ranker ranked a player in a position below their standing, this meant the ranker purposefully decided to decrease the likelihood they would be paired with that teammate in the next round. When two or more players were tied in the standings, the order of their names in the drop-down lists was determined by a random permutation of players that had been created before the experiment began (and according to which the players are coded: P1, P2, etc…)

Players presumably use the rankings to attempt to influence who their teammate will be in subsequent rounds. Substantial variation in these rankings suggests that interpersonal attitudes in the social network of players changed over the course of the experiment. Fig. 17 shows all player-player ranking time-series. In each panel of the grid, rows correspond to rankers and columns to rankees. Gray lines show standing of rankee, black lines show the rankers (rows) rankings of rankees (columns). Red and blue indicate rankings higher and lower, respectively, than the standing (default ranking). Note that there are ceiling and floor effects in the difference between ranking and standing: the rankees with highest/lowest standings cannot be ranked any higher/lower.

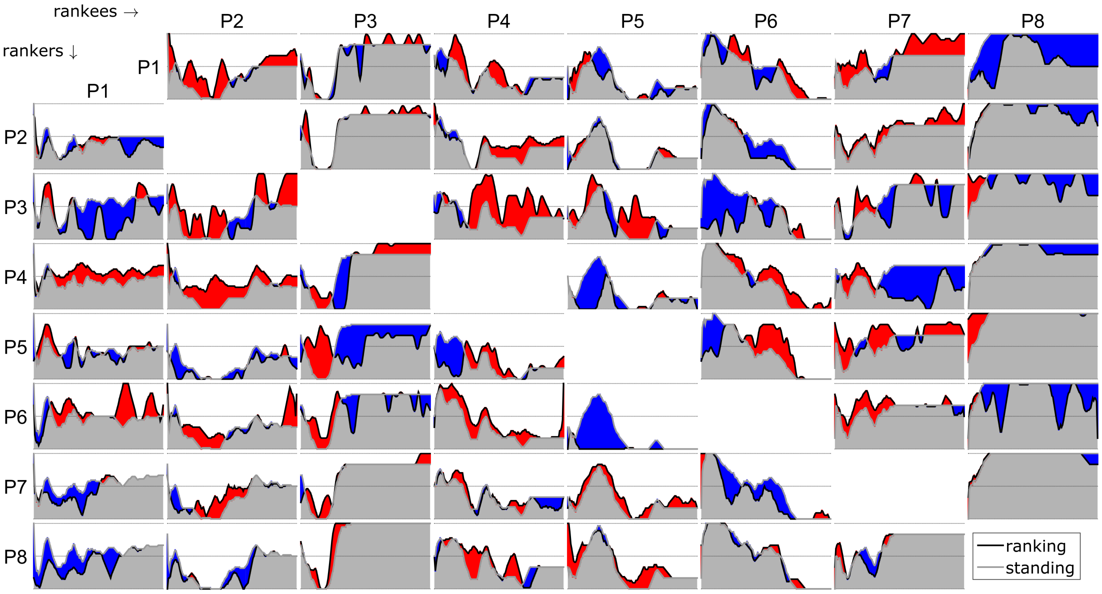
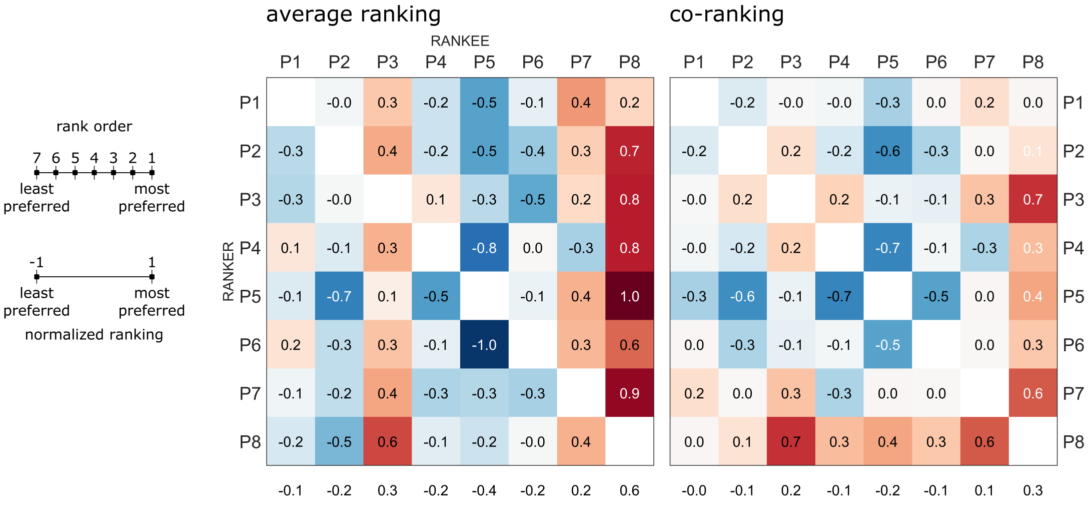


Fig. 17. Ranking time series. Rows correspond to rankers, columns to rankees. Gray lines show standings (default ranking), black lines show ranking; red/blue indicate positive/negative difference between ranking and standing.

The average player-player rankings across the experiment show some noteworthy patterns (see Fig. 18). The average rankings are asymmetric, but the *co-ranking*, defined as the average mutual ranking between a pair of players, is symmetric. Note that the values in the figure are normalized such that -1 and 1 represent least and most preferred, respectively. Of particular interest in the rankings are rankees who were ranked relatively high or low on average, e.g. P8 with an average ranking of 0.6 and P5 with an average ranking of -0.4. These players were consequently involved in some of the highest magnitude ranking asymmetries, e.g. P5-P6 (-1.0 vs. -0.1) and P2-P8 (-0.5 vs. 0.7).



average ranking

	P1	P2	P3	P4	P5	P6	P7	P8
P1		-0.0	0.3	-0.2	-0.5	-0.1	0.4	0.2
P2	-0.3		0.4	-0.2	-0.5	-0.4	0.3	0.7
P3	-0.3	-0.0		0.1	-0.3	-0.5	0.2	0.8
P4	0.1	-0.1	0.3		-0.8	0.0	-0.3	0.8
P5	-0.1	-0.7	0.1	-0.5		-0.1	0.4	1.0
P6	0.2	-0.3	0.3	-0.1	-1.0		0.3	0.6
P7	-0.1	-0.2	0.4	-0.3	-0.3	-0.3		0.9
P8	-0.2	-0.5	0.6	-0.1	-0.2	-0.0	0.4	
	-0.1	-0.2	0.3	-0.2	-0.4	-0.2	0.2	0.6

co-ranking

P1	P2	P3	P4	P5	P6	P7	P8	
	-0.2	-0.0	-0.0	-0.3	0.0	0.2	0.0	P1
-0.2		0.2	-0.2	-0.6	-0.3	0.0	0.1	P2
-0.0	0.2		0.2	-0.1	-0.1	0.3	0.7	P3
-0.0	-0.2	0.2		-0.7	-0.1	-0.3	0.3	P4
-0.3	-0.6	-0.1	-0.7		-0.5	0.0	0.4	P5
0.0	-0.3	-0.1	-0.1	-0.5		0.0	0.3	P6
0.2	0.0	0.3	-0.3	0.0	0.0		0.6	P7
0.0	0.1	0.7	0.3	0.4	0.3	0.6		P8
-0.0	-0.1	0.2	-0.1	-0.2	-0.1	0.1	0.3	

Fig. 18. Average normalized rankings and co-rankings. Rankings are asymmetric; co-rankings are the average of the rankings between player pairs.

One interpretation of the ranking is as a preferred social distance (or more specifically, preference for social interaction in the context of the game). Under this interpretation, the co-rankings represent the average mutual preferred social distance between a pair of players.

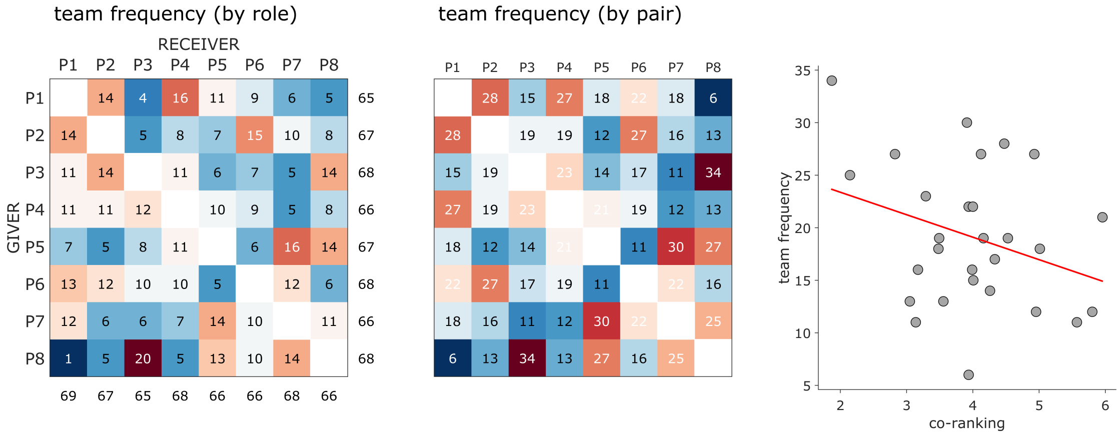
Examination of average co-rankings shows several pairs with high or low mutual preference for social interaction, e.g. P3 and P8 (also the two most-winning players), or P5 and P4.

The ranking and survey data are important for hypotheses regarding social-linguistic covariation because they provide observations of social dynamics which may be correlated with linguistic behavioral dynamics. However, there are many possible approaches to extracting parameters which represent the structure of the participant social network from the survey and ranking data. Future analyses will need to assess how these data can be used to construct parameters suitable for correlation with linguistic behavior parameters.

2.9 Team generation algorithm

Team generation was random but biased by player rankings. An unbiased random or pseudo-random algorithm could have been used to maximize the uniformity of pairwise interaction frequencies between players, and this would have the advantage of minimizing confounds between interaction frequency and other factors which could influence linguistic behavior dynamics. However, increasing the likelihood that a player would care about how they rank the other players was a higher priority. The player ranking bias had the important effect of making the rankings matter for the participants, without any need for deceiving them. In other words, participants were more likely to attend to their ranking decisions because they knew those decisions could impact who their teammate would be in the next round. This aspect of the design was explicitly emphasized in the instructions the participants were given in first session.

The algorithm for team generation was implemented as follows. First, a set of all possible candidate mappings of players to teams and roles was created. Then, any mappings which would result in an experiment-wide role imbalance for any player were discarded. In the first session a role imbalance was defined as a greater than two-game difference in games played as receiver and games played as giver. In subsequent sessions, the threshold for an imbalance was a greater than four-game difference. Hence in the first session, players would play both roles at least 2 or 3 times, and in subsequent sessions they never reached a role-imbalance of more than four games. Candidate mappings which would result in two players being on the same team in three consecutive games (in either role) were also excluded. The remaining candidate mappings were treated as objects in a multinomial distribution whose probabilities were determined by summing all the pairwise rankings from the proceeding round that were consistent with the participant teams in a mapping. For example, if a candidate mapping was P2-P3, P1-P5, P4-P7, P8-P6, then rank(P2,P3), rank(P3,P2), rank(P1,P5), rank(P5,P1), etc. were summed (rankings were derived by subtracting rank order from the number of players, so that higher rank order results in greater probability mass). These sums were normalized by dividing by the sum over all candidates, in effect creating a probability distribution over mappings. Thus when players ranked each other highly, all candidate mappings which put them on the same team would have a relatively high probability of being selected. The player-player pair frequencies across the experiment are shown in Fig. 19.



GIVER	P1	P2	P3	P4	P5	P6	P7	P8	
P1		14	4	16	11	9	6	5	65
P2	14		5	8	7	15	10	8	67
P3	11	14		11	6	7	5	14	68
P4	11	11	12		10	9	5	8	66
P5	7	5	8	11		6	16	14	67
P6	13	12	10	10	5		12	6	68
P7	12	6	6	7	14	10		11	66
P8	1	5	20	5	13	10	14		68
	69	67	65	68	66	66	68	66	

	P1	P2	P3	P4	P5	P6	P7	P8
		28	15	27	18	22	18	6
	28		19	19	12	27	16	13
	15	19		23	14	17	11	34
	27	19	23		21	19	12	13
	18	12	14	21		11	30	27
	22	27	17	19	11		22	16
	18	16	11	12	30	22		25
	6	13	34	13	27	16	25	

Fig. 19. Team frequencies. Left: team frequencies by player roles: givers (rows), receivers (columns). Right: team frequencies by player pairs.

The empirically observed effect of the rankings bias on team frequencies was not very large: Fig. 19 also shows a linear relation between average co-ranking and team frequency. The coefficient relating team frequency and average co-ranking was *b*=-2.13, indicating that an increase of one co-rank order resulted in an additional two games played across the experiment. Hence a co-rank increase of 1 might be expected to increase the likelihood of specific pairing in a subsequent round by approximately 1.5%, and so a pair of players who ranked each other as most preferred teammates would be about 9% more likely to be paired than a pair of players who ranked each other as least preferred teammates.

2.10 Player performance statistics

At the end of each round, players were shown the player standings (see Fig. 12), which were sorted according to total points. The average game time for each player was also shown. The values of these statistics over time for each player are shown in Fig. 20. P8 became the standings leader in session 3 and remained the leader through the end of the experiment. P7 and P3 vied for 2nd place in the standings in sessions 7-10. P6, who had a commanding lead in the standings after the first session, finished last in the end.

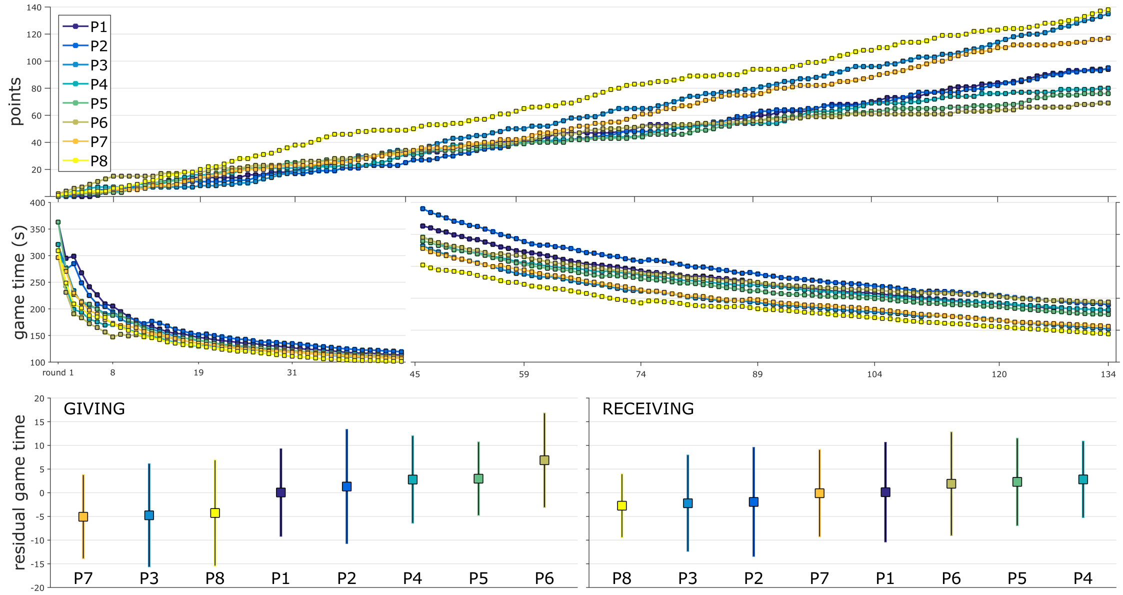


Fig. 20. Player statistics and residual game time by role. Top: player points over the experiment. Middle: player game time over the experiment; game times from sessions 1-4 and 5-10 are plotted on different scales. Bottom: Mean (±1 s.d.) residual game time from linear regression with predictors *round* (ordinal) and *teammate* × *role*; players are sorted by residual game time for each role.

Quantification of player performance is of interest for addressing certain questions about gameplay and motivations for rankings/survey responses. Fig. 20 (bottom) shows mean residual raw game times for each player and role. The raw game times were used here rather than the penalty-adjusted times, because the click and disallowed word penalties do not directly represent the performance of the giver or receiver regarding the central aspects of gameplay. The residuals were obtained from regressions with predictors *round* (as an ordinal variable) and *teammate* × *role*. The teammate-role interaction was included so that teammate effects on game times are factored out of the residuals. The resulting performance measures show about a 10 s performance advantage from the highest- to lowest-performing givers, and a 5 s performance advantage across receivers. It is noteworthy that P8, the player who received the highest overall rankings, was the highest-performing receiver and among the three highest-performing givers. In contrast, the three players ranked lowest on average (P4, P5, P6) were also the three lowest-performing players, both as receivers and givers.

2.11 Automatic speech recognition procedures

Approximately 26.25 hours (1575 minutes) of audio were collected during the experiment. In order to facilitate analyses, the HTK-HMM speech recognition toolkit (Young et al., 2002) was used to generate word- and phone-level time-aligned transcripts for each game. This procedure involves five main steps: receiver-giver channel alignment, manual labeling of training data, phone- and triphone-HMM training, recognition, and pruning.

First, the receiver and giver audio channels from each game were aligned in time. This was necessary because the laptops did not communicate with one another in real-time, but instead by writing and reading frequently from a server. Thus the giver- and receiver-audio recordings in each game begin at slightly different times (mean/standard deviation/maximum of estimated absolute lags were 303/203/875 ms). However, receiver and giver audio from each game can be readily synchronized because to some extent the speech of each player is picked up by the microphone of the other. To estimate the lag between channels, each channel was divided into

1000 ms frames with 500 ms overlap and the lags associated with maximal cross-correlation between frames were calculated for each pair of frames. Lag counts from all 10-40 bin partitions of one second were calculated, and the estimated lag was taken as the mean of lags in the highest count bin from the partition with lowest relative entropy (i.e. the partition with the most "peaked" distribution). Manual inspection of cross-channel waveform properties in randomly selected games showed maximal discrepancies of about 2 ms between channels.

Second, the instructor and graduate students in the social network map game seminar manually labeled words and phones in the giver and receiver audio of eight randomly selected games from the last three rounds the first session, and of eight games from rounds in the middle half of the fourth session. The random selection was constrained so that at least two instances of giver audio would be selected from each participant. The labeling was conducted in Praat with the aligned receiver and giver audio and a five-tier textgrid. Four of the tiers were the word and phone labels for receiver and giver audio, respectively. The fifth tier was used to mark word violations, disfluencies, and non-speech noises. It was not possible to know ahead of time what forms of variation might exist in the pronunciation of allowed words and non-penalized forms (filled pauses). A first-pass pronunciation dictionary was developed on the basis of labeler transcriptions, and was subsequently pared down to reduce labeler-specific variation. The reduced set of pronunciations was used to update the manual labeling.

Third, the manual labeling was used to estimate 5-state hidden Markov models (HMM) of triphones in the training data. The giver and receiver audio from all games was converted to 16-coefficient Mel-frequency cepstral vectors with deltas and accelerations included (26 filter bank channels, preemphasis 0.97, window size 25 ms, frame step 10 ms). Procedures described in the HTK reference manual (Young et al., 2002) were followed for estimating triphone HMMs. This involves initializing and iteratively re-estimating monophone HMMs, subsequently constructing triphones from all phone sequences observed in the training data, and then estimating triphone HMMs. HTK state-tying procedures were then applied and triphones were re-estimated.

Fourth, the tied-state triphone models were used to conduct automatic speech recognition all of the audio data from the experiment. Separate word network models were estimated from receiver/giver data, and used for recognition in the receiver/giver audio, respectively. Lastly, the output labels from the recognition were pruned and sanitized. Pruning was necessary because many false recognitions of *uh* and *okay* during silent periods are generated in the recognition process and because giver audio is sometimes recognized in the receiver channel. To address these issues, two pruning procedures were implemented. First, ignoring tokens of *okay*, word intervals which were recognized at approximately the same time in the giver and receiver audio were re-labelled as silence in the receiver audio. Second, regressions of log-RMS intensity and log-duration from each token with predictor terms *word*, *game*, and a *player* × *role* interaction. An elliptical 99% prediction region for the log-RMS intensity and log-duration residuals was calculated, and tokens of *uh* and *okay* with residual log-RMS intensity below 0 and outside of the 99% prediction region were relabeled as silence. This procedure was iterated twice. Subsequently consecutive silence intervals were merged and word-phone level consistency was validated.

Some details of the automatic speech recognition procedure have been omitted from the above description, and refinements of the procedure are ongoing. In particular, the output of the recognition process, with some manual editing, can be used provide more training data, which can ultimately be used to estimate speaker-specific HMMs. Many decisions are made in each step of the process described above, and further manipulations of these decisions remain to be explored. Nonetheless, the current scheme is remarkably accurate when recognition output is evaluated against the training data: 96.4% of giver words are accurately recognized. This rate is substantially lower for receivers, 78%. The false alarm rates are 13% and 34% for giver and receiver recognition, respectively. Fortunately, the vast majority of the false alarms for receivers involve *uh* and *okay* tokens that fail to be pruned. Disfluencies of various sorts are not uncommon in the dataset and the majority of these cannot be identified automatically. In future analyses, speaker-specific HMMs will be used and manual correction of the automatic recognition will be conducted.

3. Preliminary results

For illustration, we show here a somewhat puzzling, preliminary result. The result is odd because of how strong it is, and because it shows an unexpected pattern. Recall that there were eight location names in the experiment. For each token of a location name produced in the experiment, an average auditory spectrum was calculated from an auditory spectrogram of the middle-third of the waveform of the vowel. (The vowel was demarcated by the HMM-based alignment; auditory spectrograms were comprised of 64 erb-scaled gammatone filters, from 70-10000 Hz, using a 20 ms window and 5 ms steps). Then, for each location name, all of the spectra were transformed to principle components (spectra with outlying values in more than half of the 64 bins were excluded; outliers were identified as deviations > 2.32σ, with standard deviations estimated separately for each speaker). Then low-dimensional representation of the trajectories of vowel parameters for each speaker were estimated by taking a Gaussian-weighted average of the first three principle components of the vowel parameters over windows ranging from 1 to 67 rounds. The distance between any two trajectories on a given timescale is then the Euclidean distance between points in the space of the first three principle components. The same procedure was used to construct social-distance trajectories using the co-ranking values associated with player-player pairs.

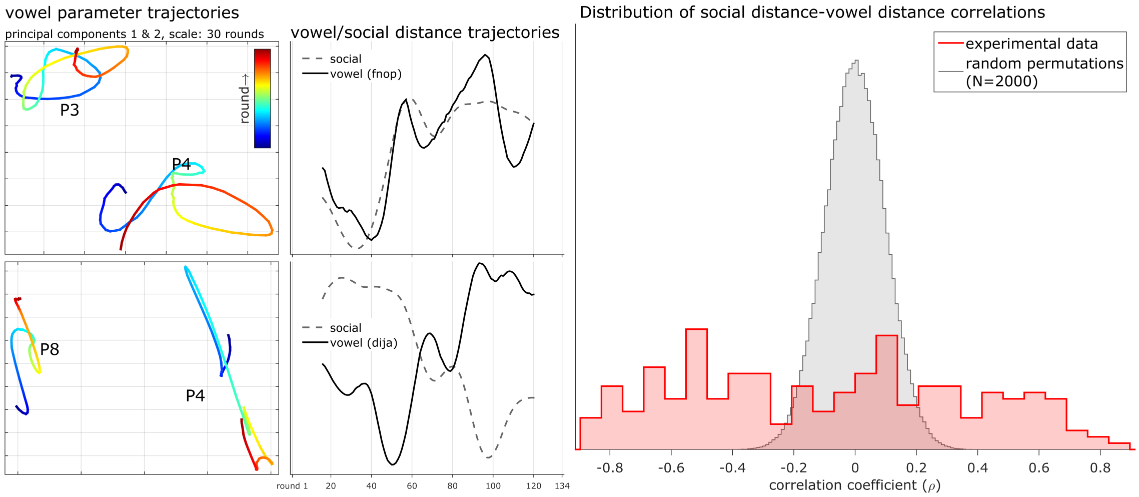


Fig. 21. Illustration of vowel-distance social-distance correlation analysis. Left, top: an example of high positive correlation between participants P3 and P4; vowel parameter trajectories (from *fnop*) in the space of the first two principle components; overlay of (re-scaled) vowel and social distance trajectories. Left, bottom: an example of high negative correlation, participants P8 and P4, first vowel in *dija*. Right: at a scale of 30 rounds, the overall distribution of correlation coefficients (red) compared to expected distribution derived from random permutations of values within trajectories (gray).

The figure above shows a pair of vowel and social distance trajectories that are strongly correlated positively (vowels from *fnop*, scale 30 rounds), and another with a strong negative correlation (first vowel from *dija*, scale 30 rounds). The figure also shows the distribution of correlation coefficients between vowel distance trajectories (with vowel always the same) and social distance trajectories measured on a 30-round timescale, and contrasts this distribution with an expected distribution derived from randomly permuting values within trajectories. The random permutation removes the time-dependence of the observations, effectively destroying any temporal correlation structure between observables.

The remarkable result of this analysis is that strong correlations between social distance and vowel distance are much more prevalent than expected by chance. What is puzzling is that both positive and negative correlations are observed. The positive correlations are expected if social dynamics influence linguistic dynamics, assuming that the measure of social dynamics (changes in co-rank) does reflect changes in social systems. But why would negative correlations also occur? The negative correlations indicate that in some cases the social distance between speakers changed in the opposite direction of vowel distance. The effect is not an accident of the scale chosen in the above analysis. When examining how correlation magnitudes vary as a function of analysis scale, we see that empirical correlation magnitudes grow faster than expected as analysis scale is increased (Fig. 22). Only on very short timescales are empirical correlation distributions similar to the expected ones.

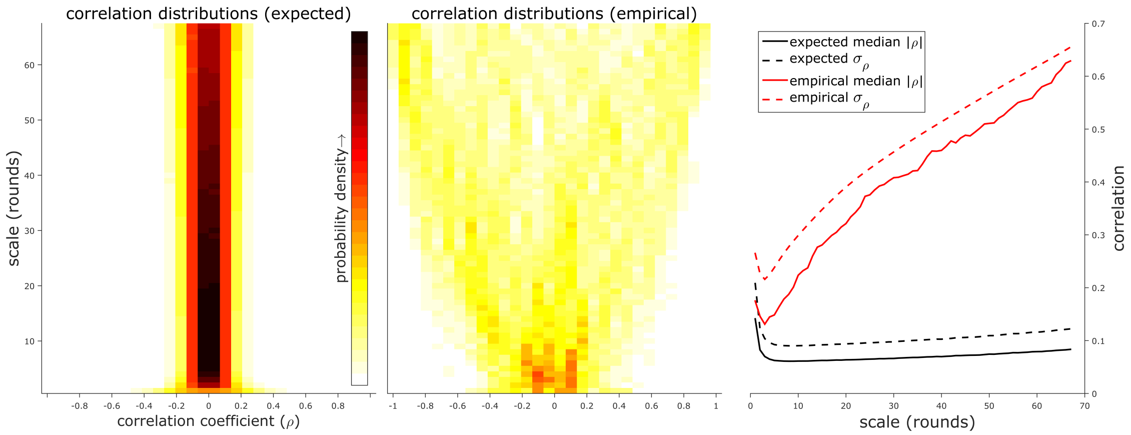


Fig. 22. Empirical and expected correlation distributions as a function of scale. Left: expected correlation distributions derived from random permutations of data. Center: empirical correlation distributions. Right: comparison of mean and standard deviation of empirical and expected correlation magnitudes as a function of scale.

The occurrence of negative correlations is not consistent with the notion that social interactions between people with mutually positive attitudes toward one another promote similarity of linguistic behavior (convergence), and that mutually negative attitudes promote dissimilarity in linguistic behaviors (divergence). However, there are many possible explanations for this inconsistency which require further effort to resolve. Maybe other factors must be considered, such as changes in speech rate and changes in syntactic patterns over the experiment. Once these are appropriately externalized, the expected patterns might emerge. Perhaps the co-ranking metric of social distance may be oversimplified: social distance effects could have nonlinearities not captured by the metric; for example, the effects of one social distance might overwhelm the effects of all other social distances for given player—more sophisticated analyses of rankings might resolve this.

4. Conclusion

The social network speech dynamics experiment provides a very rich set of data for studying how speech and social networks co-evolve. Beyond simply describing the experiment, I have emphasized that in analyzing these data, we should be more aware of the spatial and temporal scales of our analyses. One important part of this approach to analysis is a deliberate, conscious attention to the physical grounding of our analytical categories. I have also discussed two metaphors—projections and saddle-point equilibria—for conceptualizing how our analytical categories relate to the physical world.

One criticism that arises relates to the "ecological validity" of the experiment. The experiment represents an "unnatural" context and hence an "unnatural" behavior. Such criticisms beg the question of how "naturalness" is defined, and what constitutes a "valid" degree of this property. Indeed, any conception of naturalness must be a projection which emphasizes some aspects of a more complicated reality, at the expense of others. Can we statistically motivate "naturalness" of a context and behavior?

Another criticism that might arise is that information has been neglected about common social categories such as gender and sexuality, or other aspects of personality which might characterized by batteries of cognitive/psychological tests (e.g. autism quotient, working memory span, etc.). Social categories and personality traits, like all other categories, are projections. If they are useful projections, they should emerge statistically. Yet the sample size of this experiment (eight individuals) is too small to draw any conclusions from in this regard.

Ultimately the current experiment emphasizes temporal dynamics, at the expense of spatial variation (i.e. a larger population of speakers or a larger set of linguistic behaviors). Plenty of studies have examined linguistic behaviors in a large sample of speakers. Few have examined behaviors in even one speaker on the timescale of the current experiment. It is difficult and costly to collect data of this sort. Yet data on these timescales may be crucially important for advancing our understanding of speech.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., … others. (1991). The HCRC map task corpus. *Language and Speech*, *34*(4), 351–366.

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Gibson, J. J. (1979). The ecological approach to visual perception.

Hebb, D. (1949). The Organization of Behaviour: A Neuropsychological Theory.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. *Talker Variability in Speech Processing*, 145–165.

Kelso, J. (1982). *Human motor behavior: An introduction*. Routledge.

Kelso, J. (2009). Synergies: atoms of brain and behavior. In *Progress in motor control* (pp. 83–91). Springer.

Kelso, J., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*.

Kondepudi, D., & Prigogine, I. (1998). *Modern thermodynamics: from heat engines to dissipative structures*. John Wiley & Sons.

Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech*, *50*(1), 125–142.

Nicolis, G., & Prigogine, I. (1977). *Self-organization in nonequilibrium systems* (Vol. 191977). Wiley, New York.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, *7*, 101–139.

Schrödinger, E. (1944). *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press.

Shannon, C. E. (1948). A Mathematical Theory of Communication.

Sonderegger, M. (2012). *Phonetic and phonological dynamics on reality television* (Vol. Doctoral Dissertation.). University of Chicago. Retrieved from http://dl.acm.org/citation.cfm?id=2520161

Spivey, M. (2007). *The continuity of mind*. Oxford University Press.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production:/s/in Glaswegian. *Laboratory Phonology*, *9*, 65–86.

Tilsen, S. (2014). Selection-coordination theory. *Cornell Working Papers in Phonetics and Phonology, 2014*, 24–72.

Tilsen, S. (2016). Selection and coordination: the articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, 55: 53-77.

Tilsen, S., Cohn, A., & Ricciardi, E. (2014). Syllable count judgments and durations of liquid rimes in English. *Cornell Working Papers in Phonetics and Phonology*.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., … Woodland, P. (2002). The HTK book. *Cambridge University Engineering Department*, *3*, 175.

Department of Linguistics
Cornell University
Ithaca, NY 14850
tilsen@cornell.edu