Performance Approaches to Semantics in Human Language

by

Dillon Niederhut

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Anthropology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Terrence Deacon, Chair
Mariane Ferme
Terry Regier

Summer 2017

Performance Approaches to Semantics in Human Language

Abstract

Performance Approaches to Semantics in Human Language

by

Dillon Niederhut

Doctor of Philosophy in Anthropology

University of California, Berkeley

Terrence Deacon, Chair

Over the course of human history, prominent hypotheses for the source of human behavioral uniqueness have included freedom from metaphysical necessity, rational intention, and strong social altruism. However, when applied to questions concerning the evolution and current use of human language, these hypotheses generate more problems than they solve. These include everything from the statistical improbability of "hopeful monster" mutations to the unstable strategy of altruism without reciprocity. Adopting a performative view of human language removes these problematic dependencies, and explains aspects of language like phatic speech, fossils, and register, that require special pleading by one or more of the popular hypotheses.

This thesis provides two pieces of computational evidence for a performative view of human language. First, it shows that the possibility space of human language is much smaller than expected for a rule based but otherwise semantically unrestricted communication system. Second, it shows that the semantic information contained in an entire speech act is poorly predicted by the semantic information of its individual parts, as might be expected for a compositional system designed for efficient communication.

To conduct this second analysis, a novel method was developed based on the distributional theory of semantics for measuring the information content of human speech. Briefly, it is a measure of the distance between the distribution of terms in a language and the conditional probability of their appearance within a context of interest. This same method shows promise for testing other hypotheses about the nature and origin of language that involve an informational component.

This work is dedicated to the dispersed group of faculty at Berkeley who inspired me to leave my silo.

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to Terrence Deacon for his deep understanding of information, communication, and evolutionary stable systems. The influence of his ideas, especially his analyses of the nature of symbolism and his commentaries on Charles Peirce, cannot be overstated.

I am grateful to Mariane Ferme for her deep expertise of power structures and their relationship with communication and symbolism. Her critical analysis of the performance of language greatly strengthened several of these chapters.

I am grateful to Terry Regier for advising me on the proper mathematical description and interpretation of semantic information, including its conditional nature. I am especially thankful that he agreed to join this committee at short notice, and under odd circumstances.

I am grateful to Eve Sweetser and the Berkeley Gesture Group for their scholarship and discussion on extra-linguistic perspectives of human communication. The quality of their work was the inspiration for my initial investigations into semantics.

I am grateful to Slawomir Wacewicz, Przemyslaw Zywiczynski, Luke McCrohon, and the members of the Evolang community for encouraging my research into the intersection of semantics, words, and neural systems.

I am grateful to Ned Garrett, Kathleen Van Sickle, Tom Bottomley, and the Department staff for being reliable sources of advice, assistance, and humor.

I am grateful to Dav Clark, Jon Stiles, and the staff of the D-Lab for their mentorship and for their friendship.

I am grateful to James and Monica Little for our spirited debates on the nature and origin of human behavior.

I am grateful to the National Science Foundation, the National High Magnetic Field Lab, the Leakey Foundation, and the NVIDIA Corporation for their financial support.

Chapter 1

Historical Background

The modern study of human language and its semantic systems focus on a desire for efficient communication, usually squared within a cognitive system that is a rational optimizer of some set of life goals. While this is currently a popular view, ideas about the fundamental nature of human behavior and the role of language within that behavior have varied over historical time. In some ways, the mental models applied to the endeavor of understanding human behavioral distinctness have been strongly related to changes in technology and social organization.

Studying this history, and its effect on the practice of science, can create some understanding of the biases influencing research practice. This understanding, in turn, can serve as a seed to foster the precipitation of new models of human behavior and language, that react negatively to the prevailing metaphors of social and technological progress. So, we will start with a historical overview of some of the more popular perspectives on humanity.

1.1 Early Greek philosophy [c.1000BCE - 500BCE]

The early Greeks, in general, were not very interested in explaining the differences between humans and non-human animals. Instead, a lot of Greek thought was focused on explaining and instituting the differences between Greeks and barbarians, which were essentially anyone who wasn't Greek. This tribal view of human nature is easily seen in records of their wars. According to Thucydides, the Athenians killed all of the adult males and enslaved the women and children on the island of Menos after a successful siege [**gagarin˙woodruff˙1995**]. Gorgias writes:

> Triumphs over foreigners demand festive songs; but those over Greeks, laments.

In his account of the trial of Palamedes, Gorgias has the eponymous character point out that his execution will be particularly unjust, not because he is innocent, but because his accusers will be Greeks killing a Greek. The early Greeks also held barbarians as slaves, and according to Antiphon, it was regular practice to torture a slave before permitting his or her testimony in court [**gagarin˙woodruff˙1995**]. The courts themselves were divided into parts dealing with citizens, aliens, and aliens disputing with citizens [**aristotle˙ross˙1961**]. That Greeks regarded non-Greeks as something less than human can be seen in Herodotus's account of the death of Polycrates of Samos, where he writes:

> I discount Minos of Cnossus and anyone earlier than Minos who gained control of the sea; it remains the case that Polycrates was the first member of what we recognize as the **human** race to do so... [emphasis original]

It should be added that this attitude was not particular to the Greek peoples. In Herodotus's account of the Darius's invasion of Scythia, he recounts the stories of Scythians who were killed by their own people for adopting Greek customs [**herodotus˙waterfield˙dewald˙1998**].

Further evidence of a tribal outlook can be seen in the Greeks' definitions of justice (*diké*) and right (*themis*), which are concerned with how well an individual fits into the social order of his or her tribal group or city-state [**gagarin˙woodruff˙1995**]. Homer, for example, in his description of the inhuman Cyclopes, writes:

> They have no assemblies to give counsel, nor any rules of justice, but they swell on the heights of lofty hills in deep caverns, and each one makes his own law for his women and children, and does not care about the others.

Homer also describes a scene during the siege of Troy, where the soldiers have decided to flee back to their ships. Odysseus, one of the kings present at the siege, scolds them thusly:

> Sit still, my friend and hear the word of the others who are your betters, you unwarlike men, you weaklings you don't ever count for anything, in war or in council. There's no way all of us Achaeans can be kings here. Multiple leadership is no good; let there be one leader - one king

There is yet, however, one soldier who is not content, and taunts Agamemnon. Odysseus gives him an especially sharp rebuke, and the crowd responds with admiration:

> Truly, Odysseus has done thousands of noble deeds, taking the lead in giving good advice at the head of battle; but now this is much the best thing he has done among the Argives, to keep this word-shouting scoundrel from making speeches.

We see here the depiction of barbarians during the bronze age as those without social order. In the passages from the Iliad, Homer makes it very clear that Odysseus has acted justly in reminding the soldiers of their place within the social order of the Greeks. It is telling that the crowd praises him so strongly for this action, especially considering the many other "noble deeds" attributed to the character of Odysseus.

The early Greek view of the nature of humans can be seen in their creation myths, where mankind is given technical knowledge because they have no natural dispositions to help them survive in the world [**gagarin˙woodruff˙1995**]. Protagoras, in his account of the Prometheus myth, writes:

> While he was puzzling over this, Prometheus came to inspect the distribution, and saw the other animals cared for in every way, but human beings naked and unshod, coatless and unarmed. And already the appointed day was here on which the human species also had to come forth from the earth into the light. Now Prometheus has no other means for saving the human species, so he stole from Hephaestus and Athena their specialized knowledge and skill along with fire - for no one could acquire or use such knowledge without fire - and he endowed the human race with this.

It could be argued that, at least in the Western tradition, this is the beginning of the view of humans as tool-makers. Sophocles echoes this view in Antigone:

> Language and a mind as swift as the wind For making plans - These he has taught himself - And the character to live in cities under law. He's learned to take cover from a frost And escape sharp arrows of sleet. He has the means to handle every need, Never steps toward the future without the means. Except for Death: He's got himself no relief from that, Though he puts every mind to seeking cures For plagues that are hopeless. [**sophocles˙woodruff˙2001**]

Indeed, Anaximander argued that man must have originally been an animal of another species, as he never would have survived "as he is now" that is, both without natural tendencies facilitating survival and also without the technical knowledge that allows him to make tools [**burnet˙1963**].

At this point in time, the Greeks are familiar with the idea of an immortal soul[1], but equate it with the ability to move without apparent cause [**russell˙1967**]. That is, the difference between a thing with a soul and a thing without a soul is that the thing with a soul has the ability to move itself. Thales, according to Aristotle, argued that magnets possessed souls because they "move the iron" [**aristotle˙ross˙1961**]. Giving all animals the same kind of motive force meant that there must be something else to account for the difference in understanding between them. Parmenides first argued that the understanding of animals depended on the constitution of their bodies. Anaxagoras continued this line of reasoning to explain the uniqueness of humans:

> "There is a portion of everything in everything except Nous, and there are some things in which there is Nous also." In these words Anaxagoras laid down the distinction between animate and inanimate things. He tells us that it is the same Nous that "has power over", that is, sets in motion, all things that have life, both the greater and the smaller. The Nous in living creatures is the same in all and

[1] probably brought from Egypt via the Mycenaeans

> from this it followed that the different grades of intelligence we observe in the animal and vegetable worlds depend entirely on the structure of the body. The Nous was the same, but had more opportunities in one body than another. Man was the wisest of animals, not because he had a better sort of Nous, but because he had hands. [**burnet˙1963**]

As always, there are of course dissenting opinions. Antiphon, who was a legal adviser, defied the tribal outlook of his time in writing:

> We examine these attributes of nature that are necessarily in all men and are provided to the same degree, and in these respects none of us is distinguished as foreign or Greek. [**gagarin˙woodruff˙1995**]

Various early philosophers would define the soul as fire (because the body becomes cold upon death) or, like Diogenes of Apollonia, as air, because animals stop breathing when they expire [**burnet˙1963**]. However, while thinkers to the present day struggle to understand the composition of the soul, ideas like those of Diogenes appear not to have been carried on by serious philosophers past the time of the Greeks. However, the distinction of the soul as a motive force remained prevalent, and appears to be the earliest ancestor of modern conceptions of rational behavior.

1.2 Classical Greece [c.500BCE - 300BCE]

As we move into classic Greek philosophy, the force that motivates actions in humans and animals becomes divided according to different functions or appetites. In The Republic, Plato identifies two forces that motivate behavior: an appetitive soul, which moves the body to satisfy hunger, thirst, and sexual desire, and a rational soul, which allows humans to see “pure truth” [**plato˙griffith˙2000**]. It is the rational element which produces not only mathematics, but the reasoning required to create a just government, which is the topic addressed in The Republic. This rational element can be “destroyed and blinded” by the actions of the appetitive soul, rendering a human barbaric. It is a proper education, particularly in math and logic, which “purify and rekindle” the ability to think rationally. Thus, for the preservation of civilization, Plato argues that:

> . . . the overseers of our city must keep a firm grip on our system of education, protecting it above all else, and not allowing it to be destroyed accidentally.

Aristotle is often discussed in contrast to Plato; where Aristotle focuses on the origin of knowledge in perception, Plato locates it in the soul. This is the case here to a small extent. Aristotle argues in de Anima that there is nothing capable of self-generated motion, and that the movements of animals are produced in turn by their sensations [**aristotle˙ross˙1961**]. In his model, the most basic kind of soul, which is found in plants, is one concerned with acquiring nutrients, growing, and reproducing. Above that, we have the animal soul, which is capable of perception and turning sensations into imagination and movement. Last, there is the rational soul of humans. Thus, while beginning from different assumptions, Aristotle, like Plato, locates within the human soul the abilities of reason and morality. For example, in The Politics, Aristotle writes:

> For it is peculiar to human beings, in comparison to the other animals, that they alone have perception of what is good or bad, just or unjust, and the rest. [**aristotle˙ross˙1961**]

Following their belief that only civilized humans can see the nature of justice, both Plato and Aristotle consider human beings as inherently social. Much of Plato's writing about Socrates is involved with the nature of law and justice, and indeed Socrates gives his life to his belief in the law [**plato˙grube˙cooper˙2000**]. Aristotle writes:

> It is evident from these considerations, then, that a city-state is among the things that exist by nature, that a human being is by nature a political animal, and that anyone who is without a city-state, not by luck but by nature, is either a poor specimen or else superhuman. [**aristotle˙ross˙1961**]

This view continues up until the Enlightenment, when religious strife in Europe and reports of indigenous people in the Americas and elsewhere spark a belief in a primitive "state of nature" [**rutherford˙2006**].

1.3 Rome and Early Christianity [c.100CE – 300CE]

As Bertrand Russell notes, not much happens in philosophy between the death of Aristotle and the rise of the Roman Catholic Church [**russell˙1967**]. Historically, Alexander of Macedonia rises to power, conquering large portions of Asia. After him, the Roman Empire expands, at its zenith encompassing all of Greece and Asia Minor. During this time, Christianity rises in prominence, slowly at first, but in earnest during the third century. For our purposes, however, two important ideas come out of Rome.

The first idea is the rejection of tribalism for a nascent humanism. Epictetus writes:

> God is the father of men, and we are all brothers. We should not say "I am an Athenian" or "I am a Roman", but "I am a citizen of the universe."

This sentiment is echoed by Marcus Aurelius:

> But I know that the good is by nature beautiful and the bad ugly, and I know that these wrong-doers are by nature my brothers, not by blood or breeding, but by being similarly endowed with reason and sharing in the divine. [**aurelius˙hicks˙hicks˙2002**]

The philosophy of a human equality originates in Stoicism, but the idea likely arose more generally in the higher ranks of Roman society during the conquests of the army. That is to say, uniting disparate ethnic groups under one banner and to one purpose helped to create not only a sense of camaraderie, but feelings of the equality of men. Indeed, his experience in war would lead Aurelius to write of human society that

> Nature insists upon whatever benefits the whole. [**aurelius˙hicks˙hicks˙2002**]

It is possible that this began even earlier, with the conquests of Alexander of Macedonia. Likewise, in modern times, the foundations for the civil rights movement are often traced to the experiences of black Americans who fought in the Second World War.

The second important concept to emerge is the Stoic doctrine of the free will. The Stoics are the first to use the word "free" to describe volition against some sort of predisposed action, although the basic idea, that humans are free not to follow their irrational impulses, stretches at least back to Plato, who saw education as a means to revive the rational soul from slumber induced by grosser appetites. However, in the Stoic tradition, the freedom of the will takes on a character that had more to do with avoiding coercion or bad influences. Marcus Aurelius writes in his Meditations:

> None of them can harm me, for none can force me to do wrong against my will. . . [**aurelius˙hicks˙hicks˙2002**]

Russell speculates that this view may have been influenced by the social deterioration witnessed by the Stoics [**russell˙1967**]. That is to say, they could no longer point to the rule of law as an explanation for what made men good. We will see societal collapse as an explanation for changing philosophical views again in early modern philosophy.

Both of these are important as they form part of the Christian view of the nature of humanity. First, all mankind has been created as having a unified nature which holds them separate from the animals that populate the planet. Second, that because they are different, they are free to act in a way that animals are not. Note that this contrasts strongly with the earlier view that civilization and the rule of law was the important factor in determining who was human. The third principal component, that of divine possession (especially of kings and prophets), comes from Judaism. As Carrier writes:

> The Spirit of the Lord was expected to enter the body of Israel's human King, as well as the bodies of God's prophets [**carrier˙2009**].

In the Judaic understanding, this divine possession was not applied to all humans (or even all Jews), and indeed becomes a contentious issue in early Christianity concerning whether Jesus was divine, or divinely inspired. However, this becomes settled during the reign of Constantine and the establishment of Christianity as the religion of the Holy Roman Empire. During this time, the philosophy of the nature of mankind is written that will govern Western thought until early modern philosophy, and remains strongly influential, even if not acknowledged as such, today.

1.4 Catholic Philosophy [c.300CE - 1500CE]

The Catholic Church, like Plato, held that humans consist of two natures. This is not much changed from Plato, except that now the idea of sin (and not just irrationality) is attached to the body, while the spirit, which is a piece of the godhead, represents what is holy and good. Augustine of Hippo, a fourth century bishop living in Africa, remains the best representation of this idea. Briefly, in The Confessions, he writes about living a life satisfying the temptations of the flesh before coming to the Christian god and renouncing the sins of his former life [**augustine˙pinecoffin˙1967**]. To put it simply, he writes “my soul was contaminated by my flesh.” Thomas Aquinas, arguably the most important of the Catholic philosophers, puts it this way in the Summa Theologica:

> We now turn to man, a creature who is neither pure spirit nor pure body, but has a nature compounded of both. [**aquinas˙mcdermott˙1989**]

The rational mind, like god, exists outside the body, because it is only with a portion of god’s divinity in your soul that you can see past the images of things to their true natures. Augustine puts this succinctly:

> If we know only by the help of the spirit, that itself is the Spirit of God knowing in us

It follows, then, that sin is movement toward the body and its senses, where piety is movement toward a god-like sight or understanding. Again, quoting Augustine:

> I do not mean to beasts, of course, large or little, since they can see but not interrogate the beauty they have no adjudicating rationality, to assess what their senses report to them. Men, however, can by interrogation see God’s invisible things through the things which he has made…

For Aquinas, the human mind is a potential of understanding, because it is the furthest away from god’s own mind. Actual understanding arrives after a vigorous use of reasoned inquisition (ibid). He writes:

> The natures of things are known from their activities, and the activity marking human animals out from other animals is understanding…

Aquinas goes on to locate this source of understanding in a soul that is outside the body.

> The human soul, however, because it is a source of mental activity, must itself subsist, even though it is not in a body…Animal souls, however, because they do not act on their own sensation being an activity of body and soul - do not exist on their own.

And again.

> The rational soul's activity so far transcends the physical that it is not the activity of any bodily organ...

Augustine does not say much about the freedom of the will, except to point out that the freedom to choose means that one is free to choose sin over pious action:

> I bent my mind to understand what I had been told, that the reason for evil lay in the freedom of the will, that we could make evil choices... but I did not understand what this meant. [**augustine˙pinecoffin˙1967**]

Aquinas, however, treats the will extensively. For Aquinas, humans are free because their ability to use reason has liberated them from the billiard table cause-and-effect of sensation and instinct. This view predominated throughout the scholastic period and yet remains popular, so two illustrative passages from the Summa will be presented in full here.

> In other animals emotions follow instinctive judgment (sheep fear wolves because they instinctively judge them to be hostile); but in man a sort of calculation replaces instinctive judgment, making particular associations and connections. These particular connections are subject to the influence and control of general connections; we argue to particular conclusions from general premises. So reason can command the appetites of sense, both affective and aggressive, and control feeling.
>
> Things without awareness - stones and suchlike - act without judging; dumb animals judge instinctively, but not freely: sheep decide by nature, not by argument, to run away from the wolf; but men make up their own minds: in place of a natural repertoire of particular instincts they have a general capacity to reason, and since particular matters like what to do in this or that situation are not subject to conclusive arguments men are not determined to any once course. Because they reason they are free to make their own decisions.

To summarize, then, humans were different because their semi-divine nature allowed them understanding, and thus made them free.

Aquinas also mentions many older ideas concerning human uniqueness, probably acquired from Arabic commentaries on the writings of Aristotle [**russell˙1967**]. These are that humans are social, have hands, and can talk [**aquinas˙mcdermott˙1989**].

> ...man is by nature a social animal, and people living a social life need some single authority to look to their common good.

> Man's upright carriage also releases his hands for various useful purposes. And since he does not have to use his mouth for gathering food, it is not oblong and hard as in other animals but adapted for speech, the special work of reason.

One of these older ideas seems to come directly from the Prometheus myth:

> . . . so although nature can't endow him with the fixed instinctive responses, defense mechanisms and protective covering which it gives to other animals of limited awareness and powers, still it gives him reason and hands - those tools of tools - with which to make his own tools and suit every sort of purpose.

1.5 Early Modern Philosophy [c.1500CE - 1800CE]

Modern philosophy, which characterized itself as a revolutionary break with Catholic philosophy, is substantively different (for our purposes) in one regard: Hobbes, Kant, Rousseau, Vico, and even Locke see humans as inherently asocial creatures who exist peacefully in society only by the rule of law and the punishments meted out when they are broken [**cutrofello˙2005**].

> And the same are the BONDS, by which men are bound, and obliged: bonds, that have their strength, not from their own nature, for nothing is more easily broken than a man's word, but from fear of some evil consequence upon the rupture [**hobbes˙1965**].

As Rutherford explains, natural social order was assumed to be part of god's eternal law governing how everything, including humans, behaved (see above) [**rutherford˙2006**]. However, religious turmoil in the sixteenth century cast serious doubt on the idea that a social order could exist in principle, let alone in practice. Thus, if it was not natural for humans to live in society, they could only be compelled to behave in a civilized manner through fear of injury.

> It is only in society that it occurs to him to be, not merely a man, but a refined man after his kind. [**kant˙bernard˙1951**]

In the Enlightenment, this presumed primordial condition of humans was referred to as the state of nature, where freedom was absolute and all acts were permissible.

> Robberies, murders, rapes are the sports of men set at liberty from punishment and censure [**locke˙1978**].

In a civil state, certain freedoms were given up voluntarily by citizens to secure protection under the law[2]. This conception is quite removed from the ancient Greek view, which is that civilized people create laws and not that laws create civilized people.

Like the Scholastics before them, the moderns see humans as metaphysically free; however, this freedom stops being attributed to godlike nature. For if there is no divine law capable of ordering society for the common good, then the divine portions of humans would be unable to deduce the source of order in the universe (if those divine portions even exist).In other words, if reason is a portion of god's mind knowing god's work, reason disappears along when the deity does. So human nature becomes described in terms of freedom from the physical laws that govern the motions of machines. Although this view is commonly associated with Ren Descartes [**descartes˙voss˙1989**],

> . . . there remains nothing in us that we should attribute to our soul but our thoughts, which are principally of two genera the first, namely, are the actions of the soul; the others are its passions. The ones I call its actions are all of our volitions, because we find by experience that they come directly from our soul and seem to depend only on it. . .

who notably advocated for vivisection under the belief that automata like animals could feel no pain, it really was an idea common to the time. Interestingly, the common metaphor was to compare animals to machines.

> The case is not so much different in brutes but that anyone may hence see what makes an animal and continues it the same. Something we have like this in machines and may serve to illustrate it. For example, what is a watch? [**locke˙1978**].

As Burnet demonstrates for Greek philosophy and Russell notes more generally, it is common for philosophers to seize upon technological innovations as a means to understand the nature of other things around them.

If the key human trait is freedom from physical necessity, then rationality and a contemplative life is born from that same freedom. Rousseau writes:

> In any animal I see nothing but an ingenious machine to which nature has given senses in order for it to renew its strength and to protect itself, to a certain point, from all that tends to destroy or disturb it. I am aware of precisely the same things in the human machine, with the difference that nature alone does everything in the operations of an animal, whereas man contributes, as a free agent, to his own operations. The former chooses or rejects by instinct and the latter by an act of freedom [**rousseau˙cress˙miller˙1992**].

[2] note that this is the philosophy underpinning the Constitution of the United States

Like earlier philosophers, the moderns were interested in speech, but for the first time language is used as a potential cause of human rationality. In Catholic philosophy, language was “the work of reason”, not its cause [**aquinas˙mcdermott˙1989**]. Even during the Renaissance, the idea that natural language was causally related to formal logic was “ridiculed” [**losonsky˙2006**]. The more Catholic philosophers tended to see language as either independent from reason (like Descartes), or as a means of reducing natural laws to something capable of being manipulated by natural thought, such as Leibniz.

> For ideas are in God from eternity, and they are in us too, before we actually think of them, as I showed in our first discussions. If anyone wants to take ideas to be men’s actual thoughts, he may; but he will be gratuitously going against accepted ways of speaking [**leibniz˙1981**].

For John Locke, language can lead to the abstraction of general principles because, although language is a product of thought, the association between concepts and the mental symbols used to represent them in thought is arbitrary. It is necessary to use language in this fashion, as there are no innate truths hidden within the soul, which is a “blank slate”. He writes:

> For, first, it is evident that all children and idiots have not the least apprehension or thought of them. And the want of that is enough to destroy that universal assent which must needs be the necessary concomitant of all innate truths. . . [**locke˙1978**].

Thomas Hobbes takes this a step further, arguing that language is required for any kind of rational thought, and that:

> Children therefore are not endued with reason at all, till they have attained the use of speech. . . [**hobbes˙1965**]

Vico applies this link between rationality and language to the study of the entire behavioral evolution of early human civilizations [**vico˙1999**]. To Vico, the substitution of religious order for barbarism was tied to the substitution of sacred iconography for mutism. In turn, these were replaced by a government modeled after belief in paternal archetypes specifically (aristocracy) and a language which functioned on the idea of archetypes more generally (categories). The final stage of Vico’s behavioral evolution involves the transition to poetic modes of thought, where communication relies on the use of metaphor.

It is frequently the case that there are several philosophers who contradict the common beliefs of their time. However, it is worth briefly mentioning that Francois-Marie Arouet, better known as Voltaire, argued against nearly all of the concepts discussed above in his Philosophical Dictionary, ridiculing the idea of a human kind of free will,

> Such are the wretched sophisms of the wretched sophists who taught you. Here you're upset because you're as free as your dog! Come! don't you resemble your dog in a thousand ways? Don't you have hunger, thirst, wakefulness, sleep, the five senses in common with him? Would you like smell otherwise than with the nose? Why do you want to have a free will different from his?

innate knowledge,

> If a man born without his five senses could live, he would be without any idea.

and the asocial human being

> We live in society, so there is no true good for us but what is good for society [**arouet˙besterman˙1971**].

1.6 Contemporary approaches to human uniqueness

In 1859, Charles Darwin published On the Origin of Species, which outlined a model of life where every extant species had descended with only slight modifications from its forebears. Properly understood, this meant that any uniquely human quality must be a slight modification of that found in animals. Darwin himself appeared to recognize this when he wrote in The Descent of Man that "the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind" [**darwin˙1998**]. It might be thought that a shift in the understanding of human origins of such magnitude would by necessity motivate a reappraisal of long-held assumptions on the division between human and animal; however, this appears not to have been the case, even for Darwin himself. In the same book, Descent of Man, he later remarks that,

> It may be freely admitted that no animal is self-conscious.

and

> I fully subscribe to the judgment of those writers who maintain that of all the differences between man and the lower animals, the moral sense or conscience is by far the most important.

His closest supporters immediately missed the deeper implications of the theory. Thomas Huxley famously failed to grasp that evolution is not progressive in nature, and Alfred Wallace, the co-discoverer of natural selection and a great admirer of the older Darwin, could not bring himself to apply selection to human mental abilities, writing toward the end of his life that:

> . . . the moral and higher intellectual nature of man is as unique a phenomenon as was conscious life on its first appearance in the world, and the one is almost as difficult to conceive as originating by any law of evolution as the other [**richards˙1987**].

Moral behavior remained a problem for all three, who still saw the primitive humans as living in a brutish state of nature where fierce competition with each other and the forces of nature necessarily precluded cooperative behavior. Rational behavior and intelligence was also a serious stumbling block, as it appeared people required neither of these traits to survive by hunting and gathering. In an attempt to overcome these obstacles, they began inventing new kinds of selection; to wit, Darwin wrote an entire book about sexual selection and the evolution of humans, and in Huxley's Romanes lecture in 1893, he attributed ethical behavior to group-level selection [**richards˙1987**].

It was not until the late twentieth century that a clear understanding of evolutionary processes emerged. Predictably, attempts to apply this understanding to human behavior have been met with hostility; most people find the notion that their behavior can be explained by rules upsetting. What might be surprising is the content of the arguments themselves, which invoke rational thought, indeterminate will, and ethical behavior. These could have been written by Descartes or Rousseau, and would have been comprehensible to Aquinas. In a very deep sense, these beliefs in human behavioral discontinuity have continued to influence scientific research on what humans are.

Chomsky and the indeterminate will

Noam Chomsky's work in the behavioral sciences could, for instance, be taken as an extended argument in favor of a free will. Chomsky, notably, rose to fame with a critique of Skinner's behaviorist model, arguing that a human's behavior was far more complex than the examples they had available to learn from. Furthermore, he argues that most children's behaviors are not strictly reinforced in the way that an animal's behaviors are during training, but are explored and practiced without the intervention of adults. Thus, there must be some "inner knowledge" that already exists in their brains somewhere, which is only "tuned" or "set" , and not determined, by the environment.

This first idea is often raised in scholarly debates around first language acquisition in children. The general notion here is that a child is never exposed to every possible correct employment of a grammar, and therefore would be unable to deduce the rules of that grammar if they were truly naïve learners.

Chomsky's explanation from the early 2000s for how children are able to build rich inferences about language in the presence of lean stimuli is a generalized cognitive adaptation. Over the course of his career, the characterization of this generalized ability has changed, but has included mathematical reasoning and the ability to think recursively. While it might seem like Chomsky's mental trait is primarily concerned with reason, the rationale behind pointing to these traits is their supposed end result: a human with generative behavior. By generative, we mean that it may follow some rules or be under some constraints, but is free to vary so widely that it is effectively unconstrained. Recursion, for example, is used by Chomsky to argue that the possibilities of human language are infinite, because it allows the embedding of relative clauses. So

> I went to the store.

can be expanded to

> She said that I went to the store.

can be expanded to

> He thinks that she said that I went to the store.

ad infinitum.

Chomsky's more recent work is explicit in this comparison with computers, where the fundamental property of language is a BINARY OPERATOR named *Merge* that constructs hierarchical trees out of linear signals. Chomsky and Berwick occasionally refer to this property as language's "CPU " [**chomsky˙berwick˙2016**], and reiterate that it is responsible for

> the ability to construct a digitally infinite array of hierarchically structured expressions with deterministic interpretations. . .

(pg. 110).

Notably, Chomsky did not originally posit an evolutionary predecessor for this trait, nor a strong selection pressure that would have quickly driven such an ability to fixation in a population. This kind of mutation is sometimes referred to derisively as a HOPEFUL MONSTER, a sort of deus ex machina that is very unlikely to appear, let alone persist. It more or less arose out of nothing, just because it could, and therefore is not constrained by its evolutionary history.

Likewise, because this cognitive trait was not in service to nor driven by elements of the environment, one could say also that it was free from outside influence. In recent work responding to criticisms from the Evolang community, Chomsky has taken pains to explain that this change was an accidental, "slight rewiring of the brain", and that biologists and evolutionists have entirely underestimated the likelihood that GENETIC DRIFT will drive a small genetic mutation to fixation in a population.

More recently, these kinds of speculative histories commonly posit origin by exaptation. That is to say, the cognitive abilities underlying language were originally beneficial for one reason, and later were found to be useful for another reason, more or less by accident, and without the direct intervention of the environment. Possible original functions for the cognitive abilities underlying language have included navigation and tool use [**hauser˙chomsky˙fitch˙2002**, **bickerton˙2009**].

Tomasello and the ethical human

A different perspective comes from researchers like Michael Tomasello, who argue that it is the unique social adaptations of humans that explains their unusual behavior [**tomasello˙2006**]. The principle argument here is that an intention-based[3] communication system, joint attention, and social norms are really only useful in the context of cooperative relationships. An uncooperative individual would be unlikely, for example, to do anything besides ask for favors. Jointly held attention is only helpful in context where two or more individuals have the same goal, and must understand each others' intentions toward achieving that goal. A social norm can only be maintained with strategies like third-party punishments, which are altruistic in that they generate no benefit for the cooperator.

Tomasello, here sounding very pre-modern, begins by positing that nonhuman apes are not cooperative animals – that is, they frequently engage in mutualistic but not altruistic actions. When chimpanzees hunt in the wild, for example, they do not willingly share food with other chimpanzees, even those who helped during the hunt. In laboratory settings, chimpanzees will sometimes offer help to a human experimenter, but never when there is a possibility of getting food for themselves. What is perhaps even more telling is that chimpanzees do not appear to engage in social games with human experimenters the same way that a human child would.

Based on this evidence, Tomasello argues that there must have been some pressing need in human evolutionary history to engage in what he calls "collaborative" activities, or behaviors with are altruistic and require joint attention. This was perhaps something like building boats, teaching children, or hunting animals. Other scientists, such as Robin Dunbar, go further to argue that the benefit of collaborative activities created a unique need to evaluate and ensure the cooperative potential of others [**mesoudi˙whiten˙dunbar˙2006**].

At this point, humans began to build cognitive infrastructure for representing the knowledge states of conspecifics. To be sure, this is something the nonhuman apes do already, although perhaps to a lesser extent. Apes, for example, seem to understand the difference between humans who are trying and failing, and humans who are not trying at all. However, human children seem to see behaviors also in terms of roles, or with a "birds eye" view of the interaction. Thus, they are able not only to guess at the goals and knowledge states of others, but also how those related to the roles those others are playing in the context of some activity. Tomasello argues that this kind of knowledge is necessary for communication, as much of our language assumes a very large base of shared knowledge (see section 2.3).

The next step according to this view is that, once collaborative activities are commonplace due to the ease of sharing intention, human societies start creating social norms to reinforce altruistic behavior. Specifically, these norms for moral behavior act as a disincentive for any

[3] We say **intentional** here because there are many signals in nature that are not intentional, such a moth's eye-like wing patterns or a peacock's large tail feathers. Intentional communication is divorced from material circumstances in a way that many signals are not, and therefore have no guarantee of being honest or stable. If you know that a signal isn't honest, your only cue as to whether to pay attention to it comes from the relationship between you and the source of the signal.

individual reverting to the ancestral condition of selfish behavior, and thus acts as a sort of evolutionary ratchet during the evolution of collaborative activities. These social norms could be enforced by something like gossip – conversations about the moral character of a non-present party[4] – which requires some form of communication.

Pinker and the reasoning brain

The final modern view that we'll consider here is that humans are rational optimizers. Proponents of this view argue that language is a tool evolved specifically for communication and used by a rational brain.

As a species, humans can be characterized by their reliance on learned knowledge, especially knowledge by pedagogical intervention, as an adaptive strategy [**pinker˙1997**, **boyd˙richerson˙henrich˙2011**]. Given this environment, the anatomical changes necessary for language could have evolved through something like the Baldwin effect, where the limited communicative abilities that all hominids possess were used extensively to pass on knowledge about the environment, and thus came under strong selection pressure to become more accurate and/or efficient. The construction of complex tools is so often assumed to be the knowledge in question that the origin of human language is often placed between the lower (choppers and flakes) and middle Paleolithic tool technologies (bifaces and hafting), at about 300,000 years before the present.

One piece of evidence in favor of the evolution of language as a means of facilitating the transmission of cultural knowledge is the reliance of mathematical reasoning on language. Concepts like countable numbers and linear relationships seem to depend on first learning language, and then using that knowledge to bootstrap mathematical rules that are difficult to understand visually [**carey˙2009**, **spaepen˙et˙al˙2011**].

One may also argue that language is fairly accurate in reproducing mental representations of the environment. The famous Berlin and Kay color studies, for example, showed that words for colors in any language generally cluster around hues which elicit peak retinal responses [**berlin˙kay˙1969**].

There is a large amount of evidence in favor of language being designed well for efficient communication. Human cochleae, for example, are the most sensitive at the common frequencies of human speech[5]. Language has a dedicated neural pathway in the brain, that is at least partially separated from the analysis of other kinds of sounds [**hickok˙poeppel˙2007**]. Finally, language has many redundant features, so that meaning can still be conveyed properly in noisy environments, or when the person listening isn't paying very close attention (see chapter 6).

Pinker, specifically, has also argued that when language appears to be inefficient, that variety of inefficiency is still a rational way of communicating. He points to euphemism – substituting a colloquial phrase for a word or group of words – as an example of some-

[4] more commonly called "gossip"

[5] Or, human speech tends to occupy frequencies that are easy to hear, depending on your point of view.

thing that appears to be inefficient communication. [**pinker˙nowak˙lee˙2008**] However, euphemism is generally employed to hide words or intentions that are socially inappropriate or even dangerous. Pinker points to the attempt to bribe a police officer in the movie *Fargo* as an example, where the speaker wants to avoid explicitly offering money to forgo a traffic violation. Speaking in euphemism, then, gives the speaker some amount of plausible deniability should unpleasant consequences (like being arrested for attempted bribery) arise.

Chapter 2

Theoretical Background

2.1 The role of metaphors and analogy in science

It is a generally held truth that humans are not naturally given to abstract thought. Thus, Pascal Boyer explains that supernatural entities are only rarely more than one abstraction away from their natural counterparts, as a larger conceptual distance would make them hard to reason about[1]. Likewise George Lakoff has catalogued an impressive number of real experiences that humans have, like the passage of time, that are nevertheless abstract enough to warrant discussion only through metaphors of more concrete experience, like traveling along a path [**lakoff˙johnson˙2003**].

Scientists and philosophers likewise use metaphors and analogies to think about the world; the lived experience of any people constrains (or expands, depending on your frame of reference) the hypothesis space in which they are able to generate models of the universe. As a specific example, we have seen Rousseau compare an animal to a mechanical object like a clock1.5. More generally, we saw this influence Stoicism as well as early modern philosophy: to wit, the loss of social order undermined the belief that humans were inherently a social animal, or even that sociality was necessarily desirable. As Bertrand Russell explains in his History of Western Philosophy:

> The change in values is connected with a change in the social system. The warrior, the gentleman, the plutocrat, the dictator each has his own standard of the good and the true. The gentleman has had a long inning in philosophical theory, because he is associated with the Greek genius, because the virtue of contemplation acquired theological endorsement, and because the ideal of disinterested truth dignified the academic life... Modern definitions of truth such as those of pragmatism and instrumentalism, which are practical, rather than contemplative, are inspired by industrialism as opposed to aristocracy [**russell˙1967**].

[1] E.g. ghosts are humans who are usually invisible, but are not also inaudible, nor free from human concerns, nor suddenly a black hole with a circling galaxy, etc. [**boyer˙2001**]

Vico puts it briefly as:

> The order of ideas must follow the order of institutions [**vico˙1999**].

In contemporary times, the machine analogy for biological processes has been largely replaced with the network analogy, and the unmoved-mover model for freedom of the will has been replaced by the corporate model. For example, compare:

- *the mitochondrion is the power house of the cell*; with,
- *the Krebs cycle is one node in the catabolic system*; and,
- *The frontal lobes are the effector lobes*; with,
- *Frontal lobe activity is correlated with executive functions* .

It should be no surprise, then, that while thinkers of the early modern period struggled to find the source of pneumatic power that could make muscles move[2], modern thinkers struggle to find just which part of your brain is the CEO.

This highlights a broader point – like models, metaphors are never correct, but every once in a while are useful. Their utility depends on the amount of correspondence between the internal processes of the two things being compared. To take a simple example, life might indeed be like a box of chocolates in that is full of variety and surprise; but, no one would make the mistake of modeling a person's life as a box of chocolates, as lives are not sold in grocery stores, eaten bit by bit in bounded chunks, nor given from one person to another as a romantic cue.

Whether or not these internal processes are listed explicitly in the comparison, they still exist in the thinker's mind and therefore serve as intuitions about how the thing being compared will behave. To take a real example, public health officials are often quoted as saying something along the lines of *your brain is like a muscle*, to encourage people to **exercise** it. Typically, the first question the people then ask is *so when I think a lot, does my brain get bigger?*, followed by, *and how many more calories is it burning?*. Both of these intuitions, that brains grow with use like muscles, and that brains consume more energy when used, are largely fallacious. It has also led to the modern industry of smartphone applications for brain training, based on the misconception that solving puzzles are like workouts that will make your brain smarter at everything. There has yet to be any evidence that these apps actually work.

While most modern scientists do not think about human brains and behaviors in terms of muscular processes, they still use metaphors filled with hidden assumptions, albeit more modern ones.

[2] Eventually settling on the pineal gland, because it is an unpaired structure

Human brains as computers

Probably the most common metaphor currently in use about the brain is that it is a computer. While *computer* really only means a thing that makes calculations, the modern use of the word is something closer to a von Neumann machine. So, scientists talk about **working** memory (RAM) versus **long-term** memory (disk i/o), serial versus parallel processing, and the total computing power, measured in flops, of the human cortex.

This metaphor correctly captures some aspects of brains. Both computers and brains are designed to transform information. Both encode features of their environment – like keyboard presses or visual signals – into a native information format. Both can output the results of their computations into physical formats.

This has led to some interesting consequences, specifically based on the intuition that, like computers, human brains have deterministic outputs. If you started the Java runtime engine on a computer, you would be very surprised if it opened up a JavaScript environment instead. Humans, however, make these kinds of malapropisms all the time (the name conflict between Java and JavaScript being the source of more than a few). The attempted explanation for these phenomena is the competence/performance distinction, where it is posited that somewhere in the human brain is a deterministic computer, and that something happens between the computer and the behavior to generate what would otherwise be an appropriate response.

However, unlike computers, which are electronic abstractions of von Neuman machines that process data which only exist as binary electronic states, the hardware of brains is is also the software. There is no abstraction between the physical system and its functions, and therefore any modifications to the structure of a brain are also modifications to its function. Furthermore, there is no memory organ in the brain, nor is there a single center of control for processing data or allocating resources. In this context, the idea of a unary decision maker seems somewhat silly.

Human language as a tool

Another common metaphor in use is the comparison of human language to a tool. Typically, these scenarios posit rational actors who have a pressing need to communicate, and so develop a shared code in order to transmit information effectively. Implicit in this metaphor are intuitions about the intentions behind the use of language and the structure of language as a medium for transmitting information.

This matches a lot of the naïve beliefs that people have about language. Prima facie, a language is a way to convert meanings into sounds. People typically use this mapping system to give ideas or concepts that only exist in their heads to other people. Furthermore, when infants learn their first language, they seem to be very preoccupied with sharing the language around what they are doing with anyone nearby.

One consequence of adopting this point of view comes from studies of co-speech gestures, or the kinds of gesticulations you make with your body while you are talking. These are timed along with vocalizations, carry non-redundant information, and are performed **even when the other person cannot see you**. A common example of this phenomenon in modern times is someone who gesticulates at an interlocutor who is talking to them on the phone [**mcneill˙2005**]. This obviously cannot be for the benefit of the person listening, but to reinforce the belief that such behaviors are still for the benefit of communication, researchers have pointed out that stopping the speaker from gesticulating reduces the fluency of speech. This raises the question as to why these two things would have been wired together neurologically in the first place, especially if the evolutionary motivation for humans to start vocalizing was non visible or distant conspecifics, or because the hands were busy working or carrying things [**bickerton˙2009**].

Unlike tools, however, languages are not built, but grow out of social exploration and consensus. While some features of language are tightly constrained by the environment, like the serialized nature of information transfer, we don't see widespread standardization across languages the same way that we see widespread standardization across hammers, radios, or data transfer protocols. The main features of languages, like its lexicon and syntax, change with time and with populations of speakers, but this does not reduce their utility.

In this metaphor, tools are also used with intention. Language, however, is often produced without communicative intention, as in the case of self-directed speech. Even language that is ostensibly communicative in intent can show markers of unintentional production, like the use of jargon in inappropriate contexts or the incorrect assumption of shared knowledge. Along these lines, Wallace writes that one of the hardest things for a writer to do is to learn how to communicate with the explicit intention of being understood by another person [**garner˙wallace˙montgomery˙2013**].

A subtler point, though perhaps more telling, is that the language itself tends to drift toward the accidental. White notes that all expressions, no matter how vivid, eventually become tropes [**white˙1978**]. The writer Orwell, in his essay **Politics and the English Language** rails against the use of stale metaphors which no longer carry their original meaning [**orwell˙1946**]. In modern semantics, the idea that the meanings of terms lose their specificity over time is well-established.

In each case, intuitions produced by the use of metaphors like these combine with other intuitions from other metaphors, which result in conjectures about human nature that can be questionable or even demonstrably incorrect.

2.2 Humans are not generative

One of these ideas, first mentioned in subsection 1.6, is that human behavior is infinitely generative. In Chomsky's case, this is easily follows from the brains-are-computers metaphor, but this same conclusion has resulted from approaches from other directions [**chomsky˙1991**]. For example, Munroe recently estimated the number of possible English tweets to be about $2e46$, a number close enough to infinity that the entire population of humans would not be able to read them all were they to be given the entire length of time the planet Earth has existed [**munroe˙2013**]. The absurd magnitude of this number should serve as a first clue that the model which generated it is incorrect (see chapter 5).

Additionally, some evidence points against humans using a set of grammatical rules to generate speech. Extemporaneous speech tends to both ignore the native grammar, and to principally consist of short, familiar structures [**bygate˙1988**, **hymes˙1970**]. Pawley and Syder have called this speaking strategy `clause chaining`[3], where speakers combine chunks of words – typically dependent clauses – into phrases that are loosely linked, if they are even explicitly joined at all [**pawley˙syder˙1983**, **brown˙yule˙1983**]. In contrast, written language, which is pondered and edited, is much more likely to use complex noun phrases, chained prepositional phrases, and embedded relative clauses [**chafe˙1982**].

In practice, native speakers of a language use a vanishingly small subset of all the possible combinations afforded by the grammar of that language. This is, in fact, part of what makes a native speaker sound native to other speakers. So, for example, while:

> It is desired by me to have married you

is perfectly grammatical English, it also marks the speaker as someone who is not familiar enough with the language to produce the typical form:

> Marry me

Deviations from this subset at best sound odd and at worst are not interpretable, even when those deviations do a better job of following grammar rules than the natively acceptable form of the utterance [**pawley˙syder˙1983**].

Not only do native speakers of a language not necessarily follow a grammar, but the idea of a grammar itself is problematic. Grammars and other prescriptive rules about human language are typically based on written language. The kind of language that makes it into the expensive act of publication is not only heavily revised, but also typically only reflects one of any number of possible codes of any given language. The code that makes it into writing is also usually the code used by high status members of society, and thus the given grammar for that language does not include the kinds of constructions that speakers, especially lower class speakers, use in daily life. To take one example, grammars of German only include structures found in high German, the code from the north eastern part of the country, and not low (Bavarian) German, Swiss German, etc [**wardaugh˙2010**].

There is a less restrictive use of *grammar*, meaning the regularities that are typically found in language, as opposed to hard-coded and generalizable rules. It is of course obvious that language have regularities, like the *-s* ending in English to indicate plurality. However, the existence of these regularities does not necessarily indicate a generative process following rules. In fact, the most commonly used expressions in a language are the least likely to follow regular grammatical rules, in the loose sense. The *to be* verbs, for example, are some of the most intransigently irregular forms in every language. This alone might seem to be a

[3] It is explicitly this form of creating language, “phrases tacked together like the sections of a prefabricated hen-house”, that Orwell argues against [**orwell˙1946**].

coincidence; however, in laboratory experiments, humans learning a highly irregular language spontaneously regularize the less common expressions [**chater˙christiansen˙2010**]. This suggests that regularization is something that languages get for free when they are learned by non-computers who have limitations on things like memory and recursive thinking. Perhaps one of the reasons for the evolution of language is how non-computer-like our brains are.

However, it is not only irregular forms of utterance that are stored in memory instead of being regularized. Studies looking at the speed of lexical access indicate that **regularized** and common word forms are also memorized and not generated in real time by rules [**baayen˙et˙al˙2002**, **ullman˙1999**].

In fact, part of sounding like a native speaker is correctly marshalling a set of memorized chunks of speech, or what are sometimes called `speech formulae` [**wray˙1998**]. These are common phrases that are produced and used as whole, unbroken phrases, such as

> I was going to say...

or,

> ...spend the rest of your life with ...

even when they do not follow the native grammar, as the second examples often does not.[4] This native-sounding or fluency effect is not only true for grammatical constructions, but also accents (using the correct phoneme set) and expletives (like *uh* or *um*), suggesting the same cognitive system is driving syntax-y and non-syntax-y speech [**bosker˙et˙al˙2013**]. More to the point, native speakers are "at their most hesitant" when they cannot use speech formulas, indicating that much of language production, in practice, is based on memory recall and not generative rules [**pawley˙syder˙1983**].

Further evidence for this view comes from studies of second language acquisition. It has been widely observed that second language learners are more accurate in their comprehension of a second language than they are in their production of it [**wells˙1985**]. What is probably happening is that during comprehension, listeners combine their limited knowledge of lexical items with a strong prior prediction about what the speaker is intending to communicate. However, speech production, which requires whole chunks of memorized speech patterns and not just vocabulary, takes longer.

[4] As in the common construction

> He's who I want to spend the rest of my life with

or, in this example from a recently published New York Times bestseller, which seems to have been written this way until the situation was half-corrected by an editor

> So that's how my dad decided on whom he was going to spend the rest of his life with [**ansari˙klinenberg˙2015**]

The correct phrasing here would be

> So that's how my dad decided with whom he was going to spend the rest of his life

It could yet be true that human brains are generative, but also employ a less computationally expensive procedure for common phrases [**karimi˙ferreira˙2016**]. There is some evidence that complex sentences go through a quick and then a slower process of comprehension [**kahneman˙2011**]. However, sentences with very uncommon syntax, like

> The horse raced past the barn fell

can be difficult for native speakers to parse **even when explicitly informed how**. In these situations, listeners often give up on trying to parse the sentence entirely, and instead rely on situational cues [**ferreira˙engelhardt˙jones˙2009**].

2.3 Communication is a problematic pressure

Anomalies in the study of the relationship between language and information were noted in the beginnings of analytical philosophy. In a series of lectures given at Oxford, Austin struggles in the attempt to map natural language to propositional logic [**austin˙1962**]. One of the major stumbling blocks he encounters are what he calls "performative", now known as LOCUTIONARY, utterances [**searle˙1969**]. These are vocal expressions which themselves cannot be considered to have veracity, as they are more like actions than statements. Austin also noticed that several aspects of utterance only exist to negotiate concurrent social environments, prepending the warning "(a shocker this)". So it's not clear quite how to explain language from the point of view of thought.

In the opposite direction, it is also troublesome to think about thought or reason in the absence of language [**jackendoff˙1996**, **slobin˙1996**, **butler˙2011**]. To be sure, all humans have things like sensory recollection that are not directly encodable into words, but reasoning, in the way that people mean thinking through problems in terms of learned strategies, is a strategy created by linguistic acquisition, if not a linguistic strategy itself. For example, number and linear progression might seem to be based on something other than language, but these concepts are not culturally universal [**everett˙2005**]. It has been argued that learning the linear relationship, $n = n + 1$, between the countable numbers is dependent on the linguistic strategy of memorizing the sequence of countable numbers first [**carey˙2009**].

Whatever the relationship between cognitive functions and language, it seems clear that languages do not directly encode concepts, in the platonic sense [**evans˙levinson˙2009**]. As mentioned above, not all cultures have countable number; Piraha has no documented use of number terms beyond `one`, `a few`, and `a lot` [**everett˙2005**]. Generally speaking, languages typically do not encode the full set of logical relationships: Guugu Ymithirr does not encode the conditional `if`, and standard American English does not encode `exclusive or`. Conversely, languages can also be atypical in the information they **do** include. Speakers of some North American Languages habitually include the source of knowledge in their statements, with markers indicating whether they were a personal witness, for example [**mithun˙1999**]. One of the best documented instances of linguistic diversity is in the encoding of spatial direction, where some cultures use absolute cues (i.e. cardinal directions), while others may use any of a variety of relative cues (e.g. left/right, uphill/downhill) [**evans˙2003**].

Instead, someone learning a language must figure out what concepts are talked about, and how. Slobin call this "thinking for speaking", and points out that it dramatically complicates the inductive problem that new language learners have to solve [**slobin˙2006**]. This is not a superficial problem, like deciding which object a term refers to, or even which feature of that object. For example, speakers learning Navajo have to learn to modify verbs based on characteristics of that verb's object, the name of which is encoded by another term entirely.

Even if we imagine some form of non-linguistic mentalese, information transfer as an explanation for the origin of language suffers from NETWORK EFFECTS. For example, it is assumed that early humans would have already wanted to communicate with each other in ways that were not possible with the varieties of facial and vocal signaling that were already available. To begin with, it's not clear how they would have conceived of something like language in order to then desire one, having not witnessed a language before. Even if one human were to stumble on it, they would then face the problem of how to teach other early humans to speak, without being able to explain how or why. If we are assuming that the emergence of language was selected for by evolution, the situation becomes even worse, as there is no benefit of being the only person knowing a communication protocol. Even if we were to imagine that all of Language Eve's children were able to learn language, the risk of death or drift driving the language mutation out of existence are high enough to make its persistence unlikely [**niederhut˙2014˙2**].

The evolutionary landscape doesn't improve once language has gone to fixation in a population. If language was explicitly designed to be used between conspecifics, it seems unlikely that it would also be used for self-directed speech. Computers, after all, do not translate their internal system calls into HTTP. Left alone for a long enough time, most humans will start vocalizing this self-directed speech out loud, even though there is no one to hear it. In a similar vein, people talking on the phone regularly generate manual gestures to accompany their speech, even though the person they are speaking with cannot possibly see them [**mcneill˙1992**].

Theories about the origin of language sometimes try to get around the problems of network effects by positing `protolanguages`, `proto-protolanguages`, et cetera, where some combination of deictic gestures and pantomime permitted the gradual development of language faculties, which then transitioned from the visuomanual modality into the vocal modality [**hewes˙1973**, **arbib˙liebal˙pika˙2008**]. However, spoken language has not replaced other forms of communication, but is copresented along with them [**kendon˙2004**]. This corepresentation is not some afterthought, but seems to be integrated tightly with language production. Complicated gestures, for example, are set up in advance so that the visual aspect can precisely co-occur with the auditory chunk to which it relates [**mcneill˙2005**]. Furthermore, forcing speakers to abandon gesturing, perhaps by holding onto a chair back or by by keeping their hands in their pockets, disrupts everything from fluency to lexical retrieval.

A non-communicative origin of language avoids these difficulties, and it remains possible that languages became modified later on to be used primarily for the encoding and transmission of information, as exaptation is one of the more common tools in the evolutionary process. However, the extent to which human languages are primarily used for, or well-designed for, communication is open to debate. A common counterexample is the use of language in ritualized interactions, the meaning of which is not derived from the particular words used. Greetings, for example, involve standard linguistic expressions, which are often coexpressed with gestures of affiliation or submission [**wray˙1998**]. The linguistic content of these greetings, which in American can also involve questions which are not answered as a matter of course, does not carry additional information beyond the fact that it is a greeting.

In a related vein, information that is related to the linguistic elements of speech might not be caused by those elements. For example, a specific word can indicate things beyond its dictionary definition, like the social identity of the speaker, or the speaker's role with respect to the listener, which are typically not encoded by denotation. A simple example of this is the spatial dependency of word choice, made popular by a recent New York Times article, where using a word like *bubbler* instead of *water fountain* is encoding a regional identity in a way that is probably unintended by the speaker [**dialect˙quiz**].

However, it is still true that both words would be referring to the same object, unambiguously. It might be assumed, then, that while language is sometimes richer in information than intended, that the linguistic strategy of words and syntax is still transmitting useful information, and transmitting it well. However, the semantic content of speech does not map one-to-one onto the lexical items in the sentence [**chafe˙1968**]. For example, the sentence

> I never said she meant to do that

could be interpreted to mean

- Someone other than me said she meant to do that
- I am offended that you believe that I said she meant to do that
- I only implied that she meant to do that
- I said that someone else meant to do that
- I said she did something accidentally
- I said she meant to do something else

and every combination of those interpretations. While this might seem like a silly example, it is widely recognized that communication is never specific enough as-is, but requires a very rich set of prior knowledge about the speaker and the situation under discussion.

To complicate this situation further, the information contained in the linguistic elements of utterance is convolved with the non-linguistic elements. In other words, to correctly understand the information contained in language, the listener has to understand the interaction between things like syntax and things like prosody. The same sentence can mean wildly different things depending on who is speaking, who they are talking to, and how the sentence is performed [5].

Sarcasm, for example, is speech that is meant to be understood to convey the opposite of the semantic values expressed. This kind of language is marked by atypical word choices, like the frequent use of superlative modifiers [**bamman˙smith˙2015**]. More broadly, any kind of verbal irony in human language is marked by speech patterns that imply the language does not originate from the person who is actually talking [**tobin˙israel˙2012**]. For example, John Aubrey, in his 1681 biography of John Milton, observed that Milton often delineated speech meant to be heard as sarcastic by overpronouncing the letter *R*.

Another instance of this same phenomenon is the change in denotation between dialects. So, for instance, *significant* and *independent* mean very different things in statistical contexts (one could say, in the statistical dialect) than they do in standard American English. This often causes frustration among college students when reading literature from academic dialects, where sentences typically use old words to mean new things. The sentences themselves, however, like

> . . . a broad synthetic strategy that organizes and leverages the latent potentials of these complexities [**designing˙cities**]

are uninterpretable if you are reading them as if they are standard American English.

This dependence of word meaning on things like dialect and speech register points to an understanding of denotation as something other than a property inherent to a word. It would seem that the information value of a word depends more on the contexts in which it is used, like discussions about statistics versus discussions about your *significant* other, who is an *independent* person. This would also imply that word meanings can change over time, which is in fact something we observe.

One example of this is semantic bleaching, a phenomenon where a word's meaning becomes less intense over time. *Awesome* for example, meaning something that inspires awe, used to only appear in discussions of the Judeo-Christian deity, but now is used for everything from socks to Sriracha. Words don't only change in intensity, however. The phrase from the formal logic dialect, *begs the question*, used to mean a premise which only explains its conclusion when the conclusion is already true. In contemporary times, however, it is mostly used to mean something like it leads me to ask, as in

> Which begs the question, why would two hardened Kaos agents risk. . . **the carbs** [**get˙smart**]?

Another example is the growth of euphemism and language policing around linguistic slurs. Most people understand slander to be harmful words that are applied to a particular population, but the truth is more that the word is harmful because it has been used to describe that population, and not the other way around (see discussion in [**nunberg˙2012**] about other kinds of expletives). So, when *retarded* was introduced as a kinder euphemism

[5] Note that we do not mean in the Austin sense

for the mentally challenged, it grew into a slur because it referred to that population. The same thing has since happened with the word *special*, about the same population, and for the same reason.

Living languages, or languages that are still used by a speech community for communicative purposes, change their sets of lexical items along with the denotations assigned to those lexical items constantly. It should be obvious that the constant introduction of new words into a language's vocabulary degrades communicative efficiency, at least until each new word reaches saturation within a community of speakers. It is no accident that scientific terms which are required to be unambiguous, like unique identifiers in the classification of living things, are derived from **dead** languages.

What might be less obvious is that part of the intrinsic motivation behind language change is the construction and maintenance of social groupings. In other words, parts of speech communities will invent new ways of speaking in order to distance themselves from surrounding social groups [**wardaugh˙2010**]. The growth of Verlan in France, which began as a marker for immigrant communities living in the poor suburbs around urban centers, is one example of this; although, the creation of slang, and its eventual adoption by young members of the upper class is a universal phenomenon. Importantly, these dialects are often not mutually intelligible, so the social distancing is, in real life, creating a barrier to the ability of the language to transfer information accurately and efficiently.

2.4 Humans reference systems are autapomorphic

The views of language that we've been discussing so far have assumed that reference, or how things come to have meaning, is the easy question. The ethical human view of language evolution argues that apes would be communicating like us too, if only they were motivated to cooperate with one another in the way that we are. Likewise, the tool user view of language evolution posits that humans already had a rich inner life of concepts, and only needed a means by which to pair concepts to sounds.

Before humans, animals were already capable of some kinds of reference, including what Charles Peirce called indexical reference, or a pairing between a symbol and a referent [**farago˙et˙al˙2010**]. This kind of reference was demonstrated famously by Pavlov, who trained dogs to salivate to the sound of a bell, with the reference here being that *bell* → `food` [**deacon˙1997**]. More recent examples come from the vervet monkey alarm call literature, where there are separate sounds that map onto separate meanings, i.e.

- *sound 1* → `bird-like predator`
- *sound 2* → `cat-like predator`
- *sound 3* → `snake-like predator`

These sound-to-meaning pairs in vervets include whole suites of behavior, appropriate to the type of predator being avoided [**cheney˙seyfarth˙1982**].

An important thing to note here is that the sounds used by vervets have no obvious relationship with their referents, e.g. the snake alarm call does not sound like snakes. This kind of abstract relationship between signs and referents, sometimes called `the arbitrariness of the sign` is often pointed to as a defining feature of human communication, e.g. *snake* does not mean `snake` because it looks like a snake [**armstrong˙wilcox˙2007**]. Locke, for example, used this same arbitrary relationship to rebut the medieval notion that natural language couldn't be used for logic (see section 1.5). Tomasello and other gesture-first theorists explicitly include a step where communication systems based on pantomime become fully fledged human languages once they have become arbitrary.

According to a series of studies conducted on trained dolphins, their indexical reference system is capable of understanding a reference to an object that is no longer present [**herman˙richards˙wolz˙1984**]. Dolphins at least, then, are also capable of having their references displaced in space or time, which has also been pointed to as a unique feature of human languages [**bickerton˙2009**].

The association that needs to be made here is just one of proximity – either spatial or temporal – so the actual computation is fairly simple [**deacon˙1997**]. Really, it is just a kind of coincidence detection, where if one event reliably predicts another, it becomes the sign for that referent. This logic is roughly the same way that synapse strengthening works in animal brains [**bi˙poo˙1998**]. If we relax the requirements on the referent being a **meaning** of some sort, this same feat can be performed by two receptors innervating a single neuron.

If we look in a smaller set of animals, just the nonhuman mammals, there is evidence of deictic reference. So, for example, dogs and wolves are both capable of understanding that a human pointing a finger is an attempt to indicate a referent by demonstration[6] [**udell˙spencer˙dorey˙wynne˙2012**]. Deixis as a reference system, e.g. *that one*, has also been hypothesized to be a unique attribute of human language [**nunez˙2012**]. Chimpanzees not only understand pointing, but seem to employ iconic references in their natural communications with each other. So, for example, to indicate `I want you to come over here` a chimpanzee might pantomime the act of grabbing her friend's hand, even though she is out of reach [**tomasello˙2006**].

Human reference systems, on the other hand, work by what Peirce called symbolism. That is, a symbol points to other symbols, and it is a collection of symbols that together point to some feature of the environment [**deacon˙1997**]. Abstract terms like `freedom`, for example, cannot be learned by something like Pavlovian training as there is no way to point to a freedom the same way that one can point to a snake. Instead, humans learn associations between utterances including *freedom* and the environmental situations that they point to, like it being the best thing about being American, requiring sacrifice, etc. In Butler's terms, no signifier (word) acquires its reference by mimesis (index), but rather is productive (symbolic) in that the meaning is constructed before it is applied to the material (significant) [**butler˙2011**].

What is particularly interesting here is that symbolic reference is the primary system of

[6] Words like THESE and THERE perform a similar function in spoken language

reference used by humans when interacting with other humans. When given words whose referent has been established indexically, like in the vervet alarm calls, or iconically, like onomatopoeia, humans habitually **ignore** the additional information and treat it like a symbol anyway. For example, some native speakers of English will have a minor epiphany upon realizing that *cufflinks* are so called because they `link` your shirt sleeve `cuffs` together. To take another example, when children are taught visually iconic gestures, they produce them as whole symbolic units of utterance before they understand that they are motivated by visual imagery [**cartmill˙beilock˙goldinmeadow˙2012**].

Someone taking the tool view of language might say that the distinction between symbols and indices doesn't matter, because symbolic reference was just one of the abilities that evolved to allow information transfer. However, this is misguided because symbolic reference systems are not a solution to the problem of communication. In a world where early hominins already had the ability to learn arbitrary associations between signs and referents, what additional benefit would convolving the meaning of those signs with other signs provide? We could imagine, using the speculative history from Bickerton, an early human about to communicate that there was a dead animal nearby to scavenge, but first thinking to herself

> Wait a minute . . . I already have a food call that all members of my species understand, but I had better also invent a proto-gestural-language with ambiguous meanings first.

Indeed, Tomasello points out that apes already used hand signals for requests like this, such as food begging. He goes on to add that human language is not typically motivated by requests or demands, but often happens just because the people engaged in speaking enjoy sharing language with each other [**tomasello˙2006**]. This points to a different view of the motivation behind language, one that isn't dependent on immediate communicative needs, but seems to be tied up with aspects of sociality and identity. The question remains as to what this selection pressure could have been, and how it relates to the emergence of symbolic reference in humans.

Chapter 3

Language as Phonatory Culture

The Reader's Digest in 1958 printed the following joke:

> A more frightened than injured young Seabee electrician was brought into the hospital suffering from electrical burns. Shortly afterward his instructor, a chief electrician, arrived. "Why on earth didn't you turn off the main power switch before you tried to splice the wires?" asked the chief.
>
> "I wanted to save time, chief, and I've seen you stand on one leg, grab the wires and splice without turning off the power."
>
> "My God, kid," exclaimed the chief. "Didn't you know I have a wooden leg?"

It's a remarkable feature of human behavior that we seem do a lot of things just because it's what we've seen other people do. It might not seem remarkable to a human that this kind of behavior is the norm, but other animals do not share this predilection, including the nonhuman apes. This tendency to `imitate` actions like standing on one leg, as opposed to `emulating` the goals of those actions, has been hypothesized to be one of the major contributors to cultural accumulation in humans.

The accumulation of cultural behaviors, in turn, has been hypothesized to be the key adaptation for human evolutionary success [**boyd˙richerson˙henrich˙2011**]. The cultural accumulation of knowledge has allowed humans, who have no biological adaptations to most of the ecosystems on the planet, to spread across the entire globe. It has also allowed us to mobilize food sources that are poisonous or hard to digest. Furthermore, it is unlikely that a human, no matter how intelligent, could discover for herself all of the environmentally specific behaviors to live in a new place before dying of some combination of starvation and exposure.

We are calling these cultural behaviors – as opposed to cultural knowledge – because human behaviors are often performed without knowledge of why. Henrich and Henrich give the example of Fijian food taboos, which ostensibly serve to prevent pregnant women from accidentally ingesting marine toxins [**henrich˙henrich˙2011**]. When asked to explain their taboos, the women provide conflicting explanations, possibly confabulated on the spot.

Boyer encountered a similar difficulty during interviews about religious beliefs, where he was repeatedly told in some form or another that the explanation didn't matter: things **just were** that way.

So a culture is a set of behaviors that are learned socially, and usually implicitly. Those behaviors are convolved with a rich context of social relationships and situational information, like whether or not you are currently pregnant. The rationale, if there is one, behind any culturally acquired behavior are typically not described explicitly, and links between behaviors and contexts are arbitrary. Finally, the behaviors themselves are practiced before they are performed, and are sometimes performed just because it's fun to share them.

These statements should sound like many of the features of language that we touched on in the theoretical background (see chapter 2). It's interesting that a lot of research is devoted to studying aspects of language, like its acquisition, when so little is known about how humans acquire any other aspect of their cultural heritage. It seems likely that the progression would be similar. This chapter will expand on the idea that human language is explicable as culturally acquired, phonatory behavior, first by dwelling on the similarities between language and culture, and then by outlining a theatrical metaphor for linguistic behavior.

In this chapter, human language will be discussed under the name *phonatory culture* in an attempt to avoid immediately summoning prior intuitions about how and why language is special when compared to other human behavioral systems. Here, we are taking an overtly continential view of language, following authors like Butler who emphasize that human language is, strictly speaking, a muscular action, although one with "specific linguistic consequences" [**butler˙1999**]. In defense of this view, there have been several recent papers providing evidence that aspects of language acquisition and production are better explained by general social learning mechanisms than they are by hypothetical language-specific systems [**frank˙lewis˙macdonald˙2016**, **yurovsky˙wagner˙barner˙frank˙2016**]. In a similar vein, some hypothesized **language-specific** mechanisms like regularized grammar have been shown to be byproducts of pre-existing mental processing, like the fact that human memory has limits [**chater˙christiansen˙2010**]

3.1 Language is more social than necessary

By definition, all communication systems are social in that they are designed to be used by more than one entity. Human language is **extra social** in that social aspects of human life appear in their communication practices in ways that go beyond the application of communication for existing requirements. For example, Dunbar's finding that stories about social interactions have more accurate recall than non-social stories with similar features is social, because it facilitates social interactions [**dunbar˙1998**]. It is not extra social however, because the structure of the language has not changed because of these social tendencies.

One example of how language is extra social in a way that matches cultural behavior but not the design of communication protocols is bias during acquisition. When more than one variant of a single behavior is presented to a cultural learner, she will prefer the behavior used by the individual who is more successful, in something called `prestige bias` [**richerson˙boyd˙2006**]. If you were designing an individual to learn arbitrary communication protocols, a simple solution would be to adopt the most frequently observed protocol, and not the one used by, for example, the fanciest-looking computer. But, in cultural acquisition, such biases make sense. To become successful, it makes sense to imitate those around you who seem to be more successful than others; and, apparent success is a fairly honest signal of actual success.

In a similar manner, this cultural learner would also be more likely to adopt the behaviors of her social group. Certainly, biasing a learner to prefer adopting protocols by personal similarity would immediately lead to fragmentation and defeat the purpose of having a communication system in the first place. Again, this makes sense from a cultural point of view. Members of your social group are more likely to be demonstrating responses to the kinds of situations that you are more likely to come across. Additionally, social groups of humans often use similarity in behavior as an honest signal of belongingness.

So, for example, upper class New Yorkers speak in a different manner than lower class New Yorkers [**labov˙2001**]. Things like word choice and phoneme production are interpreted as signals of status, whether this is conscious or not. When speaking to individuals of another social class in an affiliative setting, people `accommodate`, or incorporate elements from, the other linguistic code. This tends to be unidirectional, with lower class speakers adopting higher class language, possibly reflecting a conscious desire to appear conciliatory, and/or an unconscious propensity to imitate high prestige behaviors.

One interesting consequence of this is political arguments around language use. The French, famously, have the Acadèmie Française dedicated to deciding what words and phrases should be used by authentic French people. During the founding of America, Noah Webster rewrote the rules of spelling, as a way to distinguish American English from the language of its colonizer, Great Britain. Similar examples can be drawn from around the world. China, for instance, has implemented a single writing system for all of the languages within its borders as a way to unify the country, and calls each of them dialects. However, the dialects (Mandarin, Cantonese, Wu, etc.) are not mutually intelligible when spoken aloud. When Turkey modernized in the twentieth century, the government adopted the roman character set (as opposed to Arabic or Cyrillic, both of which were used in Turkey) as a way of aligning themselves with the western hemisphere. After Kenyan independence, Jomo Kenyatta chose Swahili as the official language of the country, even though it had almost no native speakers and fewer competent speakers than English, as a way to both distance itself from its colonial history and the sectarian attitudes of the ethnic groups in the country.

Even in the same language, how the features of that language are used depends on the social role of the speaker. Research in this area typically focuses on age and gender (as they contribute greatly to societal roles), but we can imagine that these differences are also found across other roles. Depending on the language (and the language community), women may use more facilitated speech, more polite speech, more direct speech, explicitly masculine speech, or even employ silence as a communicative practice more often than males in the speech community. Younger members of any speech community tend to create new words and syntax patterns more often than older members. Very old members, however, tend to adopt the speech patterns of the women of that community, no matter what that pattern is.

It isn't only your own social role that matters when deciding how to employ your language, but also your social role with respect to the person to whom you are speaking. The use of `polite` speech, for example, depends both on how well you know your interlocutor and the power difference that exists between you. Most of the romance languages incorporate some kind of formal versus informal address in pronouns and verb conjugation, but this is likely a superficial change dependent on a deeper social practice, as they also tend to include other modifiers like

- use of the future tense (*would you like* instead of *do you want*)
- use of indirect requests (*it would be great if* instead of *I want*)
- use of politeness modifiers (e.g. *please*).

In Javanese, for example, there are six levels of politeness marking, and not just two or three [**wardaugh˙2010**]. When a government official is speaking to a member of the middle class, they will use the most formal register. When they are speaking to a member of the lower class, like a servant, they will drop two or three registers lower, to a modest register. The choice of words, politeness modifiers, and verb conjugations aren't the only things that change from register to register either. Polite speech, both in Java and elsewhere, is also marked by changes in timbre, prosody, and speech rate, suggesting that the difference between polite and impolite speech is more of a holistic and multimodal construction.

3.2 Language is a performance

The politeness modifiers are only one example of extra-linguistic behaviors being modified concomitantly with the linguistic ones – utterance, as a rule, incorporates movement of the body. However, it also routinely involves work done between individuals. This kind of work ranges from spatial changes between interlocutors[1] to active efforts to help the other person portray a role they wish to fill [**goffman˙1982**].

This particular organizing metaphor began as a restructuring of theory around theater and dance, but has since been expanded into a way of thinking about all kinds of behavior[2]. Particularly in human interactions, performance theory gives us a way to think about the embedding of extra-sociality in behavior, by casting all kinds of everyday behavior as a performance.

Let's illustrate this with a quick example. Goffman relates a story of workers performing as *service staff* [**goffman˙1959**]. When among the seated diners, the service staff help each

[1] E.g. if only of them is sitting down

[2] Not even necessarily human. Burke writes that performance reaches from "ritualization in animal behavior (including humans) through performances in everyday life" [**burke˙1966**]

other maintain an outward facing image of polite servience. When away from the customers, in the kitchen or the back of house, they help each other maintain a different role, typically of a disgruntled employee. The important thing to note here is that at no point is anyone **not playing** at being something or other, and that the choice of role depends on time, place, and which other people are around.

The performative nature of human behavior does not only apply to outward facing actions, but to internal conceptions as well. Butler argues that gender identity, and even the personal belief in a unified identity, are themselves the result of internalizing performances of gender norms and interactions [3] [**butler˙1999**]. The norms themselves are produced by the uncritical reproduction of observed behaviors, in a way that is both meaningful as an action and meaningful as an interpreted action [**butler˙2011**]. The performance of femininity by a man in drag, then, is fundamentally similar to the performance of femininity by woman, in the sense that both are imitations of a cultural suite of behaviors which carry with them the consequences inherent to all symbols.

These kinds of roles can be seen in language behavior as well. When speaking to someone employing different speech patterns, it is common for a speaker to change their own linguistic behavior. If they are trying to build an affiliative relationship, the speaker will produce CONVERGENCE behavior, like matching speech rate, tone, and even dialect. In Labov's famous studies of New York department stores, for example, salespeople converge on the speech patterns of their presumed buyers [**labov˙2001**]. However, since the salespeople, who typically come from a lower socio-economic class with a slightly different dialect, do not have direct access to the speech patterns they are adopting, they tend to exaggerate the most obvious differences.

Conversely, speakers who wish to showcase their differences are very likely to adopt DIVERGENCE behavior. One example of this is modern movements to establish the legal legitimacy of languages in colonized areas, like Celtic in Ireland or Cherokee in the United States. A more everyday example is speech patterns associated with stable power imbalances. Teachers have a kind of lecturing voice used in class that is distinct from the way the students talk. This is typically true even when the teacher herself would otherwise be speaking in the same speech register as the students.

Another way to put this is to say that speech behavior is based on the role we would like our interlocutor to think we are playing, or the desire to be interpreted as a particular kind of body [**lepage˙tabouretkeller˙1985**, **butler˙2011**]. This doesn't need to only be *I am different than you*, but can also accommodate more complicated social dimensions like kinship. Notably, this kind of social information is not incorporated into structural models of language behavior [**wuthnow˙1987**].

An interesting consequence of this is the source of social norms surrounding taboo words. These are words whose use is discouraged not because the meaning of the word itself is offensive, but because its use is indicative of belonging to an unappreciated social group.

[3] The identity here resulting from the need to convey who is the self and who is the other inside a dialiectic situation

The *n-word* fell out of use among middle and upper-class white people in America because it was used heavily by poor white people [**nunberg˙2013**]. Using the word would have made it seem like they were adopting a role of a poor person, which was of course not their intention. Lower class white people, who tend to converge on the speech patterns of the upper classes, eventually dropped the word from use as well.

To take another example, the *a-word* was considered transgressive because it was favored by American GIs, whose behavior as a whole was a bit rougher upon returning from the war [**nunberg˙2012**]. This gave them a sort of authenticity, and their speech patterns became used by members of the lower class in instances where they adopted divergent speech patterns.

3.3 Language can be ritualized

Like all performances, the production of language can become ritualized. This exists in an obvious way, e.g. in religious rituals with prescribed speech acts, but also in less noticeable ways. There are stereotyped language behaviors for all kinds of daily minutiae, like greeting friends or ordering food.

It's possible that this kind of ritualization grows out of the way new speech patterns are negotiated among communities. As Schechner points out, people do not typically spend their days constantly doing brand new things [**schechner˙1977**]. They practice language structures, and repeat only the ones that elicit the desired reaction. So, the most common parts of anyone's day are repeating behaviors that have been practiced often and found to reliably produce a desired outcome.

Another possibility is that stereotyped speech patterns are a way to facilitate the acquisition of language by new learners [**wongfillmore˙1976**]. Generally speaking, stereotyped sequences are much easier to learn and reproduce than the rules that underlie combinatorial systems. Additionally, there is evidence that adults, when interacting with young children, scaffold their learning by combining ritualized utterance with other ritualized motor behaviors in appropriate contexts.

In either case, the end result is that language competent speakers produce language that is filled with chunks of stereotyped speech patterns (see section 2.2). These tend to be invariant in both their linguistic forms and their tonal contours, both of which are expressed fluidly[4] [**peters˙1983**, **weinert˙1995**]. They are also regularly associated with other stereotyped actions, including co-speech gestures.

[4] Text-to-speech engines are trained on individual words and not whole constructs, which is part of why they sound funny

Chapter 4

A Performance Model of Human Behavior

We consider here that language is simply a phonatory type of cultural behavior, and that language production and comprehension is a cultural kind of performance, in an explicitly theatrical sense. As an organizing metaphor like `language is a tool`, performance theory is useful in that it easily produces complex behavior from a lean set of assumptions about the cognitive abilities underlying those behaviors.

4.1 Minimum Assumptions of the model

The assumptions necessary for a performance model are as follows.

1. Learn gregariously

   If language is indeed phonatory culture, and culture is learned from conspecifics, the obvious requirement is that early humans must have been capable of social learning. A first barrier to the accumulation of cultural traditions in nonhuman animals is that most social learning happens within the context of the mother-infant relationship [**huffman˙mahallage˙leca˙2008**, **sargeant˙mann˙2009**]. Modern humans have expanded their sources of social learning to include juvenile peers and non-related adults, so there must have been a point in history where humans became more tolerant of conspecifics [**vanschaik˙deaner˙merrill˙1999**]. It's possible that the kinds of traits that are referred to as pedomorphic in humans evolved precisely because juvenile appearance is a way to facilitate tolerance by adults.

   Many authors posit an additional requirement for the ability to map observed behavior onto actions performed by their own bodies, possible also including predictive elements like FORWARD STATE ESTIMATION [**arbib˙2005**, **anderson˙cui˙2009**]. The latter is not likely to be required, given something like C-induction, and the former is a plesiomorphic trait for humans [**roberts˙et˙al˙2012**, **wolpert˙ghahramani˙jordan˙1995**].

2. Assume intention

   The next requirement is a propensity to assume agentic, or goal-directed, behavior. It might not be necessary to be able to infer what that goal is, just that a person performing the behavior hopes to achieve some outcome which may or may not be observable [1]. Because you are conspecifics in the same habitat, it is likely that you will have that goal at some time in your life as well, and so the complexity of learning any particular behavior can be deconvolved from concurrently learning the reason for the behavior. This strategy, where the indexical referencing is essentially ignored, has been called C-INDUCTION [**chater˙christiansen˙2010**].

   Some authors have argued that there is a further requirement to be able to model the somatic states of conspecifics [**iacoboni˙2009**]. The reasoning is typically that two individuals are capable of sharing indexical references to goals only if they are also capable of acknowledging shared internal states like hunger and sorrow. While it is possible that modern humans do do this, this is not strictly a necessary condition, as the same shared reference requirement can be satisfies with the much simpler LOOKS LIKE ME $\propto$ IS LIKE ME rule [**carey˙2009**].

3. Attend to action chains

   Finally, the performance itself must be the focus of attention. That is to say, the learner would need to focus on the links between the individual behaviors that are linked together in a performance, and not on the outcome itself. To misappropriate a common idiom, early humans had to learn to look at the finger pointing at the moon. This is a critical change, as the primacy of relationships between actions is fundamentally the same as the primacy of relationships between symbols – outside of the system of actions, any individual component is non informative. Importantly, attending to action sequences instead of rewards is something that nonhuman primates struggle to learn [**tomasello˙2006**].

4.2 The performance model is lean

Commonly, hypotheses about the nature and origin of language involve very fat assumptions about the state of the world or the state of human nature. For instance, humans appear to believe that any language they hear is both factual and germane [**grice˙1989**]. However, using language as an HONEST SIGNAL is not an evolutionarily `stable strategy`, because there is no causal mechanism tying reference to belief [**axelrod˙1984**]. For example, the author can write that *airplanes are mammals* regardless of his own beliefs about the matter. Other authors have attempted to explain the existence of Grices maxims by some combination of arguing that humans evolved to be obligate altruists ane/or that there are strong social punishments for unhelpful people [**mesoudi˙whiten˙dunbar˙2006**]. Both of these

[1] This is essentially the FUNDAMENTAL ATTRIBUTION ERROR.

are bold claims to make, especially in the light of what we know about human behavior specifically, and animal behavior more generally [**cluttonbrock˙2011**]. To put this another way, it is obvious to researchers who study living things besides themselves that the claim that humans are uniquely altruistic is an extraordinary one with only ordinary evidence.

Neither of these explanations is necessary if the listener is not assuming honesty or helpfulness, but instead simply that the interlocutor is not altering her normal behavior. Imagine for instance, that you are trying to learn to tie your shoes, so you wait until another person stoops to tie her laces and you watch. You would indeed be very surprised if she tied her own shoes incorrectly just to fool any potential observers. Language taken as performative behavior only requires agentic action, and does not require that performance be done with altruistic intention.

It is typical for speculative histories of language origins to point to one of two alternative hypotheses. One, that there was no communicative behavior until the appearance of language, after which everyone wanted to communicate; or two, that people were already communicating, just not vocally. The first hypothesis begs the question as to how communication could undergo positive selection when there was no one else with whom to communicate. The second hypothesis is more tenable, but is left explaining why any animal would abandon a perfectly good communication system (usually manual) to switch to a system already used for eating and breathing. However, performance of any behavior, whether it be something like nut cracking or grooming, typically involves elements like posture and vocalizations already. A communicative system exapted from phonatory behaviors has no need to bootstrap itself past network effects nor explain how behaviors are coordinated [**niederhut˙2014˙2**].

Finally, the performative model has no need to explain how indexical reference evolved into a symbolic system. The system itself starts with symbolic reference built in, after which the associations to concrete phenomena can be created using the same kind of mutual information criteria believed to underlie indexical reference. Likewise, performances are composed of motor subunits that can be extended linearly or nested hierarchically, and so there is also no need to explain the evolution of grammatical systems of reasoning. In support of this view, there have been no evolved “language areas” of the brain that are new in humans, only normal motor planning areas which seem to be doing something spectacular [**jurgens˙2002**, **striedter˙2005**]. Evidence from language acquisition studies supports the view that domain general learning rules, like the inference of hierarchical structure, are sufficient to explain the how humans learn words and syntactic structure [**perfors˙tenenbaum˙regier˙2011**, **yurovsky˙wagner˙barner˙frank˙2016**].

4.3 The performance model covers edge cases

Besides requiring fewer initial assumptions, a performance model easily handles phenomena that are difficult to explain in other language models. For example, models that characterize language as a tool for efficient communication or the side-effect of rational thought require some sort of special pleading to describe linguistic fossils[2]. Efficient information transfer would predict the rapid elimination of fossils and other formulae, whereas an internal grammar can parse fossils, but parsing them provides no additional information as to the intended meaning. On the other hand, behavioral fossils are generally common.

It could be argued that, from an information cost point of view, linguistic fossils are not significant for selection processes. In other words, if we are trying to minimize the number of syllables per unit of information generally, fossilized phrases occur too rarely to be selected against as a significant cost. However, part of the cost of information transfer is storing hard rules for edge cases that do not follow the convention. This is less of a problem for unconventional components of speech like irregular verb conjugations as these tend to be used frequently and thus are easily learned. It is also not a problem for a view where behavioral wholes are imitated, as this would be a normal part of the process.

Indeed, first language learning in humans seems to begin with the repetition of exactly these kinds of formulaic utterances, and irregular verb forms are produced earlier than regular verb forms [**wray˙1998**, **devilliers˙devilliers˙1973**]. Children continue to use formulaic speech as a learning strategy in a way that adults do not, which may explain why second language learners have such difficulty sounding like native speakers [**weinert˙1995**]. There are reasons to believe that these kinds of ritualized utterances are really the basis of language, and that the rules governing grammar are generalized every time a speaker learns their first language [**langacker˙1987**, **chater˙christiansen˙2010**].

A performance model also strongly predicts the use of self-directed speech, or language behaviors that are not directed to any recipient at all, and therefore cannot be communicative[3]. This can be thought of as rehearsing behavior, where a performance is practiced in private before it is used in front of others [**schechner˙1977**]. The practice itself may be related to the suite of activities undertaken by the human brain when not actively engaged in another task, known as the default mode but more commonly referred to as daydreaming [**raichle˙snyder˙2007**].

We can in addition point to code-switching as a bad information transfer practice, but a good performance one, because performance depends not only on semantic content but also on identity and belongingness. Aggramatical structures, and especially the relationship between aggramatism and frequency, are a sore point for a grammar-based model, but are predicted by a performance model. Finally, the many-to-many mappings between utterances and possible meanings is fully expected in a performance model, where the associations of that performance depend on what else is happening, who is listening, and why.

[2] These are phrases like *eke out a living. A living* is the only thing one can *eke*, and thus the whole phrase essentially acts as a single word

[3] Which, by definition, are not predicted by an information transfer model

4.4 The performance model makes testable predictions

The performance model, by reframing the question of what language is and can be, makes predictions which are testably different from prior models. A simple prediction that the *brains are computers* model makes is that human language is countably infinite. The performance model, in contrary, predicts that the total size of any language is limited by the size of behavioral repertoires of its speakers. Observing relatively small bounds on the size of the semantic space of a language would be evidence in favor of the performance view.

In phonatory culture, the meaning of any cultural behavior is tied to non-linguistic aspects of its performance, like audience, time, place, and other kinds of context. A behavior that is performed across many contexts provides poor indexical reference to any one, and so the number of inferences afforded by a behavior should decrease the more often it occurs[4]. Concomitantly, the magnitude of information associated with a word should decrease the more often it is used. Conversely, from an efficiency point of view, you would want to most common terms used to also be the most meaningful, as this would reduce the amount of effort needed to communicate an idea successfully.

The performance model also takes a strong stance on the idea of composability. The *language is a tool* view posits that language exists to convey meaning, which in turn is done by combining language elements together. However, the meaning behind cultural behaviors are not so easily decomposed into individual motor behaviors. Observing that the semantic value of an utterance cannot be reproduced by simply combining the semantic value of its constituent words would be evidence in favor of the performance view.

[4] One is reminded of **The boy who cried wolf**

Chapter 5

The Size of the Language State Space

It is an often repeated assertion that human languages are unbounded in their generativity. A recent search on Google Scholar[1] for example, returned over 2000 results for academic papers containing the exact phrase “infinite number of sentences”. One imagines that the total number of papers containing a similar sentiment is substantially larger. The claim of infinite generativity relies on the metaphor that a human brain operates much like a computer. Chomsky, for example, writes

> language, is, at its core, a system that is both digital and infinite

, and works by

> the infinite use of finite means.

[**chomsky˙1991**]

One form of reasoning behind this claim relies on some combination of an open system of reference tied to a recursive syntax. In a fully recursive language, there are no bounds on how long sentences can become, or how deeply their elements can be nested [**hauser˙chomsky˙fitch˙2002**]. In this case, the total number of possible sentences able to be constructed is discrete but also unlimited, and therefore infinite. An open system of reference means that it is always possible to add new symbols that refer to referents or even other symbols [**deacon˙1997**].

Another form is to keep the underlying mechanism as a black box, but to still assume that humans are using language to disambiguate signals in their environment. Randall Munroe, author of the popular comic strip xkcd, recently used Claude Shannon’s estimate of the entropy of English characters to estimate the total number of possible English tweets (see Eq. 5.1) [**shannon˙1948**]. He arrives at the conclusion that there are $2.3 * 10^{46}$, a number large enough[2] to be considered infinite in practicality, even if not in actuality [**munroe˙2013**].

[1] scholar.google.com

[2] For scale, the number of seconds in a human life is roughly $3 * 10^9$, and the total number of seconds that have passed since the birth of the universe is roughly $5 * 10^{17}$

To be sure, this kind of random process is not used to model language in practice, but does provide a reasonable estimate for the size of language under a generative model, especially given the lack of alternative predictions to test.

$$N = 2^{1.1 * n_{characters}} \tag{5.1}$$

The infinite generativity argument is often raised against children learning to speak just by example. The reasoning here is that any size exposure to an infinite system comprises 0% of its total variance, and it is therefore impossible to make correct inferences about the rules governing the creation of that language. In other words, because the amount of global variation is so large compared to the possible size of local exposure, humans must have some sort of prior knowledge about the structure of the language to bootstrap their acquisition. Evidence that young learners hear a larger fraction of their first language would remove some of the motivations behind innate language arguments.

However, both of these approaches assume that human language samples its semantic possibility space rectangularly, when we know that humans make preferential use of some combinations of their language [**zipf˙1945**, **weinert˙1995**]. There is the equally troublesome observation that humans do not use all or even most of the syntactically correct ways to express an idea, but prefer to stick with one standard means of expression [**hymes˙1970**, **bygate˙1988**]. This holds true even when that favored expression does not conform to the agreed syntactic rules [**pawley˙syder˙1983**, **wray˙1998**].

Munroe's estimate might be a good approximation of the true number of unique English tweets if humans decide what to tweet by randomly sampling the space of all possible tweets. This seems unlikely for a few reasons. On a superficial level, most tweets are going to be about temporally relevant things, often constrained by conversations and audience [**marwick˙boyd˙2010**]. Thinking a little deeper, the possibility space of language only reflects combinations of things within the lived experience of humans[3]. Finally, at the most profound level, the possibility space of human language at time t must be limited to one or only a few steps beyond the space at $t - 1$ because of the way that cultural knowledge is transmitted and transformed.

Additionally, there is a large cognitive constraint in terms of the size and recursive depth of linguistics constructs and their capacity to overload human memory. This is usually noted but also usually ignored, following an old argument about the importance of studying competence versus studying performance[4] [**chomsky˙1965**]. We do not plan on taking up this particular debate here, as there is not room enough to adequately cover the evolving arguments on both sides. Outside of the context of that debate, it is important to bear in mind that there is a real physical limitation on the length and depth of human language[5]

[3] In the ancient world, the most fantastical myths still were composed out of normal things, like a desert guardian which is a combination of parts that belong to normal animals, see also [**boyer˙2001**]

[4] which is itself inspired by an even older argument of Plato's about the ideal versus the real

[5] usually given at 3 levels of recursion, and/or 30 words

which practically restricts recursive depth and therefore also the reality-constrained possibility space.

To assess the probability that human language is practically infinite, we employed an empirical method to estimate the function governing the probability of encountering a new sentence after observing a sample of n sentences. This method was applied to a corpus of conversational, text-based data collected online. We used this function to calculate the sample size at which hearing a new sentence becomes statistically improbable[6], which we take as our estimate for the total number of unique sentences. We then use this function to make estimates for the timing of some landmarks in language acquisition.

5.1 Methods

Processing

Data were collected in Python 3.5.1 on Ubuntu Server 14.0.4, and were analyzed with Microsoft R Open[7] based on CRAN release v. 3.2.3, "Wooden Christmas-Tree", employing tidyr and dplyr for data reshaping [**r**, **tidyr**, **dplyr**]. Tables were produced with xtable, and figures were produced with ggplot2 [**xtable**, **ggplot2**]. The python program for generating the data, the raw data, and the R markdown file necessary to reproduce this analysis along with its tables and figures are available at github.com/deniederhut [**knitr**].

It is desirable to conduct analyses on corpora that are both large and ecologically valid. However, there is an inherent trade-off between these two features, as the closer one gets to spontaneous conversation, the harder the data become to transform and process. As an example, the largest readily available corpus, Google Books, is constructed from the scanned text of library collections. As a result, it tends to have a larger vocabulary, fewer idioms and slang words, more standard grammar, and more obscure and technical topics than the English employed by contemporary speakers [**pechenik˙danforth˙dodds˙2015**].

To strike a balance between validity and size, a sample of $8M$ comments of conversational English was collected from the online discussion forum, AskReddit. These comments were separated into individual sentences using NLTK's sentence tokenizer [**nltk**], resulting in a sample size of $15M$ sentences and $250M$ words, making it somewhat larger than the Brown corpus at $1M$ words. These sentences had an average length of 77.8 characters, and 17.1 words, based on applying NLTK's word tokenizer to the corpus.

This population of sentences was sampled without replacement by randomly drawing size n of sentences, then iteratively downsampling it by $\frac{n}{2}$ until the sample size was less than 10. At each iteration, three methods of determining the uniqueness of a sentence were used to compare each sentence in the subsample with the rest, and the total number of unique sentences, the method of calculation, and the size of the sample were recorded. Because these three methods differed greatly in computational complexity, the slower methods were

[6] although not impossible

[7] https://mran.revolutionanalytics.com/open/

not applied to sample sizes that would have taken longer than 120 hours wall clock time to compute.

Similarity metrics

The fastest of these three methods was to use Python's built-in SET constructor, which defines a duplicate sentence as one in which every byte matches. In other words, this metric considers

> This is my house

and

> This is my   house

to be two unique sentences. This method is desirable for running in linear time, but undesirable in being very conservative. Below, we will refer to this as BITWISE similarity. Bitwise similarity was applied to all samples, including the entire corpus of $15M$ sentences.

One potential way to avoid classifying two sentences as unique when they differ only by whitespace is to use a similarity metric for the character strings. Our second test method uses fuzzywuzzy's implementation of Levenshtein distance to judge the similarity of two strings based on fuzzy matching of the character combinations [**fuzzywuzzy**]. Below, we refer to this as LEVENSHTEIN similarity. Because Levenshtein similarity was quite a bit slower than bitwise similarity, it was not used on samples $>= 150,000$ sentences.

The third method uses gensim's implementation of cosine similarity to judge the semantic similarity of two sentences. In this method, each pair of sentences was tokenized, and vectors were built from IDF weighting them against a tertiary document containing the 25 most common English unigrams, in their relative ecological proportions. Then, the cosine similarity of the two sentence vectors was calculated. This was our current best approximation of our intuition about what makes a sentence unique, but it may still overestimate the difference between two sentences. Below, we refer to this as COSINE similarity. Because cosine similarity was the slowest of all the test metrics, it was only applied to relatively small ($<= 23,000$) samples.

These latter two methods had their parameters tuned such that

> This isn't my house

and

> This is not my house

were judged to be the same sentence, and

> This is my house

and

This is not my house

were judged to be different sentences. This was achieved by setting the match strength to 90% and 50% for Levenshtein distance and cosine similarity, respectively.

Model

To estimate when a person will stop encountering new sentences as a function of the number of sentences already observed, we must first estimate the probability function governing the appearance of new sentences. We chose to model this function as the log probability of drawing a new sentence from a population after sampling n sentences. The proportion of unique sentences at each sample size was calculated by simply dividing the number of unique sentences from each method by the sample size from that iteration. To coerce the relationship between proportion unique and sample size into the characteristic S shape of a logit curve, the sample sizes were log transformed (Fig. 5.1).

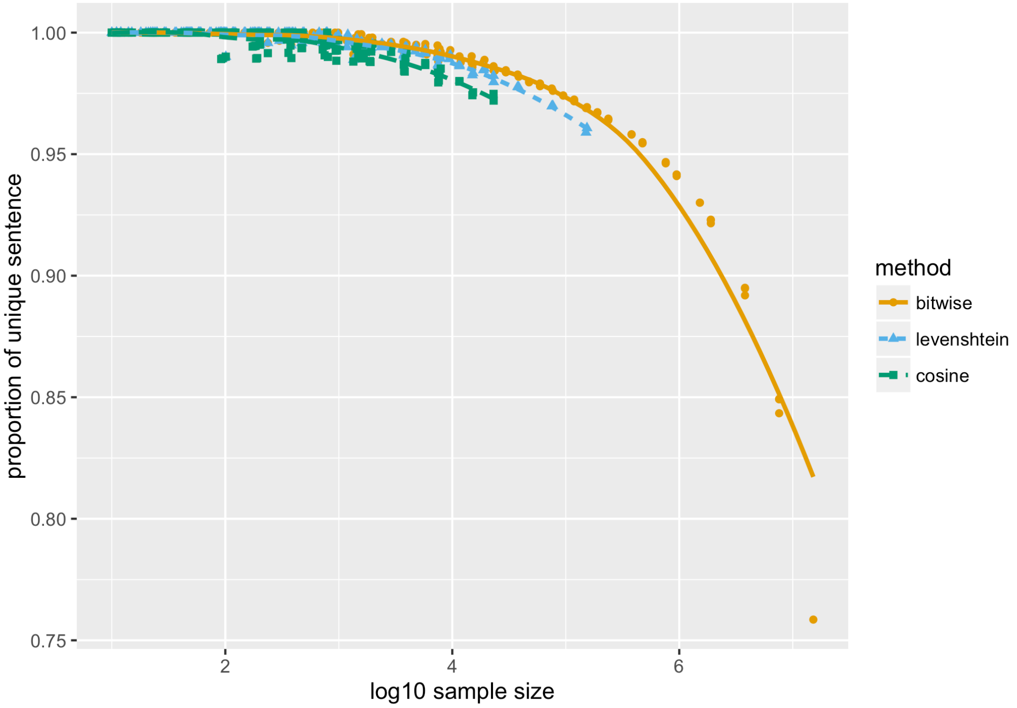

Figure 5.1: When sample size is log transformed, the probabilities take on the shape of a logit function

When the logit of the proportion is taken, this relationship linearizes (Fig. 5.2). Data from each of the similarity methods were fitted with a linear model predicting logit proportion from log sample size.

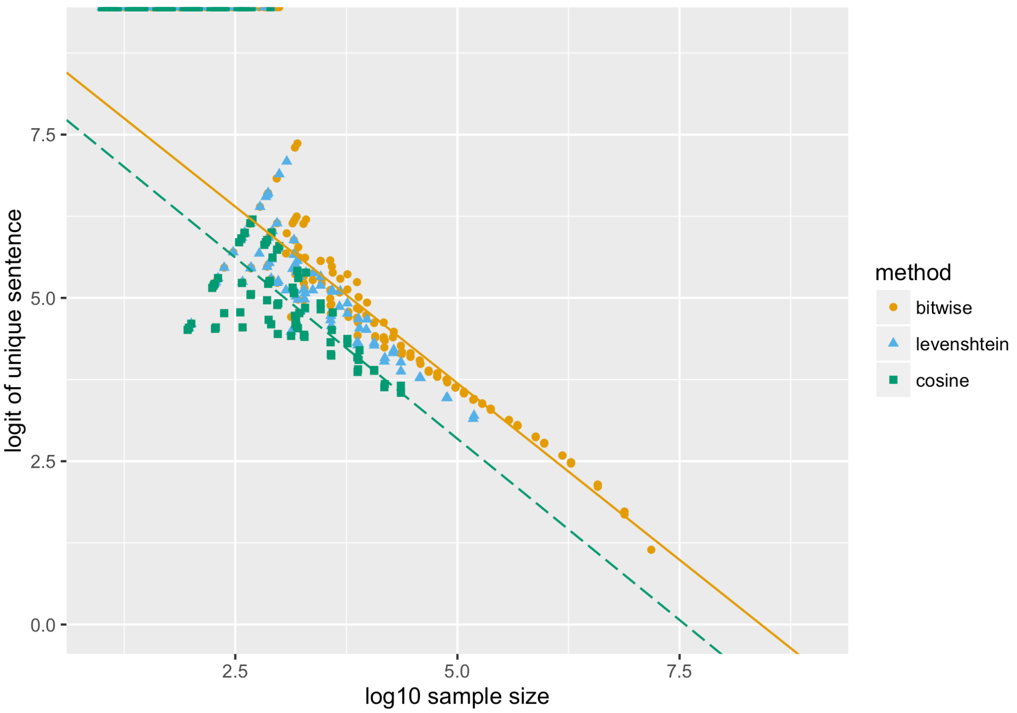

Figure 5.2: At every sample size, the logit probability for finding a cosine-unique sentence is an order of magnitude lower than bitwise-unique

These models were then used to produce estimates of the number of unique sentences observed by major developmental landmarks, including time to first word, first sentence, and first story, which occur around the ages of 1, 3, and 5, respectively. These were then divided by the number of unique sentences necessary to have been observed to produce a 1% chance or less that the next sentence encountered would be novel. The values produced by this method were compared to the proportion of sentences observed using the information theoretic measure, multiplied by 99% of its total size to keep the scaling consistent.

5.2 Results

The proportion of unique sentences at a sample size n can also be thought of as the probability that the next sentence you encounter will be unique, given that you have already observed n sentences. We can then say that the probability of encountering a new sentence decreases nonlinearly with the number of sentences you have already heard (Fig. 5.3). We take this as evidence that humans do not randomly draw sentences to speak from the sample space, but rather prefer to use some sentences very frequently, but most sentences very rarely.

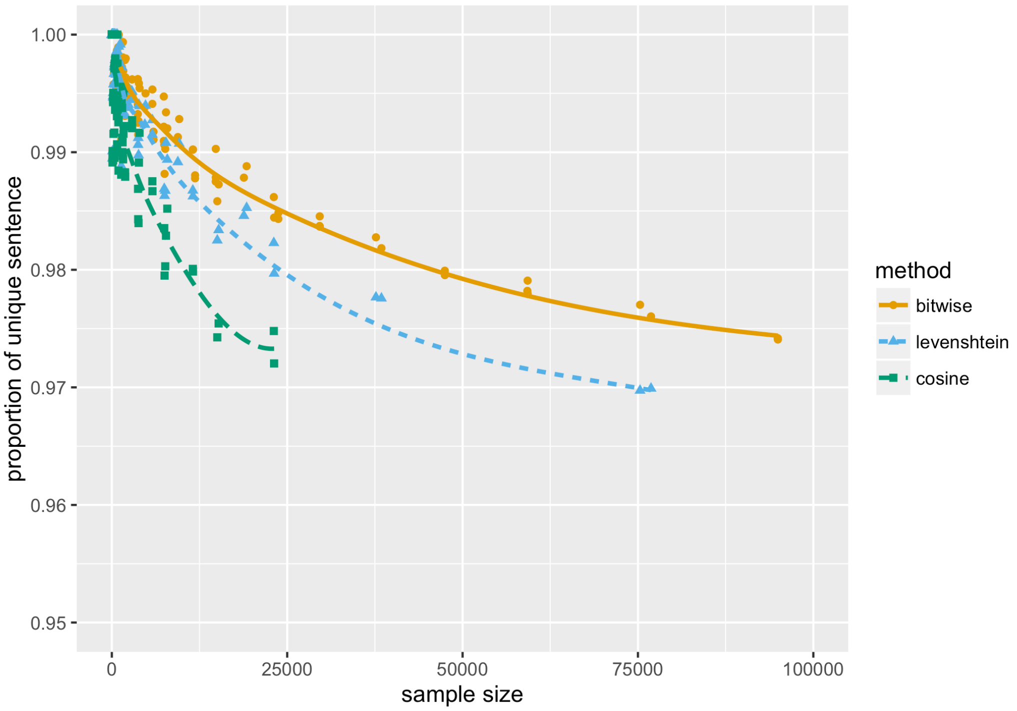

Figure 5.3: The probability of encountering a unique sentence decreases nonlinearly with the number of sentences sampled

The linear models predicting logit probability from log sample size have slopes that are roughly parallel, differing primarily in their intercepts (see Tab. 5.1, Tab. 5.2, Tab. 5.3). To achieve the same probability of hearing a new sentence, a learner must hear roughly an order of magnitude more sentences for the bitwise measurement than the cosine measurement. The intercept for the Levenshtein measurement is in between the other two, which may have been expected based on the permissiveness of their matching.

It is interesting that these slopes are largely parallel, converging around -1.1. We have no intuitions about why this should be the case, or why the number itself appears to mirror Shannon's estimate for English character entropy. The parallel slopes do, however, make us feel more confident in extending the predictions of the cosine metric to the range covered by the bitwise metric.

To achieve a 1% likelihood of hearing a new sentence (bitwise), one must hear $4.64 * 10^{12}$ sentences. For the Levenshtein measure, this number drops to $7.77 * 10^{11}$, and down further

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.0997	0.1324	68.75	0.0000
log(n)	-1.0812	0.0292	-37.00	0.0000

Residual standard error: 0.342 on 113 degrees of freedom
Multiple R-squared: 0.924, Adjusted R-squared: 0.923
F-statistic: $1.37 * 10^3$ on 1 and 113 DF, p-value: $< 2 * 10^{-16}$

Table 5.1: Linear model to compute logit p of encountering a new sentence given sample size n and a bitwise similarity metric.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.0077	0.3088	29.17	0.0000
log(n)	-1.1440	0.0819	-13.97	0.0000

Residual standard error: 0.279 on 35 degrees of freedom
Multiple R-squared: 0.707, Adjusted R-squared: 0.699
F-statistic: 84.5 on 1 and 35 DF, p-value: $< 7.3 * 10^{-11}$

Table 5.2: Linear model to compute logit p of encountering a new sentence given sample size n and a Levenshtein similarity metric.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.3906	0.4315	19.45	0.0000
log(n)	-1.1095	0.1207	-9.19	0.0000

Residual standard error: 0.334 on 51 degrees of freedom
Multiple R-squared: 0.793, Adjusted R-squared: 0.789
F-statistic: 195 on 1 and 51 DF, p-value: $< 2 * 10^{-16}$

Table 5.3: Linear model to compute logit p of encountering a new sentence given sample size n and a cosine similarity metric.

to $5 * 10^{11}$ for cosine-based comparisons (Tab. 5.4).

5.3 Discussion

While the total theoretical number of unique sentences in English is very large, the number that are actually used in conversation is smaller by several orders of magnitude. Using the Shannon method for this corpus, we can use our measurement of average sentence length, 78 characters, to get an estimate for the theoretical number of English sentences, $5.9 * 10^{25}$ (see Eq. 5.1). While slightly smaller, this is not appreciably different than Munroe's estimate for tweets.

p	bitwise	levenshtein	cosine
0.90	2.42E+06	8.97E+05	3.82E+05
0.50	2.61E+08	7.47E+07	3.65E+07
0.05	1.38E+11	2.80E+10	1.65E+10
0.01	4.64E+12	7.77E+11	5.06E+11

Table 5.4: Number of sentences a person needs to hear before they have a *p* probability of hearing a new sentence, computed by different similarity metrics.

However, our estimates for the number of probable unique sentences using the bitwise method, $4.64 * 10^{12}$, is more than ten orders of magnitude smaller [8] (Tab. 5.4). The less conservative cosine estimate is another order of magnitude smaller still, at $5 * 10^{11}$. We consider a difference of fourteen orders of magnitude between the empirical estimate and the theoretical estimate to indicate that the theoretical one may be unrealistic.

To be sure, five hundred billion unique English sentences is still a very large number – much larger than any person will encounter in their lifetime. However, there is no need to hear every possible sentence in order to infer enough about the structure of a language to start producing your own sentences in it. It should be possible to do so based on a relatively small sample, especially when that language has regular grammar rules that apply to the more uncommon utterances; and, when many of the sentences heard in daily life are ritualized utterances surrounding typical social situations.

Let's assume for the sake of argument that the average person hears about nine hundred sentences per day, or $3.4 * 10^5$ per year (a rough estimate based on research from [**mehl˙et˙al˙2007**]). We can use this value to do a back of the envelope calculation of the probability of encountering a new sentence by age.

Using just this estimate for number of sentences per year, we can apply the logit model in a fairly straightforward manner to calculate that a 7 year old child can expect only 90% of the sentences they hear to be novel [9]. By their 60s, a person has only an 77% chance of hearing something new.

We can also use this model to ask a slightly different question that will allow us to make direct comparisons with the Shannon estimates – what is the fraction of all possible sentences a person can expect to have heard at any age? This is computationally trickier, but a few key milestones have been been plotted for both the theoretical estimate and our ecological estimates (Fig. 5.4).

It is clear here that a poverty of the stimulus type argument clearly applies for the Shannon estimates, as the proportion of language experienced by a human during their life remains 0%. The bitwise and cosine estimates, however, grow visibly larger over the course

[8] Another way to look at this, is that at the Shannon estimate for the total number of possible sentences in English, the probability of hearing the same sentence twice is $4 * 10^{-10}$, or less than one millionth of one percent.

[9] unless stated otherwise, these estimates will be based on the bitwise equation to keep them conservative

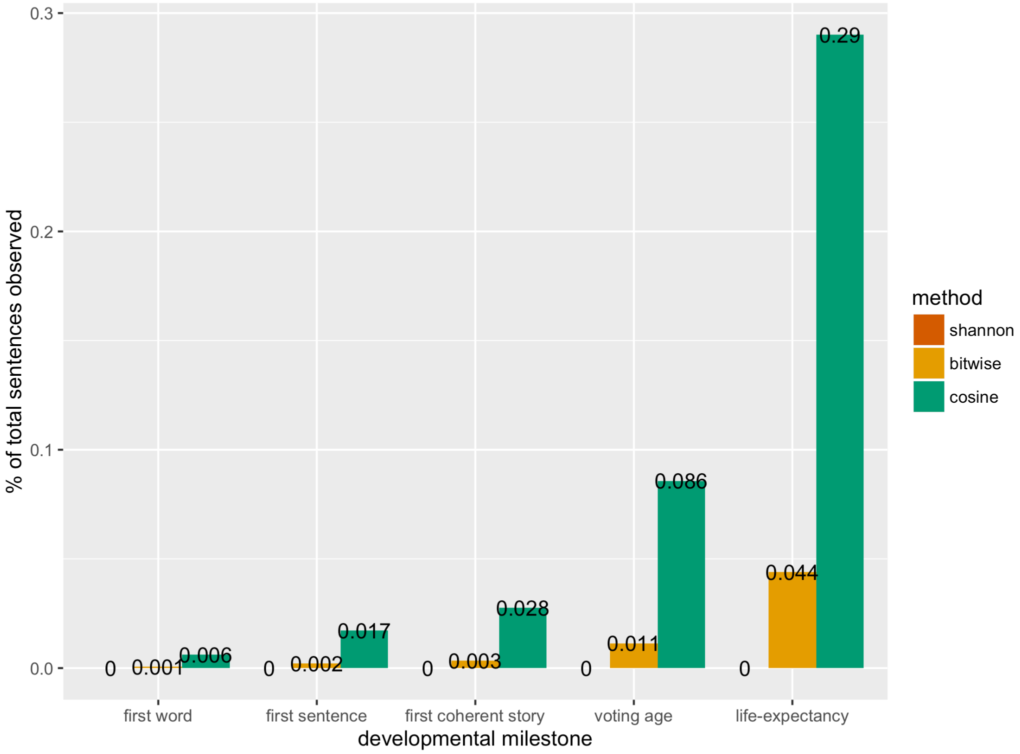

Figure 5.4: Percent of total likely sentences heard by a human at developmental milestones, using three different methods of calculating “total likely”.

of the lifespan, until a human can expect to have experienced 1/300th of the semantically (cosine) unique possibilities of their language by the end of their life. This seems to be a more realistic estimate on its face, and paints a different picture of the relative quantity of language received by humans while learning to speak.

One interesting consequence of calculating time-to-competence this way is that the rate of exposure to English matters a great deal. It’s possible that the well-documented effect of reading on increasing language fluency is simply a side effect of the fact that it is faster to read words, roughly 5/3 as fast in English, than to listen to them spoken aloud [10]. If a person were to replace all of the time they spent listening with time spent reading, the proportion of the total possible sentences in English they would be exposed to would roughly double over the course of their life, with their likelihood of surprisal would only drop from 77% to 72%.

[10] Also, perhaps, that it is likely easier to find a day’s worth of reading than a day’s worth of conversation

5.4 Conclusion

There are some concerns with the estimates presented here. First, the English corpus is drawn from a particular sociocultural community which largely consists of 18-35 year old educated white males living in the United States, and thus fails to capture the full diversity of the English language. Additionally, the corpus was collected over a relatively short period of time – just 6 months. An actual human is exposed to language that changes over time, presenting something of a moving target. Finally, the amount of exposure is only one factor in learning a language, which must also include factors like pragmatic usage and social context.

However, our estimates for the number of likely sentences in English benefit from being grounded in observations of real use of language. They are generated from a probabilistic model of encountering sentences, which is more ecologically valid than a model which treats the distribution of sentences to be rectangular in their probability of appearance. Finally, they also produce more reasonable values for language exposure.

These estimates were used to produce three major sets of findings. First, that a large portion of any person's language exposure is to the same set of common sentences. We saw this above where an adult in their 80s has a 23% chance that the next sentence they hear will be something they have heard before. Halving this probability of surprisal drops the requisite age by a factor of seven[11]. Second, we deduce that the total number of likely sentences is much smaller using empirical estimation, suggesting that theoretical estimates greatly overestimate the number of viable sentences in a language. Third, the models fitted to empirical data produce a more reasonable account of how much language a child needs to hear to make inferences about the structure of that language.

[11] in other words, an 8 year old is only twice as likely to be surprised by an English sentence as a 60 year old

Chapter 6

Computational Approaches to Semantics

6.1 Trade-offs between cost and ambiguity

Historically, inquiry into the means by which information is encoded in natural language has depended on a series of related models. Each of these posits a pair of communicators, one speaker and one listener, and the communication path between them, which is typically a linear sequence of auditory signals. Different researchers focus on different aspects of this particular model, but the conclusions drawn tend to be similar or at least mutually reconcilable.

The earliest perspective taken on this model focused on the communication channel itself. Shannon, in a series of psychometric experiments, calculated the additional information provided by a character given a preceding sequence of characters. This, in turn, was used to estimate that the information density of English is about 1.1 bits per alphabetic character. If the goal is to maximize the amount of information transmitted per character, this value is somewhat less than desirable.

On the other hand, maximizing the amount of information per English character also means that you are maximizing the information loss if one of those characters should not be transmitted successfully. This has lead to the model of a NOISY CHANNEL, where it is possible to drop or mistranslate information transmitted in that channel. In this view of the communication model, the goal is to maximize the amount of successfully transmitted information given some rate of information loss. Maximizing the amount of transmitted information requires building in enough redundancy into the signal that it can still be reconstructed successfully. In the best case, the information that is meant to be transmitted will be split evenly across characters in the signal, so that there is not one single text character or discrete moment in time that will cause catastrophic information loss [**jaeger˙tily˙2011**].

Another perspective on this model is to place attention on the speaker and listener, instead of aspects of the channel, and to assume that both individuals are driven to expend the smallest amount of effort during the interaction [1] [**zipf˙1949**]. For the speaker, this could mean something like using a single sentence, single word, or single emphatic grunt to convey some intended amount of information. For the listener, this could mean something like receiving an encyclopedia entry, or living in a hypothetical world where the language has such a large vocabulary that each individual life experience has its own special word.

These two partners (or combatants, depending on your mood) reach a compromise that maximizes the information transmitted whilst balancing the efforts of the speaker and of the listener. In practice, of course, human utterance does not consist of no words nor of encyclopedias passed back and forth, and neither is is a language consisting of a single sign with infinite inference nor infinite signs with no disambiguation.

In fact, it has recently been put forward that shared knowledge of the environment is a crucial secondary channel of information [**piantadosi˙tily˙gibson˙2012**]. This effect exists both at the general scale of cognitive primitives, and the more specific scale of interaction context. The former, C-INDUCTION, explains why humans don't need to explain to one another what they mean by *I* or o *comfortable*. The latter, introduced under the name FRAME SEMANTICS, explains why native English speakers don't need additional disambiguation around *table* when used in the following:

1. She put her glass on the table
2. Let's table this discussion
3. Table twelve wants their check
4. The water table is dropping rapidly

It has been argued that the common understanding of a situation is actively used as a method by which to reduce the amount of additional information that needs to be vocalized. This shortcut includes both the Gricean observation that people tend not to transmit information that is already common knowledge, and also the fact that native speakers do not have trouble acquiring the correct reference for *table* given an appropriate understanding of the communicative context. By assuming shared knowledge, speakers are able to shorten the total amount of information, to use shorter and/or more common words that take less effort to decode, and to push some of the redundancy requirements imposed by a noisy channel into the inferences likely to be engendered by a particular situation.

6.2 Methods for measuring information

These different perspectives leave unanswered the question as to how best to quantify the amount of human understanding transferred in any given utterance. Approaches here tend to focus on one of two directions: unpredictability and description.

[1] Zipf calls this THE PRINCIPLE OF LEAST EFFORT

Surprisal

One approach is to use a collection of one or more of a given size unit of the language signal to predict the next individual piece of the language signal. If the predictions are always accurate, then the next piece is not adding any information, in the sense that it is not providing you with knowledge that was not already present. On the other hand, if the next piece of the language signal is difficult to predict, or unexpected even in what had been easy situations, then that piece gives you quite a bit of information.

The basic idea is to construct a chain of conditional probabilities, of length n, of units, and then to use this chain to predict the unit in the $n-1$ position. This approach has the benefit of being simple both in conceptualization and in implementation. With a large enough sample of language to train against, this approach boils down to counting occurrences and computing the maximum likelihood of any particular combination. On the other hand, the required sample size to witness a stable estimate of every combination of length n grows exponentially, which is not computationally tractable.[2]

This method can be applied in the forward direction, outlined above, to generate a LANGUAGE MODEL, which can be used to probabalistically generate arbitrarily large sequences of text. One could think of this as predicting the next word given a particular context, where the context consists of the previous n words. This can also be applied in the backwards direction, which provides a measure of how expected some context is given a particular word. In both cases, the measurement of information is related to entropy, and indicates some degree of order. Surprisal, or unexpectedness, is then essentially the same thing as randomness.

This has an important consequence. Where most people understand INFORMATIVE in the colloquial sense to mean bearing useful information SURPRISAL really only means random, or more random than expected. This does not mean that it is semantically valuable, as non-sequitur would also fall into this category. To take a concrete example, imagine we have some context about looking for the location of a library. *It is down the street* has less information, in the Shannon sense, than *elephants are pink on Tuesday.*

Recently, the language model has been used to reexamine Zipf's assertion that the length of a word is best predicted as an inverse relationship of its frequency in a given corpus. Piantadosi show that a more accurate predictor is the amount of information contained by a word, using a related metric which conditions this probability on a small number of preceeding words [**piantadosi˙tily˙gibson˙2011**]. The same author has also shown that orthographic length has an inverse relationship with the number of contexts to which a word may apply, as measured by applying the reverse language model described above [**piantadosi˙tily˙gibson˙2012**]. The authors argue that Zipf was essentially correct in that people aim to minimize communicative effort, but that the effort comes from interpretation and not the amount of time or muscular work required to produce some sequence of speech.

[2] In practice, various methods are used to impute data that is not measured accurately or at all. Popular choices include linear interpolation and redistributing probability mass.

Word sense

A second approach is to collect some descriptive reports of how useful a range of words are. An early effort comes from Zipf himself, who described an inverse relationship between the total number of dictionary entries for a word and its orthographic length [**zipf˙1945**]. One could imagine a similar exercise being undertaken with the semantic graphs recorded in WordNet or FrameNet.

Additional approaches include recording self-reports of "complexity" on an anchored Likert scale, measuring implicit mappings between fake words and visual forms of manipulated complexity, and building characterizations of language spontaneously generated in response to some prompt. Lewis and Frank have described a relationship between the orthographic length of a word and the specificity of its meaning using methods like these [**lewis˙frank˙2016**].

6.3 Updates to the communicative model

More recent work has updated this communicative model by including a prior assumption that the speaker and listener are communicating specifically about one particular label from a known domain. From the speaker's perspective, this means that there is a definite label in mind, and the task is to provide the smallest amount of information that correctly causes the listener to discover the same label. From the listener's perspective, the task is to use the given sign to narrow down a distribution of possible labels to the single most likely one, given expectations about the mapping between the structure of the semantic space and the structure of the lexical space.

Regier, Kemp, and Kay review a number of studies that look at the alignment between lexical partitioning and semantic partitioning of kinship labels, color labels, and generalizations of object names as feature vectors [**regier˙kemp˙kay˙2015**]. By restricting the semantic space to a well-defined and bounded domain, the authors of these studies have been able to construct measures of simplicity and informativeness that take into account the underlying structure properties of that domain.

For example, Kemp end Regier present a cross-linguistic analysis of the partitioning of relations into kinship labels [**kemp˙regier˙2012**]. The measure for complexity builds on the inherently hierarchical nature of family trees, and incorporates knowledge of age and gender. In this study, and in others like it, the authors tend to find that lexical partitioning is near optimal, e.g. [**xu˙regier˙2014**]. That is to say, any given attested partition will have other partitions that are simpler, and other partitions that are more informative, but not partitions that are simultaneously both.

Chapter 7

Introducing the Zipf statistic

The debate around whether language systems evolved because it was beneficial for early hominins to communicate well is currently settling in favor of a communicative origin and communicative optimization. However, there has yet to be a method for calculating the semantic magnitude of a particular speech act in a way that matches human intuition about informativeness versus randomness. Were such a method to exist, hypotheses that posit a utility motive for the origin of language could be tested in natural experiments involving the acquisition or change of that language.

7.1 Natural word use follows a Zipfian distribution

In conversational English, a few words are used many times, and the rest are used with relative rarity (Fig. 7.1) [**zipf˙1936**]. Even words that seem commonplace, like *carpet*, *scale*, and *weird*, appear fewer than once in every twenty thousand words. To put this in context, it has been estimated that the average person speaks 16,000 words per day [**mehl˙et˙al˙2007**]. In other words, for a given person on a given day, it is unlikely that a common word like *carpet* will be heard.

Intuitively, words that are used very frequently don't seem to mean all that much. The word MY, for example, really only tells you that something is being possessed by the speaker, but does not tell you anything about that something or about the speaker. To take another example, the word YEAR probably indicates a period of 365 days, but might not have the expected start date (i.e. a fiscal year) or might only refer to nine months out of those 365 days (i.e. an academic year).

On the other hand, uncommon words feel like they mean a whole lot. For example, were the speaker to use the word MULTICOLLINEAR, a native speaker of English familiar with statistics could assume the following statements (ordered by decreasing likelihood):

1. the speaker is a fairly educated person who does statistical analyses, speaking to one or more persons who are fairly educated and understand statistics

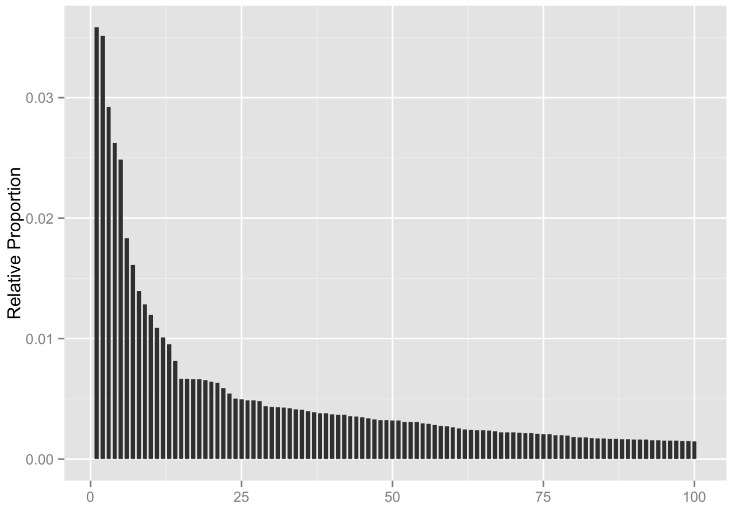

Figure 7.1: Relative proportion of the 100 most common words in English, from the sample described below.

2. the speaker, in particular, is well informed about the general linear model
3. the context under discussion is a model with two or more predictors that are highly correlated, and thus has unstable linear coefficients
4. the speaker will go on to mention ways to measure this, like the variance inflation factor, and ways to correct it, like principal component analysis
5. the speaker does not work with very large data sets or machine learning methods

In this particular case, a single word is giving us a rich set of inferences about who is involved in the conversation, their knowledge state, the topic under discussion, and what will happen next. This is much more information than was provided by *my* and *year*.

Ideally, one would determine the semantic magnitude of a particular word or phrase by measuring all of the attributes in the context in which that word was used, and then calculate some kind of conditional probability that a phrase indicates a feature in that context. However, *context* is difficult to define and even harder to measure. Successful studies that match language use to the environment apply themselves to very restricted domains, such as the application of discrete color terms to the continuous color spectrum [**regier˙kemp˙kay˙2015**]. To continue the example above, measuring the education level of every person involved in a conversation, along with the major topic under discussion and the fields in which the speaker does not work in a dataset large enough to draw meaningful inferences about language is not currently feasible.

However, that fourth inference above – measuring the frequency of associated words like *variance inflation factor* – is easy to measure in large quantities of language data, and requires no metadata about the language. Additionally, it should provide some subset of the same information as measuring the actual environment. To continue the example above, *multicollinear* is only probable in a context involving generalized linear models whose predictors have high covariance, a context which itself is likely to produce *variance inflation factor*. So, the existence of *multicollinear* should predict the presence of *variance inflation factor*, with some loss of information.

7.2 Semantic value is a change in that distribution

We can imagine, then, that a rare word is one which refers to a rare context; and conversely, that rare contexts tend to be described with rare words. In that case, we may posit that the semantic value of a word is related to the distance between the words associated with it relative to their frequency of use in the language as a whole [**wittgenstein˙1953**, **salton˙1971**, **deerwester˙1990**]. To state this another way, an informative word is one which changes the probability distributions of the words surrounding it each time it appears.

This argument differs from prior work which assumes that the semantic content of a word is defined by the words that appear nearby [**gentner˙1983**, **furnas˙1983**]. The model in this paper is that nearby words point to the same latent variable, which is the difficult-to-measure, real-world context of the speech act. The implementation, however, is very similar to modern methods in computational semantics like pairwise mutual information or latent semantic analysis (LSA) [**lund˙1996**, **turney˙pantel˙2010**]. The difference here is that we are not trying to assign a similarity between the distributions of one single word versus another single word; we are attempting to assign a single value that demonstrates the relationship between a single word and the distribution of all words in the English language.

Generally, one can calculate the distance between two non-parametric distributions using the Chi squared statistic. Specifically, we can compare the word distribution of all events that contain a single word of interest with the overall English distribution to get a single measure of the difference between the two distributions:

$$\sum_{i}^{k} \frac{\left(p_{sample} * n_{sample} - p_{population} * n_{sample}\right)^2}{p_{population} * n_{sample}} \tag{7.1}$$

where k is the list of unique words in the population; i is one word in that list; n_{sample} is the total number of words used in communicative events that also include the word being measured; and, p_{sample} is the probability that any one in n words is word i.

To make this a little more concrete, imagine looking at the informativeness of the word MULTICOLLINEAR. Communicative events that contain the word MULTICOLLINEAR also tend to contain words like VARIABLE, CLUSTER, and MODEL. These words have very high relative proportions in our sample, but low relative proportions in our population. So, for each of these, we might add something like:

$$\frac{(1E+01-1E-06)^2}{1E-06} = 1E+08 \tag{7.2}$$

to the total sum. However, these events also contain many common words like TO, OF, and A, at close to their relative proportion in the total population of words. Each of these words adds a smaller value to the total sum.

This method of measuring distance is favorable in that the appearance of rare words is heavily weighted. It is disfavorable in that it is also sensitive to the total number of words, n. A very common word like MY has a very large n, so even small differences between the proportions of a word in the sample and the population produce values that are in the zeroth or first of magnitude. A very uncommon word with a small n will frequently produce values that are equal to the expected frequency of each word in the population, which is typically below 1E-05 – five orders of magnitude less. When summed across the number of unique words in the population (the length of k), the effect of n dominates the calculated value.

7.3 Correcting the magnitude of that change produces a statistic

To produce a useful statistic from the chi squared values, one first needs to correct the bias produced by the size of the sample. Then, the distribution of the test statistic can be characterized to produce population parameters for the expected mean value and variance. Both of these steps require real word linguistic data that has been decomposed into a distribution of frequency counts.

An English word distribution was created using a Python library written by the author, available at `https://github.com/deniederhut/redicorpus`, by randomly sampling comments from the discussion board at reddit.com/r/AskReddit over a period of 22 weeks. This particular discussion board was chosen both for its high traffic rate and the broad topics and conversational nature of the discussions there. This resulted in a total sample size of 4.29E+06 communicative events, with 1.08E+08 total unigrams and 2.93E+05 unique unigrams which appeared more than once.

To describe the sampling distribution of chi square values, comments were randomly sampled from the total corpus twenty times each at the probabilities 1E-02, 1E-03, 1E-04, and 1E-05. In each of the 100 samples, term frequency was set to equal the number of comments, and the chi squared value of the term frequencies in the sample was calculated. As predicted above, the magnitude of the chi squared statistic is dependent on the size of the sample used to calculate the statistic (Fig. 7.2).

The relationship between chi squared values and sample size linearizes when each variable is square root transformed. This linearization may be explained by Zipf's observation that the number of meanings a word can have increases as a function of the square root of its

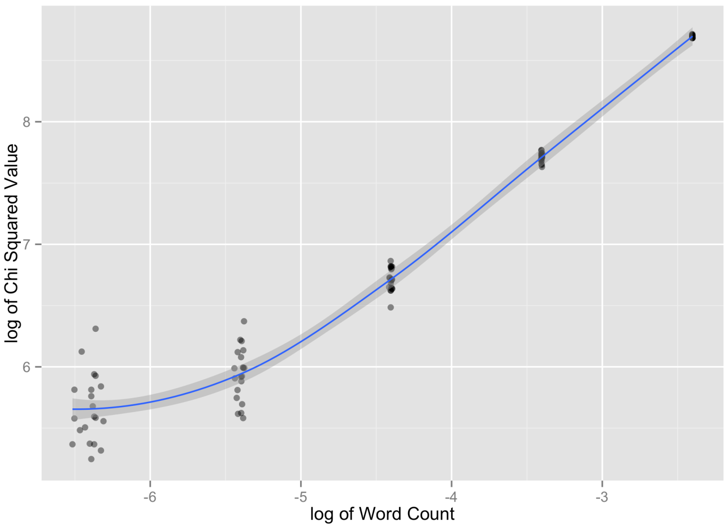

Figure 7.2: Chi squared values produced by comparing word distributions are convolved with the size of the sample.

relative proportion [**zipf˜1945**]. Happily, this transformation also causes the variability in chi squared values to become a constant. After linearization, the chi squared values from each random sample are almost perfectly predicted by the total number of words in all the comments used to compute the value (Fig. 7.3).

The stable variability means that a simple metric can be created using the chi square value, corrected for word count, and divided by the constant standard deviation. For the sake of brevity, we'll call this the Zipf statistic.

$$z_{correct} = \frac{(-274.85 + \frac{\sqrt{n}}{6.37}}{288.33} \tag{7.3}$$

7.4 The metric statistic conforms to expected behavior

As a proof of concept, the Zipf statistic was calculated for several words from the corpus (Table 7.1). Generally speaking, it produces values concurrent with our intuition. Very common words, like *my*, *day*, and *feel*, have negative Zipf statistics that increase linearly in magnitude with the square root of their relative proportion in the population. Words with a relative proportion around 1E-05 have Zipf statistics that are close to zero, or close to the corrected values derived from randomly sampling the entire population. Words that appear

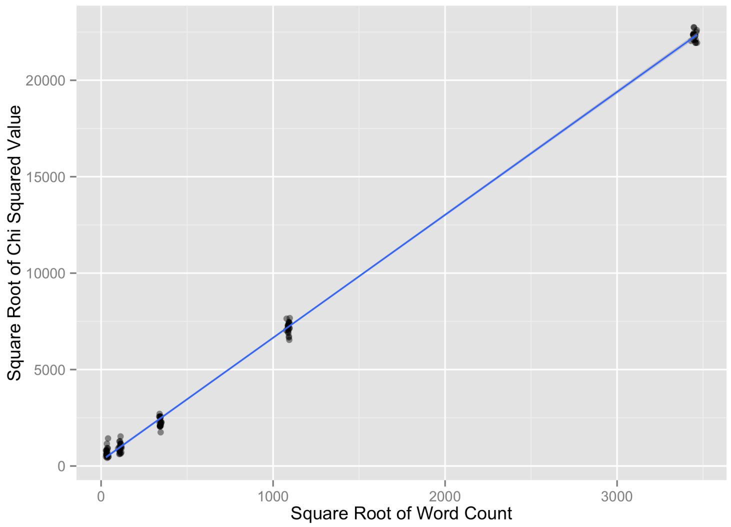

Figure 7.3: Square-root transforming the chi squared values and sample sizes linearizes both the relationship and the variability in y.

less frequently than this have an increasing change of producing positive Zipf statistics, indicating words with high semantic content.

Term	Relative Proportion	Test Value
my	1.30E-02	-9.41E+01
day	2.12E-03	-5.82E+01
feel	1.71E-03	-5.04E+01
record	1.24E-04	-7.13E+00
unclear	1.11E-05	-8.12E-01
hither	6.09E-07	6.80E-01
dill	2.92E-06	8.25E-01
omlette	3.42E-07	1.78E+00
multicollinear	3.69E-08	3.28E+00

Table 7.1: Zipf test values conform to expected behavior.

The Zipf statistic provides a method for quantifying the relative informativeness of any word, given a language corpus that is divided by communicative events. Additionally, this is an objective method that can be implemented largely without human intervention, and in any language that is easily tokenized or lemmatized. It measures the semantic value of a given word by comparing the frequencies of other words used in the same context with the distribution of all the words that appear in the corpus.

7.5 Hypotheses made tractable by this method

If the evolution of human language was driven by a need for effective communication, we might expect evolutionary pressures to produce cognitive systems that prioritize the acquisition of information-heavy terms. Specifically, we would hypothesize that words with high semantic value would:

- be learned sooner in infancy; and,
- spread more quickly through a population; and,
- be preferentially adopted across languages.

More generally, the Zipf statistic should be useful in testing predictions of language change and use that include social and cognitive factors. For example, one could ask whether the diffusion of linguistic variants is better predicted by the utility of the word, or its use as a marker of social identity [**eisenstein˙2014**]. It should also be possible to create historical data on the rate of semantic bleaching of words, and to investigate if a relationship exists between that rate and populations employing the word.

Chapter 8

Features that Correlate with Semantic Value

Vector space models have become a popular way to represent the semantic relationships between terms [**salton˙1971**, **salton˙wong˙yang˙1975**]. This is in part because they extract knowledge about the semantic structure of a corpus in an unsupervised manner and with minimal prior knowledge about the language [**turney˙pantel˙2010**]. They also have a mathematical representation that is very close to the distributional theory of semantics [**wittgenstein˙1953**, **harris˙1954**].

Vector space models have principally been used to establish the semantic similarity between documents, based on this distributional theory. This has been a very successful model for discovering things like latent topics in a group of documents, or the similarity of a word with itself across time [**blei˙ng˙jordan˙2003**, **hamilton˙leskovec˙jurafsky˙2016**, **xu˙regier˙malt˙2016**]. Variations of two and three dimensional matrices representing the relations between terms, documents, and topics, have until recently been the state of the art in natural language processing.

However, the information provided by these comparisons is relative, by definition. This means that, for example, vector space matrices have been successful in deciding whether two news articles have the same topic, but less successful in deciding how informative either article is. A metric called the Zipf statistic was recently proposed for summarizing how semantically valuable a term is [**niederhut˙2016˙1**]. It works by composing a vector space matrix comparing the context surrounding a term to an entire corpus[1]. Such a statistic allows us to quantify the amount of information contained in any ngram, as an absolute magnitude.

In the Zipf statistic, a document is made by concatenating all of the comments that include a given term, **t**, after one instance of **t** has been removed from each comment. Then, the vector of this document is compared against a vector computed by concatenating all comments in the corpus, **c**. The distance between these two documents is computed with

[1] One might think of this as a `context-corpus matrix`

a chi-squared statistic, which is favorable in that it is more sensitive to the appearance of rare terms than cosine similarity. However, it produces a value that is convolved with vector length, so the statistic for **t** must be also normalized (see chapter 7).

Conceptually, the Zipf statistic calculates a conditional normalized distance between the distribution of terms that a native speaker expects in a document. If the native speaker knows nothing about the document, this distribution is the same as the relative distribution of terms across an entire corpus. If a speaker knows one or more terms in that document, this distribution changes to the conditional probability of terms given that knowledge. If a term doesn't change a speaker's expectations about what will appear, it is not very informative, and therefore the distance will be small (when normalized, negative). However, if a term causes large changes in what a speaker expects to be in the rest of the utterance, that term is very informative, and will have a large distance (when normalized, positive).

Term	Zipf Test	Tf-idf	Cosine	Entropy
txt	7.48E+00	1.10E-05	8.94E-01	9.21E+00
crass	6.54E+00	2.24E-05	9.77E-01	9.01E+00
vegan	-2.06E+00	2.22E-04	9.63E-01	9.28E+00
sugar	-6.30E+00	8.69E-04	9.83E-01	9.62E+00
dude	-1.91E+01	5.01E-03	9.84E-01	9.56E+00
how	-5.31E+01	3.71E-02	9.89E-01	9.62E+00

Table 8.1: The Zipf statistic is a qualitiatively unique measure of the information contained by text.

The Zipf statistic produces results that fit our intuitions about the relative semantic value, or informativeness, of terms. For example, *vegan* produces a Zipf statistic a whole order of magnitude larger than *how* (see table 8.1). If all we knew about a person was that they had said *vegan* yesterday, we would also know that they are either a vegan themselves or were complaining about other people who are vegans. If the former, the resulting conversation is also more likely to include a discussion of sustainability, animal welfare, and/or personal health. If the latter, the resulting conversation also likely involved jokes about atheists and crossfit enthusiasts. Conversely, if all we knew about a person was that they had said *how* yesterday, the only reasonable inference we can make is that they might speak some amount of English.

Other more common, and easier to compute metrics, do not necessarily capture the same kind of information. For example, Tf-idf is primarily about establishing a comparison between how often a term occurs within a document, as opposed to between documents. We can visualize this by plotting the termwise Zipf statistics against these other metrics (Fig. 8.1). These other measures do not tend to have a monotonic relationship with the Zipf statistic.

Prior research measuring the information content of words has principally used two strategies. The first is to enumerate a number of possible meanings or contexts for a given word,

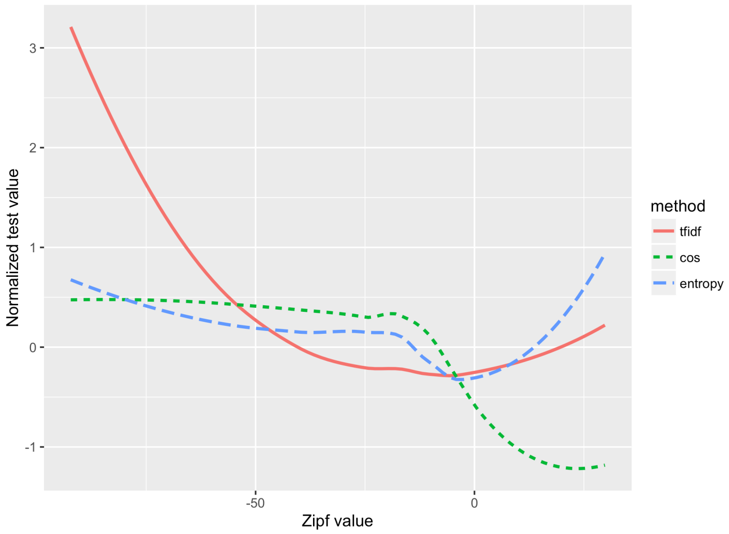

Figure 8.1: More common metrics of the informativeness of text do not have monotonic relationships with the Zipf statistic.

and to conduct analyses on the enumeration counts [**zipf˙1949**, **lewis˙frank˙2016**]. The second is to build language models that look backwards; that is, to estimate the conditional probability of a context given a specific word [**piantadosi˙tily˙gibson˙2011**]. Our approach here is much closer to the second metric, with two important differences.

The first is that we do not use a language model; and because of this, we can include the full information of an entire utterance. In Piantadosi, for example, the CONTEXT of a word is taken to be the two or three words immediately preceding it. In a sentence like:

> Looking at the sleeping volcano, she couldn't tell **that it had** *erupted* just last year.

the context, in bold, consists mostly of noninformative stop words which don't seem to have anything to do with volcanoes erupting [**piantadosi˙tily˙gibson˙2012**]. However, the computational cost of computing language models increases exponentially with the number of words considered, so using the entire sentence as context is currently computationally intractable [**jurafsky˙martin˙2008**]. The use of vector space models allows us to more or less ignore this constraint.

The second difference is that we do not use the common measure of informativeness, which is the reciprocal entropy or SURPRISAL. The principle reason is that entropy does not distinguish between words that are highly informative and words that are non-sequitur. Instead, we have used a standard nonparametric method for establishing the magnitude of the difference between two distributions.

The implied assumption across much of linguistics is that the semantic value of any higher order linguistic unit is constructed combinatorially from its lower order units [2]. That is to say, the information conveyed to a native speaker by some phrase is a linear combination of the semantic value of the words in that phrase. Or, as stated in [**chomsky˙berwick˙2016**], that "words are the atoms of language". For example, we can reason about the difference in meaning between the sentence:

Dog!

and the sentence

Brown dog!

as the modifier *brown* being added to the object *dog* and creating a unit whose semantic value is some combination of the information contained in each word. In this case, we can infer that the speaker is referencing a specific dog by pointing to its color, which is more semantically valuable than referring to some dog.

We have no intuition about the specific kind of relationship that is responsible for making *brown dog* more semantically valuable than *dog*, but we do insist that, in a combinatorial model, the value of the combination must be strictly greater than either component. In a linear model, this would mean that predicting the semantic value of the combination should involve the combined raw coefficients of the constituent terms being greater than one. A reasonable place to start would be to hypothesize that, in a combinatorial model, the resulting value is a strict linear combination of the values of the terms that comprise it.

$$V_{brown} + V_{dog} \propto V_{brown-dog} \tag{8.1}$$

An alternative hypothesis is that the meaning of an utterance is not built from the bottom up, but actually resides at the top of the organizational hierarchy [**deacon˙1997**]. It contains semantic information that is not available at lower levels, because of the emergent properties of symbolic reference systems. For example, we can reason about the difference in meaning between the sentence:

Dog!

and the sentence

Blue dog!

as the whole unit meaning something that is neither the color *blue*, nor the animal *dog*. In this case, our inference about the speakers communicative intentions can't be broken down into the individuals terms, but relies instead on knowing something about the political history of the United States. In this case, we can argue that:

[2] This is often called the PRINCIPLE OF COMPOSITIONALITY

$$V_{blue} + V_{dog} \subset V_{bluedog} \tag{8.2}$$

A model of communication where the semantic magnitude of an ngram is determined by its frequency in the population would seem to disagree with a model of communication where informativeness of an utterance wholly considered is created by successively concatenating terms. Here, we test the assumption that semantic value is constructed in this linear fashion, and describe an alternative view of the ontology of semantics with some supporting data.

8.1 Methods

Corpus

An English corpus was compiled with redicorpus[3]. The library has two basic components: a bot that builds daily corpora from web pages; and, a querying utility that returns term-comment vectors and the semantic test described in Niederhut [**niederhut˙2016˙1**], among other things.

Redicorpus was used to build an English corpus of $\sim$ 4,300,000 comments by randomly drawing comments from the 100 most popular discussion posts each day from AskReddit, for a period of 22 weeks. AskReddit is a discussion board on the popular website Reddit which supports open-ended discussions. It was chosen for its high rate of traffic – about one million unique viewers per day, according to internal metrics[4] – and for its conversational tone. While typed comments are not as ecologically valid as extemporaneous speech, they avoid the validity concerns of book-based corpora, which have been shown to overrepresent technical language [**pechenik˙danforth˙dodds˙2015**]; and, of periodical-based corpora, which overrepresent current events.

These comments were stemmed using NLTK's implementation of the Porter Stemmer [**porter˙1980**], word tokenizer, and ngram utility [**nltk**]. This resulted in a corpus of $\sim$ 100,000,000 unigrams, with $\sim$ 300,000 unique unigrams that appear more than once. Data collection was halted at this point to prevent memory overflow errors when building vectors out of the corpus.

Sampling

The subset of terms appearing more than once in the corpus was randomly sampled to produce $\sim$ 600 unigrams and $\sim$ 600 2-3 grams, in order to reduce the time and cost of running the computations. For each term, relative proportion, gram length, vector length, and Zipf value were calculated and entered into a data set. To fill in any missing comparisons, the bigrams in the sample were split into their component unigrams, and each unigram that was not already in the data set was retrieved from the corpus and added.

[3] v 0.0.2, available at github.com/deniederhut/redicorpus

[4] https://www.reddit.com/r/AskReddit/about/traffic

The corpus, which was created from raw HTML, included undesired terms like pieces of URLs. For preprocessing, this data set was stripped of terms that included numerical values or punctuation. Outliers, which we define as terms showing a Zipf statistic value greater than 30, were also removed prior to analysis. This resulted in a total sample size of 1,497 observations.

Processing

Data were collected in Python 2.7.8 on Ubuntu Server 14.0.4, and were analyzed with Revolution R Open[5] based on CRAN release v. 3.2.1, "World-Famous Astronaut", [**r**]. Tables were produced with xtable, and figures were produced with ggplot2 [**xtable**, **ggplot2**]. Data and R files to reproduce this analysis along with its tables and figures are available at github.com/deniederhut.

8.2 Semantic value is not compositional

A model of communication where the semantic value of an ngram is determined by its frequency in the population would seem to disagree with a model of communication where informativeness of an utterance wholly considered is created by successively concatenating terms. We wanted to test this latter possibility. In such a compositional model of semantics, it is implied that the value of utterance should be a function of component words. For example, the semantic value *brown* should be related to the semantic value of *dog* modified somehow the semantic value of *brown*.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9872	0.4360	4.56	0.0000
first.zipf	0.0288	0.0077	3.75	0.0002
second.zipf	0.0324	0.0078	4.17	0.0001

Table 8.2: Component unigrams are a weak predictor of the semantic value of the bigrams they create (a strong predictor would have a coefficient close to 1)

To test this hypothesis, a subset of the data was created, representing all of the bigrams from the previous data set with the Zipf value of the whole bigram, it's first unigram, and it's second unigram. If the component unigrams are contributing linearly to the semantic value of the bigram, we would expect to see coefficients of 1. Instead, the coefficients for the component unigrams, while statistically significant, are two orders of magnitude smaller, and the variance explained by the model is only 13% (Table 8.2). However, it was entirely reasonable to suspect that the relationship between the informativeness of unigrams and the informativeness of the whole utterance would be more complicated than the simple addition

[5] https://mran.revolutionanalytics.com/open/

of coefficients, so the data were run again in a full factorial model. In this model, the coefficients all became negative, and the main effects were no longer statistically significant (Table 8.3).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9137	0.4548	2.01	0.0463
first.zipf	-0.0009	0.0092	-0.09	0.9247
second.zipf	-0.0007	0.0096	-0.07	0.9447
first.zipf:second.zipf	-0.0013	0.0003	-5.15	0.0000

Table 8.3: In the ful factorial model, the main effect of the semantic value of constituent unigrams is no longer statistically significant.

8.3 Frequency is the best predictor of semantic value

Semantic value is inversely proportional to frequency

Because of the `long tail` of terms in a language, it is easiest to see the relationship between term frequency and other metrics when the terms are plotted as the base ten log of their relative proportion (Fig 8.2). It is evident that, as a term's relative proportion in the corpus increases, its Zipf value decreases. In the log plot, there is a clear inflection point where the slope changes from neutral to negative. Using a spline smoother, we find the inflection point to occur at a relative proportion of $1e-04$, or, where a term appears once in every 10,000 terms. If the Zipf statistic is a valid measure of semantic value, this would indicate that terms start to become less informative when they appear more often than words like *teenage* and *cancer*.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0228	0.3132	-3.27	0.0011
sqrt(tf)	-780.6324	18.9317	-41.23	0.0000

Table 8.4: The suqare of term frequency is a strong predictor of semantic value.

The relationship between relative proportion and Zipf values becomes approximately linear when the relative proportion of terms is square root transformed (Fig. 8.3). When regressed on the square root of relative proportion, Zipf values decrease at the rapid rate of -780 per unit increase in the predictor, with an R^2 of 0.53 (Table 8.4).

Semantic value of ngrams is weakly predicted by length

It was possible that higher order ngrams were driving the term frequency effect, as they are intuitively more informative than lower order ngrams and objectively less frequent in the

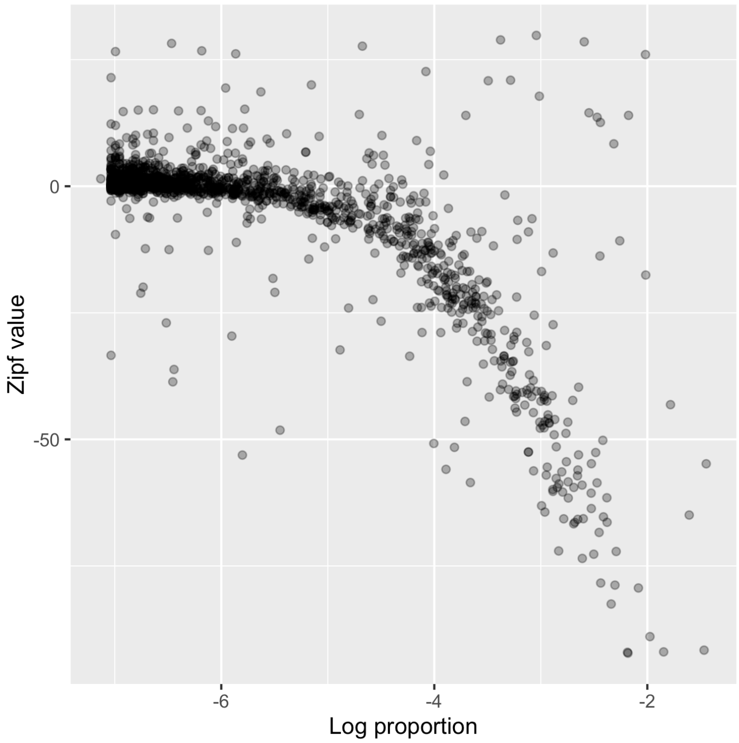

Figure 8.2: Plotting Zipf values against log term proportions shows a monotonic decrease in semantic value with increasing likelihood of term use.

corpus. To control for this potential causal confound, the gram length of each term was added to the linear model. Inclusion of gram length does not appreciably change the coefficient for relative proportion, and only increases the R^2 by 1% (Table 8.5). Furthermore, it's own coefficient is fairly small, and it disappears when common function terms are removed from the data set (see subsection 8.4).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.2219	0.7228	-5.84	0.0000
sqrt(tf)	-752.6501	19.6355	-38.33	0.0000
len	1.9111	0.3898	4.90	0.0000

Table 8.5: The number of unigrams in a unit of speech is weaker predictor of informativeness than the frequency of the term's appearance.

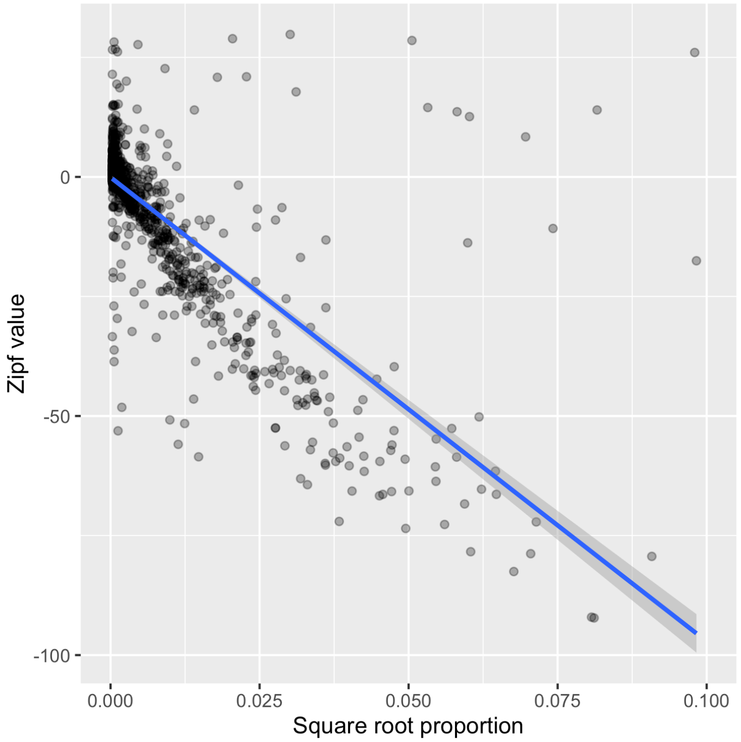

Figure 8.3: The relationship between Zipf values and term proportion linearizes when plotted against the square root of proportion.

A frequency-only model reduces error better than more complex models

We wanted to compare the predictive accuracy of these additive models against the term frequency model described in subsection 8.3. To accomplish this, additive, full factorial, and proportion-only models were calculated for the bigram subset of the data. For each model, root mean square error (RMSE) and median absolute error (MAE) were calculated. The model that only used relative proportion minimized prediction error better than both of the models that include information about the component unigrams (Table 8.6).

8.4 The frequency effect is robust

The effect is not driven by stop words

It remained possible that the term frequency effect was being driven by `function words`, which have high relative proportions but are not considered to bear much information. To

	RMSE	MDAE
additive.unigrams	2.54	1.00
factorial.unigrams	2.34	0.89
tf.only	2.23	0.88
additive.stopwords	2.54	1.01
factorial.stopwords	2.32	0.86

Table 8.6: A prediction equation that only uses term frequency performs better than models that include information about how informative the component unigrams are.

test this possibility, the English stop words corpus implemented in NLTK was used as a proxy for function words [**nltk**]. The entire analysis was re-run with stop words removed, and the same pattern of results remained with a single exception – in the initial model, removing stop words causes the effect of gram length to become statistically insignificant (Table 8.7). Notably, the stop words comprised the majority of the high leverage cases in the upper right portion of Fig 8.3, and their removal nearly doubles the coefficient of term proportion from -780 to -1300. This is surprising both in how much it improves the model (from an R^2 of 0.53 to 0.70) and in how much semantic value is afforded function terms by the Zipf statistic.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3695	0.4978	2.75	0.0060
sqrt(tf)	-1308.3610	23.8950	-54.75	0.0000
len	-0.1231	0.2590	-0.48	0.6346

Table 8.7: Removing common function terms does not change the pattern of results, except that now the contribution of the number of unigrams is no longer significant.

As a further test, a new variable was created that represented the number of stop words in each bigram. Whether used by itself or adding it to the simple additive model, including the number of stop words did not create a model that performed better than term frequency alone (see Table 8.6). A full factorial model performs slightly better for MAE, but this must be balanced against the consideration that this model uses more than three times the number of predictors than the simple proportion model, and may be reflecting overfitting.

The effect is not driven by the Zipf statistic

It remained possible that the term frequency effect was specific to the Zipf statistic itself and would not be reproducible with other methods of calculating semantic value. To test for this possibility, cosine similarity was calculated for each term in the data set against the frequency vector for the entire corpus, and the analysis was re-run.

Because this constitutes a part/whole relationship, the cosine values were uniformly positive and tended to be close to 1, with the top three quartiles of values falling at or above 0.9. This made the relationship very difficult to linearize[6]. Furthermore, Zipf and cosine do not seem to measure semantic distance in quite the same way, as there is no overlap in the the 10 most informative terms in the corpus as calculated by Zipf and by cosine similarity.

Nevertheless, the cosine similarity results recreate the same basic pattern as the Zipf statistic results: relative proportion of terms is inversely correlated to informativeness, and gram length is not a good predictor (Fig 8.4) [7].

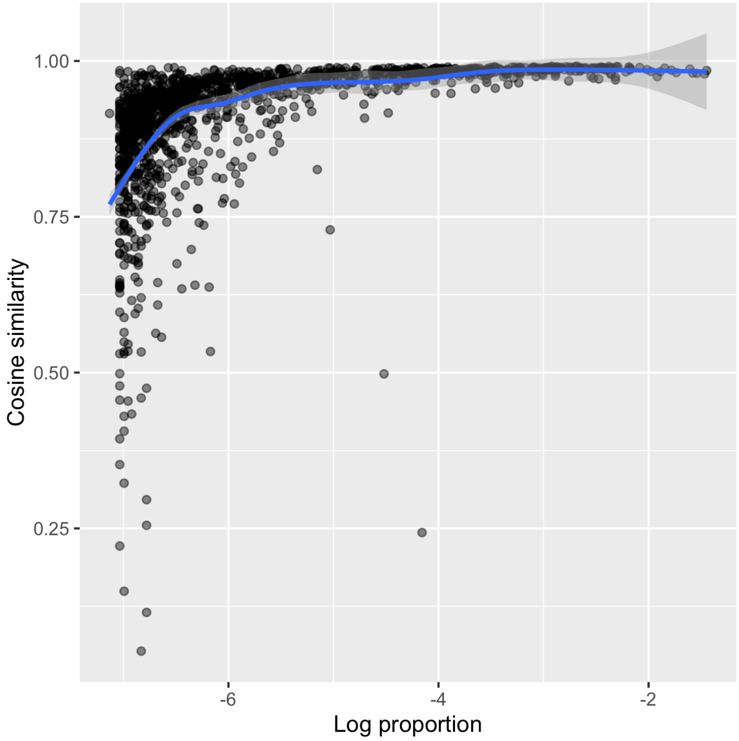

Figure 8.4: Plotting cosine similarity between term and corpus to the 50th power plotted against log term proportion shows the same monotonic, inverse relationship.

The effect is not driven by using count frequency

In the same way that these results could have been an artifact of the metric, it was also possible that these results were driven by some factor that had to do with the power law

[6] ultimately requiring a power transform to the fiftieth

[7] Because cosine is a measure of similarity and Zipf is a measure of dissimilarity, equivalent results means opposing slopes

governing the distribution of terms, and not with the actual informativeness of those terms. Up until this point, it also would have seemed reasonable to conclude that less frequent terms were more informative because they were only used in specific contexts, but only because this fits with our intuitive beliefs about language and not because we have estimated the specificity of terms explicitly.

To address both of these concerns, we computed a metric to capture the generality of term usage. If a term is only used in specific situations, you would assume that the mean in its frequency of use would be much smaller than the variance in its frequency of use. In effect, this is the normalized moment of each term (see equation 8.3), and is providing a conceptually similar to although mathematically different from the metric employed in [**piantadosi˙tily˙gibson˙2012**].

To put it concretely, if there is a discussion of modern metal music, you would expect the term *Nightwish* to be mentioned frequently in that context, but almost never in any other conversation. Since discussions of modern metal music comprise a small portion of all discussion topics, the average daily appearance of *Nightwish* will be close to zero, and so will its generality. On the other hand, a term that is not specific to a discussion topic like *well* will have an average rate of appearance that is much greater than the variance in its appearance, and this will have a generality metric greater than one. Indeed, *Nightwish* has a generality of 0.14 in our data set, and *well* returns a measure of 26.5.

$$\frac{mean(dailyfrequencycounts)}{sd(dailyfrequencycounts)} \tag{8.3}$$

The relationship between generality and term frequency linearizes as a negative square, which could be thought of as something like the context-specificity of a term (Fig 8.5). While the output is noisier, it is evident that the Zipf value of a term decreases monotonically with its specificity. The analysis was re-run with this specificity measure, and the same pattern of results remains, at roughly the same R^2 of 0.60.

The effect is not driven by the corpus

Finally there was a risk that these results were driven by a kind of test-retest overfitting. An unpublished result from another laboratory showed a failure to replicate when the corpus used to calculate pointwise mutual information and the corpus used to generate term frequencies were different [**doyle˙2015**]. To control for this possibility, the subset of unigrams was taken from the overall data set and term frequencies for each term in this reduced data set were retrieved from the Microsoft Research Ngram Corpus [**microsoft**]. Relative proportion remained a significant predictor (Table 8.8).

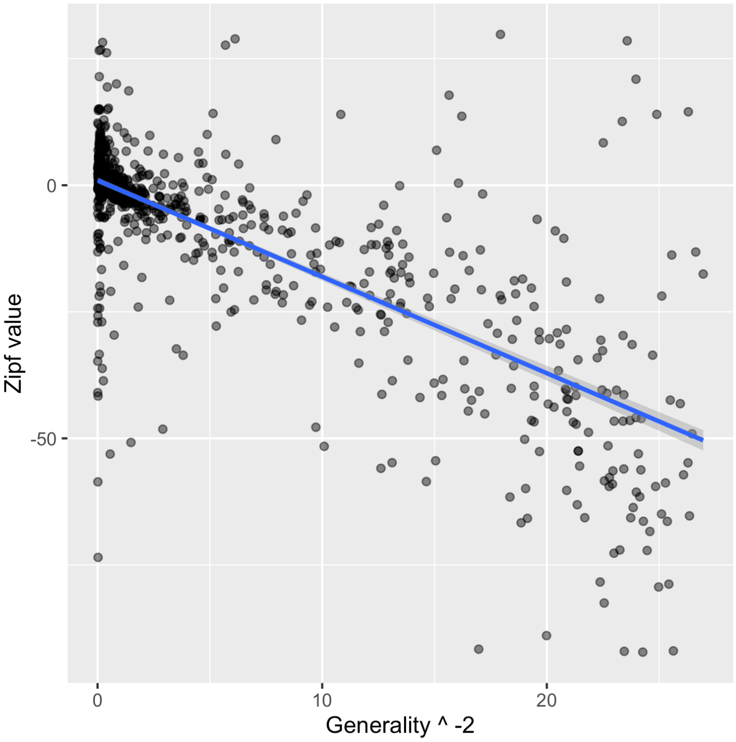

Figure 8.5: Context specificity plotted against Zipf values shows a monotonic decrease in semantic value with decreasing specificity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13.1228	1.1295	-11.62	0.0000
sqrt(p)	-589.1914	52.7313	-11.17	0.0000

Table 8.8: Replacing the corpus for deriving term frequencies does not remove the effect of term frequency.

8.5 Conclusion

Our initial observation that the informativeness of a term decreases with its frequency in a corpus agrees with previous work using language models and human judgments [**piantadosi˙tily˙gibson˙20** **lewis˙frank˙2016**]. In Lewis and Frank, the dependent measurement was complexity, and the independent measure was orthographic length. In Piantadosi, Tily, and Gibson, it was the probability of being associated with a specific context, with the independent measure being one of a variety of estimates for how difficult the word is to produce (long, infrequent, etc.). We believe that this shows the broad reach of this effect.

We differ from previous work in our hypothesis about the nature of the causal connection between the two. To be explicit, Piantadosi argue that words that are easy to produce because they are common or short are applied to new contexts in lieu of generating a new word, and thus become less meaningful [**piantadosi˙tily˙gibson˙2011**]. In contrast, we have argued that words are already associated with contexts, and that the modification of a word to become easier to pronounce should happen only after the word acquires a new context (and thus becomes more common). In other words, a word is not informative because it has a definition that causes it to be used rarely, but rather a word is a reference to a peculiar set of contextual elements because it has only been used in the presence of those elements, which is later reflected in its definition. With the increasing availability of large amounts of timestamped linguistic data, testing this difference in hypothesized order of events is a tractable question.

This observed relationship seems to disagree with a set of findings from studies examining how words are applied to items in a category [**regier˙kemp˙kay˙2015**]. In these results, the most common words are also those with the most specific meaning. *Brother*, for example, has a more restricted meaning in the English kinship semantic space than *grandparent*, both because the latter is gender neutral and because it can refer to people in multiple lineages of the family. However, *grandparent* is more semantically valuable, according to our definition, because it almost certainly predicts further communication specifically about kin or genealogy, possibly including information about health status. In contrast, *brother* can refer to non-related males (as in *guy* or *dude*), objects that appear to be similar (as in *twin* or *clone*), or, by happy circumstance, a particular kind of printer.

Additionally, we have provided one piece of evidence against the principle of compositionality. Specifically, we have shown that the semantic value of bigrams is not predicted by the semantic value of its constituent unigrams. Furthermore, we have shown that the square root frequency of a term is a better predictor of its semantic value than its length, whether or not it is considered a stop word, the semantic value of its unigrams, or any combination of these factors. The question remains as to why some combinations appear to construct their semantic value in a combinatorial manner. Perhaps in the case of *brown dog*, the apparent combinatorial relationship is reflecting the nature of the event and not the underlying symbolic relationship.

We have further demonstrated that the relationship between semantic value and frequency is not an artifact of the metric. Cosine similarity, when applied in a document-corpus manner, gives a similar pattern of results even the exact shape of the distribution it produces is not the same. This result is also not simply dependent on term frequency, as another metric of rarity, the normalized moment, produces a similar result. Finally, removing common function words, which theoretically could have been driving the effect, actually makes it stronger.

We have endeavored to show a potential use of the Zipf statistic to produce a robust finding concerning semantic value. However, we anticipate that this method for computationally determining the semantic value of ngrams could be useful for testing any hypotheses that posit an effect of **meaningfulness**, such as:

- diffusion rate of new terms
- likelihood of adoption of loan words
- loss rate of semantic value (i.e. bleaching)
- likelihood of substitution during dyadic interactions.

This metric also creates the intriguing possibility of automatically detecting euphemism, colloquialism, and dialect by comparing terms by Zipf value across different corpora.

Chapter 9

Future Directions

We have seen thus far that the metaphors used to explain humans and their behavior has changed over time, often concurrently with advances in other technological fields. The currently popular computational metaphor has succeeded in motivating the analysis of some aspects of human behavior, but fails in others. Notably, these include the extra-social, multimodal, and self-directed aspects of human language use. These difficulties can be avoided by instead adopting an organizing metaphor of language as performance of phonatory culture. We have conducted three experiments here that offer evidence in favor of this perspective.

The first prediction we make here is that human behavior is not generative, and therefore human language will not be as varied as would be expected were it produced by a generative system. We developed a novel method for estimating the number of ecologically possible (as opposed to theoretically possible) sentences in English. This metric returned an estimate for the number of possible English sentences that is several orders of magnitude smaller than a theoretical estimate.

The second prediction we make is that rarely performed behaviors will bear more information for an observer than more frequently performed behaviors. We derive a novel method for estimating the semantic value of arbitrarily long sequences of English words, and compare that value to the relative proportion of the appearance of a large number of unigrams, bigrams, and trigrams. The data suggest that there is an inverse relationship between meaningfulness and commonality; and, furthermore, that this relationship is independent of and stronger than the number of individual words in the sequence.

The third prediction we make is that the symbolic reference system in human performance cannot be decomposed into individual, separate pieces, without some loss of information. Using our method for estimating the semantic content of terms, we show that it is not possible to predict the semantic value of a sequence of words based on their constituent terms.

This organizing metaphor suggests different interpretations of common observations around natural language. A particularly interesting observation in this context is that linguistic models that specifically segregate language into higher order grams perform better in machine learning contexts that pipelines that only use individual term counts. The typical explanation is that including bigrams and trigrams is a way of incorporating some kind of grammatical information into the model, but these results suggest that trigrams work better because they are qualitatively different in their ability to capture the symbolic components of speech acts.

The results presented here suggest future lines of research focused around the idea of symbols and performance in the processing of naturally generated language. One possible direction is to investigate the extent to which current language models are improved by using non-overlapping, variable word-length formulae, much in the same way such models currently allow the use of non-overlapping, variable character-length words. In a similar vein, it would be interesting to investigate whether incorporating viewpoint and audience improve language models that already exist [1]. Another would be look for evidence that the size of the semantic space in any given language is related to the size of the community of speakers of that language. Or, in other words, to look for evidence that the total number of sentences or ideas is related to population size.

[1] Some of this work is being done already, see [**bamman˙smith˙2015**]