Chapter 4: Cycles

Christopher Ahern

> [T]o say what a word means in a language is to say what it is in general optimal for speakers of that language to do with that word, or what use they are to make of it.
> –H.P. Grice (1989, 299)

Distinguishing between the formal and functional Jespersen cycles clarifies what needs to be explained. The functional cycle can occur independently of the formal cycle, so we need an explanation for it regardless. This motivates our focus in this chapter on the functional cycle. That is, we want to know why one form displaces another, taking over the meaning of plain negation. We also want to know why this new form can be displaced in further functional cycles. In particular, we want to explain why in the history of English we observe emphatic *ne...not* increase in frequency and displace pre-verbal *ne*. To do so, we build a mathematical model of the pragmatic pressures that lead to this transition.

First, we provide an interpretation of the notion that the incoming form is *emphatic* with respect to the incumbent form, that it conveys some special meaning. Namely, we note that the incoming form is initially restricted to contexts where the proposition being negated has just been introduced into the discourse, but expands to contexts where it is merely inferable from the discourse, and eventually to contexts where the proposition is brand new to the discourse. Second, we discuss experimental evidence that suggests why the incoming form spreads across contexts in the way it does. Speakers have difficulty in separating out their own private knowledge from what is common knowledge between themselves and their interlocutors. Given this difficulty, speakers adopt a heuristic that biases them towards their own perspective when assessing how closely connected a negated proposition is to the discourse. This leads speakers to treat propositions as more connected to the discourse than is warranted. Third, we incorporate these facts into a signaling game, determine the equilibria and dynamics, and show how the number of signals used interacts with speaker bias. Finally, we fit the resulting model to data from the functional cycle in Middle English and discuss the implications of the fitted parameters in light of the experimental evidence.

The main contributions of this chapter are twofold. The first contribution is that we offer the first explicitly dynamic model of the functional cycle that explains why the discourse constraints change in the manner that they do. While previous accounts have noted the constraints on the different forms of negation in the functional cycle, they have not explained the increase in the incoming form beyond the somewhat circular claim that the incoming form is overused. Here we argue that the functional cycle is a byproduct of our cognitive limitations in tracking common knowledge. Importantly, while the the driving force of the cycle is rooted in the cognition of individuals, it leads to change because of the social interactions between individuals in a population. In Grice's terms, the optimal use of and response to different forms of negation are both moving targets. Their coupled movement is what underlies the functional cycle.

The second contribution of this chapter is that the model we present offers an information-theoretic foundation for grammaticalization. For example, the intuitive notion of *bleaching* is simply the loss of information carried by a signal as it becomes obligatory. Grounding qualitative terms in this quantitative framework offers a new perspective on diachronic changes in how meaning is signaled. Moreover, it allows for a broader conception of meaning as the information carried by linguistic signals. This broader conception of meaning has the potential to unify our analyses of semantic, pragmatic, and sociolinguistic meaning.

1 Emphasis

We begin with the intuitive notion of *emphasis*. We know it when we hear it, but what it amounts to is often left implicit in accounts of the functional cycle. There are two general functions that have been suggested as candidate interpretations. The first is that emphatic negation widens and strengthens negation to preclude exceptions (Kadmon and Landman, 1993). That is, emphatic negation signals a stricter *standard of precision* for interpreting a proposition (cf. Austin 1962; Lewis 1970; Landman 1991; Krifka 1995). The second interpretation is that emphatic negation serves to deny an expectation, or mark the contradiction of a potentially implicit assertions (Detges and Waltereit, 2002; Kiparsky and Condoravdi, 2006). That is, emphatic negation signals the relationship between the proposition being negated and the preceding discourse. In what follows we focus exclusively on this second interpretation, but return to the first later on as an important point of comparison.

Broadly speaking, this second interpretation of emphasis depends on the joint attention of speakers and hearers (cf. Chafe 1974; Prince 1981). More precisely, emphatic negation has been found to signal that the proposition being negated is *activated* (Dryer, 1996; Schwenter, 2005, 2006). A proposition is *directly activated* if its contents have just been explicitly introduced into the discourse, so it is present to the joint attention of speakers and hearers. A proposition is *indirectly activated* if its contents can be inferred from the discourse either via an entailment or implicature. A proposition is *non-activated* if its contents have not been explicitly introduced into the discourse or it cannot be inferred from the preceding discourse.

There are two important things to note about activation. First, activation does not entail belief, nor vice versa. Participants in a discourse believe propositions that are not activated, and not all of the activated propositions are believed. This distinguishes activation from the notion of *common ground* (Stalnaker, 1978), which consists of the set of propositions that both interlocutors believe to be true, or at least accept as true for some purpose (Stalnaker, 2002, 715-720). Second, and related to this first point, emphatic negation can be used both to negate an activated proposition or restate it. For example, if the activated proposition is p then negation can be used to negate that p. If the activated proposition is $\neg p$ then negation can also be used to restate that $\neg p$.

Crucial for our purposes is the fact that activation can reasonably be identified in historical corpora with the use of modern translations. Moreover, activation has been shown to have the same effect diachronically in the histories of French (Hansen, 2009; Grieve-Smith, 2009), Italian (Hansen and Visconti, 2009, 2012), and English (Wallage, 2013): the incoming emphatic form in all of these languages is initially restricted to use with either activated or directly activated propositions. For example, Wallage (2013) shows that in Early Middle English *ne...not* is overwhelmingly restricted to negating activated propositions. But, over time it spreads to negating non-activated propositions as well.

The following examples from the *Penn Parsed Corpus of Middle English* (Kroch and Taylor, 2000) cited by Wallage (2013) demonstrate the transition. First, the emphatic form is restricted to contexts where it is used to negate directly activated propositions. Where $p =$"They are deceived", emphatic negation can be used to deny the explicitly stated proposition that p. In all of the examples that follow, p and $\neg p$, or both are bolded in the translation of the passage.

(1) Alle ðo men ðe swinkeð on ðessere swinkfulle world, alle he swinkeð for sumere
All the men that labour in this toilsome world, all they labour for some
hope ðe hie habbeð, ðe hem oft eaten ande beswinkð ... Ac ðo ðe swinkeð for
hope that they have, that them often at end deceives ... But those that labour for
ðessere eadi hope, hie *ne* bieð *naht* becaht
this blessed hope, they NEG are not deceived.
"All the men who labor in this toilsome whorld, they all labor for some hope they have which often **deceives them** in the end...But those who labor for this blessed hope, **they are not deceived**."
(CMVICES,33.385, 1200 CE) The incoming form can also be used to restate a proposition that has already been explicitly negated. For example, where $\neg p =$"You don't know yourself", it can be restated.

(2) If u ne cnawest e seolf ... If u *ne* cnawest *naut* e seolf
If you not know the self ... If you NEG know not the self
"If **you do not know yourself**...If **you do not know yourself**"
(CMANCRIW, II.80.941-948, 1230 CE)

Subsequently, the use of the incoming form is extended to being used to negate propositions that

are only indirectly activated. For example, we might suppose that a virtuous religious rite with all the sin-cleansing properties of baptism would have some post-mortem benefits. If we take the proposition resulting from this inference to be $p =$"[That rite] opened to them the bliss of heaven", then the incoming form can be used to negate the proposition resulting from the inference.

(3) and te lage hadde to alle te mihtes te haueð nu fulluht for ðat clensede te man
and the law had then all the virtues that has now baptism for that cleansed the man
of sinne: swa doð nu fulluht ac it *ne* openede hem *noht* te blisse of heuene alse
of sin: as does now baptism but it neg opened them not the bliss of heaven as
fulcneng doð us.
baptism does us.
"And **that rite had then all the virtues which baptism now has**, for that cleansed man of sin even as baptism now does, but **it opened not to them the bliss of heaven** as baptism does to us."
(CMTRINIT, 87.1165, 1225 CE)

Similarly, the incoming form can also be used to state a negative inference. If we suppose that renouncing one's sins requires being done committing them, $\neg p =$"I cannot renounce my sins", then the incoming form can be used to state the proposition resulting from this negative inference.

(4) Ich nam noht giet sad of mine sines and forti *ne* mai ich hie *noht* forlete.
I not-am not yet sated of my sins and therefore neg can I them not renounce
"**I am not yet sated of my sins** and therefore **I cannot renounce them**."
(CMTRINIT, 75.1028, 1225 CE)

Finally, the incoming form can be used to negate propositions that are entirely new to the discourse. That is, *ne...not* is used to negate a proposition that is not readily identifiable either directly or indirectly from what has come before. It is useful to note that all of these examples come from roughly contemporaneous documents.

2 Experimental evidence

While we observe this trend in historical corpora of English and other languages, this does not explain why the incoming form spreads across the contexts in the way it does. To understand why, it is useful to consider experimental evidence demonstrating particular communicative biases on the part of speakers. Namely, speakers' private knowledge persistently influences how they signal meaning.

For example, Wu and Keysar (2007) had pairs of participants play a communication game in fixed roles of speaker and hearer. Speakers and hearers jointly learned names for a set of abstract shapes. Speakers then learned several names for additional shapes privately. The experimenters varied the number of shape names that the speaker and hearer learned together, what the exper-

imenters called the *informational overlap* between participants. The participants then played a game where the speaker directed the hearer to select a target shape from a set. Across trials the target shapes were evenly distributed between shapes whose names were learned together, shapes whose names were learned privately by the speaker, and shapes that were new to both participants. Presumably, using a name to refer to a shape is only felicitous if the name of the shape was learned together by both speaker and hearer. But, surprisingly, in trials where the target shape's name was private knowledge, the name was the first thing speakers said in 5% of trials where there was a low informational overlap and in 28% of trials where there was a high informational overlap. That is, speakers relied on their private knowledge more where there was a greater degree of informational overlap.

Note that this use of the privately known names was not a result of speakers forgetting the context in which the names were learned. Heller et al. (2012) replicated these findings and showed that speakers were incredibly accurate at distinguishing between names learned together versus names learned privately. Rather, Wu and Keysar (2007, 4) suggest that these results point to speakers using a kind of *overlap heuristic*: when the informational overlap between yourself and your interlocutor is sufficiently extensive act *as if* they have all the same information as you. In fact, this is kind of combined *co-presence* and *community membership heuristics* proposed by Clark and Marshall (1981) to resolve the problem of *common knowledge*: that everyone knows that p, that everyone knows that everyone knows that p, *ad infinitum*.[1] This heuristic and the speaker bias that it creates are a means for solving the difficult task of keeping track of what is private versus common knowledge. Assuming that an interlocutor knows roughly the same things about the world reduces the cognitive burden and simplifies things a great deal.

However, given that speakers' use of private knowledge varied across conditions, we might wonder whether this bias is specific to certain contexts. For example, speakers might be able to pay closer attention to the discourse and keep better track of things. That is, there might be two modes of thinking regarding the discourse (Keysar et al., 2003; Kahneman, 2011). However, this bias is not subject to conscious manipulation. Wardlow Lane et al. (2006) had pairs of participants play a communication game in fixed roles of speaker and hearer. Four shapes of varying size and color were presented to the participants. One shape was visible to only the speaker, blocked from the view of the hearer by an occluder. Speakers were instructed to communicate information about a target shape visible to both participants so the hearer could identify it. In the test conditions, the item that was visible only to the speaker was the same shape as the target item, but varied along some relevant dimension (e.g. size, color). For example if the target shape was a blue triangle then the shape that was only visible to the speaker was a green triangle, and none of the other shapes visible to both participants were triangles.

Wardlow Lane et al. (2006) found that speakers modified their description more in the target

[1]This technical term was introduced into the Philosophical literature by Lewis (1969), but has a long history under various names. Whereas Clark and Marshall (1981) use the term "mutual knowledge" to refer to common knowledge, as it is commonly used *mutual knowledge* only requires that everyone know that p. Thus anything that is common knowledge is also mutual knowledge, but not *vice versa*.

condition. That is, speakers were more likely to say "The blue triangle" to refer to the target shape if there was a green triangle visible only to them. Speakers' private information leaked into what they said and how they said it. This happened even though speakers had direct evidence that only they could see the shape that contrasted with the target shape. In fact, this over-modification happened to an even greater extent when speakers were explicitly instructed to conceal information about their privileged information. Speakers have a difficult time inhibiting their perspective even when they want to, suggesting that speaker bias is persistent fact about communication.

These experimental results show how speakers' private knowledge persistently influences how they signal meaning. In particular, they show that speakers tend to assume that their interlocutors are overwhelmingly similar to them. While this experimental evidence deals with the referential domain, the results can be extrapolated to the propositional domain and activation in particular. That is, activation is defined by the joint attention of speakers and hearers. But, neither speakers nor hearers know whether or not a proposition is actually being attended to by an interlocutor. This problem is a perfect candidate for solution by the kind of heuristic described above. Namely, speakers can simplify the problem by assuming that what their hearers attend to is sufficiently similar to what they attend to.

One potential concern about this kind of heuristic is that mistakes seem unlikely. If a proposition is directly activated, then it is directly activated because it has just been mentioned. Similarly, if a proposition is indirectly activated, then it is indirectly activated because there is an entailment or implicature that does so. A response to these objections, particularly the first, lies in the nature of activation. While we have described discrete categories, Dryer (1996, 481-482) rightly conceives of both direct and indirect activation as continuous measures. For example, while the utterance of a proposition directly activates it, this activation decays over time. As he puts it, the proposition is *deactivated* as time goes on and it passes from the joint attention of speakers and hearers. Similarly, while the utterance of one proposition may indirectly activate another via an inference, some propositions may be more *accessible* than others via inference. Some inferences are natural, whereas others are non-sequiturs. In fact, we might take activation as a whole to be constituted by degrees of inferability. Recently uttered propositions are high on the scale, whereas brand new propositions are at the bottom end.

So, speakers might treat a proposition as more activated than is warranted for several reasons. For example, suppose that a speaker keeps thinking about a proposition p, but her interlocutor does not. This means that p is being deactivated as it has passed from the joint attention of both participants. However, the speaker still dwelling on p is not aware of this, and only has her own perspective to consider. Thus, she may still treat p as highly activated even though it is not. Similarly, a speaker may easily infer a proposition p from the preceding discourse due to her attention to particular aspects of the discourse. But, her interlocutor may only make the same inference with great effort. Thus, a speaker may treat p as highly activated even though it is not.

If speakers have a tendency to overestimate activation and use *ne...not* more than is warranted by the actual degree of activation, then how will hearers respond? Given hearers' response, how will speakers respond in turn? What does this mean for the functional cycle? To answer these

questions we translate these experimental findings into a mathematical model that can be used to investigate meaning over time. This model consists of two components. First, we define the *stage game* that describes the interactions between speakers and hearers and captures speakers' bias towards overestimating activation. Second, we define the *game dynamics* that describe how a population of speakers and hearers change over time while playing the stage game.

3 Signaling

We start by defining the components of the stage game that we will use to analyze the functional cycle. We define the states, messages, and actions along with their interpretations. We then turn to the utility function of senders and receivers as they relate to the preferences of speakers and hearers. Once we define the game we can determine its equilibria and the dynamics of how speakers use different forms of negation to signal the activation of the proposition being negated.

First, let the set of states $T : [0, 1]$ be the degree of activation of the proposition being negated, where $t = 0$ indicates a brand new proposition and $t = 1$ indicates a proposition that was the last thing uttered. Second, let the set of messages that the speaker sends be a finite set $M = \{m_1, m_2\}$. We can think of these as the incumbent and incoming form in the functional cycle respectively. So, for English, m_1 is *ne* and m_2 is *ne...not*. Finally, let the set of actions $A : [0, 1]$ be the action taken by the hearer to interpret the message. For example, an action a_i can be thought of as an initial guess by the hearer about the level of activation t_i of the negated proposition.

With these components defined, it is important to make a conceptual clarification about the role of states as degrees of activation. In signaling games, the state is taken to be some piece of private information that the sender has about the state of the world. This is not quite accurate given that we assume that speakers never know the actual degree of activation. That is, they can never peer inside hearers' heads to verify what propositions are being attended to at any given moment. At best, speakers have a subjective estimate of the degree of activation. However, this subjective estimate is systematically related to the actual degree of activation.

To see this, suppose that both speakers and hearers have some subjective estimate of the activation of a proposition, call them t_S and t_R. For example, both might take their estimate to be the approximate amount of attention they are paying to a given proposition. Now, given that activation is defined in terms of the joint attention of speakers and hearers, then the actual state of activation t must be some function of the two subjective estimates. If a speaker's estimate of a proposition is that it is not activated $t_S = 0$ because she is not attending to it, then by definition it is not activated $t = 0$. In contrast, if a speaker's estimate of a proposition is that it is highly activated $t_S = 1$, this does not mean that it is indeed activated since the hearer's estimate could be lower $t_R < 1$ because the hearer is not attending to it to the same degree.

A simple way of capturing the relation between the subjective estimates and the actual degree of activation is that $t = min(t_S, t_R)$. The actual level of activation is the highest degree that both participants would agree to based on their own subjective estimates. This is arguably what

underpins our intuition that a form is infelicitous because a hearer did not or could not have had a sufficiently high subjective estimate of activation. If this is the relation between the subjective estimates and the actual state, the speaker's estimate stands in a particular relation to the actual state. Namely, the speaker overestimates the degree of activation. To see why, note that if $t_S \leq t_R$ then $t_S = t$, and if $t_S > t_R$ then $t_S > t$. Together these imply that $t_S \geq t$. In other words, on average the speaker overestimates the actual degree of activation.

Now, if speakers only have access to their own subjective estimates of the state of activation, then it makes sense that their preference are determined by that estimate. As far as the speaker is concerned, t_S is the actual degree of activation of the proposition being negated. It makes sense then that the speaker would want the hearer to infer that degree of activation. In particular, we suppose that speakers prefer for hearers to infer the degree of activation that is closest to their subjective estimate. The following utility function satisfies this constraint, it is maximized exactly where $a = t_S$.

$$U_S(t, a) = 1 - (a - t_S)^2 \tag{1}$$

While this function captures the speaker's preferences, it also introduces a new and undefined parameter, the speaker's subjective estimate. A simple way to address this is to posit a general functional shape $f(t) = t_S$ that is defined in terms of the actual state and other parameters but captures speakers' tendency to overestimate the actual state by guaranteeing that $f(t) > t$. There are infinitely many functions that satisfy this constraint, but a particular simple functional form has a natural interpretation in our case.

In their seminal work on signaling games of *information transmission*, Crawford and Sobel (1982) introduce a bias parameter $b \geq 0$ into the utility function of senders that indicates how aligned the goals of senders and receivers are. For example, where $b = 0$ their preferences are perfectly aligned, but for $b > 0$ they diverge. We can apply this directly to our case if we think about the bias parameter b as speakers' tendency to overestimate activation. Then the following simple linear function allows us to incorporate the tendency to overestimate into the speaker's utility function.

$$t_S = f(t) = t + (1 - t)b \tag{2}$$

This yields the following utility function which satisfies the constraint that $t_S \geq t$.[2] It is maximized for an action $a = t + (1 - t)b$. This means that speakers prefer that hearers take an action slightly higher the actual state of activation.

$$U_S(t, a) = 1 - (a - t - (1 - t)b)^2 \tag{3}$$

[2]This differs from the formulation in Crawford and Sobel (1982) where $U_S(t, a) = -(a - t - b)^2$. Their form allows senders to prefer actions $a > 1$, which has no interpretation in our model. That is, the degree of activation and interpretation of the degree of activation are both constrained to the unit interval. We also add a constant so that all payoffs are positive for the dynamic analysis in the next section.

By changing the variables we obtain a functional form that depends only on states, actions, and the new bias parameter. In fact, this bias parameter has a natural interpretation in terms of speaker bias as a measure of how good or bad speakers are at keeping track of common versus private knowledge. The case where $b = 0$ corresponds to speakers developing the ability to read minds and accurately assess the actual state of activation. For $b > 0$ speakers have a tendency to overestimate the degree of activation and prefer a higher action on the part of hearers.

Now, we might wonder whether speakers have access to the action taken by hearers. That is, if speakers cannot peer into the heads of hearers to determine the state of activation, does it makes sense to assume that they can somehow infer the reasoning process indicated by the action hearers take to interpret the message? However, speakers have incredibly rich sources of feedback from hearers. For example, this feedback includes the amount of time hearers take to respond, facial expressions, and backchannel cues (e.g. "Mhmm" versus "Huh?"), as well as verbal responses such as as requests for clarification or continuations of the discourse. All together then, it seems reasonable that speakers can recover the action taken by hearers.

With speaker preferences defined, we can think about hearers. Suppose that hearers want to accurately infer how the proposition being negated relates to the prior discourse. For example, a hearer does not want to overestimate the degree of activation of a proposition and expend too much effort on trying to discern how it is connected to the discourse. Similarly, a hearer does not want to underestimate the degree of activation and miss out on information regarding how a proposition fits into the discourse. The following utility function satisfies this constraint, it is maximized exactly when $a = t$.

$$U_R(t, a) = 1 - (a - t)^2 \tag{4}$$

That is, hearers do best when they accurately infer the actual degree of activation. Now, we might wonder again whether it is reasonable to assume that hearers have access to the actual degree of activation. They cannot read minds any more than speakers. However, hearers gain information as they reason about different potential degrees of activation.

To see why this is the case, first assume that speakers and hearers have the same reasoning capacities. That is, they would agree on what potential degrees of activation make some kind of sense given the discourse. So, we assume that speakers and hearers can both identify a particular set of degrees of activation as making sense. Note that this does not require that speakers and hearers expend the same amount of effort in identifying the degree of activation, but rather that speakers and hearers have the same reasoning capacities. Here we will suppose that there is some function of states given the discourse that defines whether both speakers and hearers can identify a particular degree of activation. Namely, let $g(t)$ be a convex function such $g(t) = 1$ if and only if t can be identified by both speaker and hearer. Further, suppose that the lowest identifiable degree of activation corresponds to the actual state of activation, for all $t_i < t$, $g(t_i) = 0$. This simply means that the subjective estimates of speakers and hearers serves as a lower bound on what degrees of activation both speakers and hearers can reasonably identify.

Now, suppose that hearers reason in the following manner when they take an action to interpret

the speaker's message. The hearer takes an action a_i, corresponding to an initial guess that the degree of activation of the proposition is t_i. By the same reasoning for speakers above, hearers overestimate the actual degree of activation $t_R \geq t$, so it makes sense that they would compensate for this by choosing an action such that $t_i < t_R$. Given this action, there are two possibilities. The degree of activation for the proposition could be identifiable or not. If the initial guess of the degree of activation is identifiable $g(t_i) = 1$, suppose that the hearer reasons about lower and lower degrees of activation; likewise, if the initial guess of the degree of activation is not identifiable $g(t_i) = 0$, suppose that the hearer reasons about higher and higher degrees of activation. In both cases the hearer will consider a degree of activation t_j such that $g(t_j) = 1$ and for all $t_k < t_j$, $g(t_k) = 0$. That is, she will eventually recover the actual degree of activation. Note also that the amount of effort put into finding the actual degree of activation grows with the distance between the initial guess and the actual degree. So, if speakers and hearers reason in a sufficiently similar manner, then hearers will be able to recover the actual state of activation.

The preferences we have defined for speakers and hearers are a way to represent the experimental evidence we described in the previous section. That is, they allow us to state, in mathematical terms, the stage game which is the shape of the interaction. But, this shape by itself does not make any predictions about the behavior of individual speakers and hearers playing the game or the trajectory of a population of speakers and hearers over time. To understand these we need to determine the equilibria of the stage game and how a population changes while playing the stage game under a particular game dynamics.

4 Equilibria

Now that we have defined the components of the game, we turn to analyzing its properties. Determining the equilibria of the game we described will allow us to address several questions about how signals are used in the functional cycle. Broadly speaking, we want to know two things. First, we want to understand the relationship between senders' bias towards overestimating activation and the use of signals at equilibrium. In particular, we want to know how large this bias can be while still allowing for multiple forms to be used at equilibrium. Second, we want to know whether particular equilibria constitute evolutionarily stable strategies. If so, we want to know the strategies that are evolutionarily stable. If not, we want to know what strategies could invade the population and disturb an equilibrium. Broadly speaking, this amounts to determining the conditions for the functional cycle to occur. We begin by defining speaker and hearer strategies, the expected utility of different strategies, and then determine the evolutionarily stable strategies of the game.

The set of speaker strategies is all potential mappings from the unit interval to a discrete set $S : [0, 1] \rightarrow M$. This is problematic given that the domain is uncountable. To simplify things we consider the following condensed representation. Let $\mathcal{P}_n(T)$ be a partition of the state space into n equal length subintervals $t_0 = 0 < t_1 < ... < t_{n-1} < t_n = 1$. For each properly defined subinterval, (t_{i-1}, t_i) the sender uses the message m_i. A speaker's strategy is then a function from

this partition to messages $S : [\mathcal{P}_n(T) \rightarrow M]$. Intuitively, this is simply a way of carving up the state space into discrete contiguous regions and using those regions to determine which signal to send. For example, consider the case of two messages $\mathcal{P}_2(T)$, where m_1 and m_2 correspond to *ne* and *ne...not* respectively. For $t \in (0, t_1)$ a sender will use *ne* and for $t \in (t_1, 1)$ a sender will use *ne...not*.

In fact, this kind of threshold strategy is extremely close to what we observe in historical data. That is, as the functional cycle proceeds, the incumbent form *ne* is not evenly distributed across activation contexts, but largely negates non-activated propositions. The probability of using the two negative forms is overwhelmingly conditioned by the activation of the proposition being negated. For example, for Midddle English from 1150-1250 CE the conditional probabilities are the following (Wallage, 2013, 12)

$$p(\textit{ne} \mid \text{NON-ACTIVATED}) = .85 \tag{5}$$

$$p(\textit{ne...not} \mid \text{ACTIVATED}) = .84 \tag{6}$$

This means that specifying speaker strategies in this manner is both a useful and empirically accurate abstraction. Interestingly, this also gives some empirical credence to the push-chain scenario conception of the functional cycle; *ne* actually appears to be pushed down the scale of activation by *ne...not*.

The set of hearer strategies is all potential mappings from the set of messages to the unit interval $R : M \rightarrow [0, 1]$. Since the domain is finite, this is more straightforward than the set of speaker strategies. For each message m_i the hearer takes an action a_i. So, for example, a_1 would be the hearer's response to message m_1, in this case *ne*, and a_2 would be the hearer's response to message m_2, in this case *ne...not*.

Now that we have defined the set of speaker and hearer strategies, we can ask what strategies constitute evolutionarily stable strategies. Given that signaling games are asymmetric, this amounts to identifying the strict Nash equilibria of the game. This can be done by determining what strategies jointly maximize the expected utilities of speakers and hearers.

$$E[U_S(s, r)] = \int_T \left(1 - (r(s(t)) - t - (1 - t)b)^2\right) p(t)dt \tag{7}$$

$$E[U_R(s, r)] = \int_T \left(1 - (r(s(t)) - t)^2\right) p(t)dt \tag{8}$$

These are exactly analogous to expected utility in the discrete case, where we summed over all the possible states. Again, $r(s(t))$ is the receiver's respond to the sender's message and yields an action, which determines the utility for both sender and receiver given a state.

We estimate the prior probability over states as a *beta distribution* over the set of states, parameterized by two shape parameters α and β, and often written as $\mathcal{B}(\alpha, \beta)$. Figure 1 shows the distribution for several values of α and β, including the uniform distribution $\mathcal{B}(1, 1)$. These two

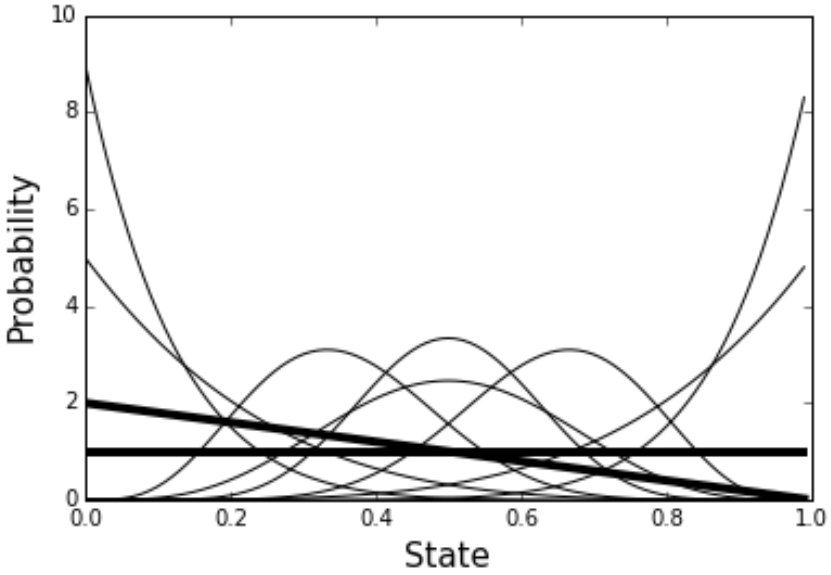

Figure 1: Beta distribution for various parameter values of α and β, including the uniform distribution $\mathcal{B}(1,1)$, and the empirical distribution $\mathcal{B}(1,2)$.

shape parameters gives us quite a bit of flexibility in modeling different potential prior probability distributions. We use them extensively in the analysis that follows, so we note two things that should offer an intuitive conceptual foothold. First, the expected value of a beta distribution is given by $\frac{\alpha}{\alpha+\beta}$. So, for example, the expected value of the uniform distribution is $\frac{1}{2}$ given that $\alpha = \beta = 1$, which is what we see in Figure 1. Second, if we fix $\beta = 1$ and let α vary, then the distribution will be more and more skewed to the right as α grows larger. This also follows from the fact that the expected value $\frac{\alpha}{\alpha+1}$ gets closer and closer to one as α grows. Note that the mirror image case would hold if we fixed $\alpha = 1$ and let β vary.

We use the historical data in Table 1 from Wallage (2013, 12) that shows the number of propositions by activation to estimate the estimate the prior distribution over states. The degree of activation in these examples is estimated using translations of the texts. However, there are reasons to be confident that activation can reliably be identified in historical corpora. These figures agree with similar estimates from contemporary corpora. For example, in a sample from a corpus of British English, Tottie (1991) finds that negation is only used 14% of the time with directly activated propositions. Likewise, in a corpus of American English, Thompson (1998) finds that negation is only used 5% of the time with directly activated propositions. These results suggest that the prior distribution is stable, with the preponderance of negation being used with brand new non-activated propositions. Intuitively, this distribution makes perfect sense: the majority of conversation is about introducing new information rather than treading the same old ground of what has already been said. If the prior distribution is indeed stable, then we can estimate it from the

PERIOD	NON-ACTIVATED	INDIRECTLY ACTIVATED	DIRECTLY ACTIVATED
1150-1250	393	203	52
1250-1350	346	296	42
1350-1420	294	179	60
TOTAL	1033	678	154

Table 1: Distribution of sentence activation in PPCME (Kroch and Taylor, 2000) from Wallage (2013)

data pooled across time periods. A good fit to the data is the prior probability distribution $\mathcal{B}(1,2)$, also shown in Figure 1.[3]

To calculate the maxima of the expected utility functions, let $\langle s^*, r^* \rangle$ be an equilibrium strategy profile. In this case, the speaker strategy is defined by $s^*((0, t_1^*)) = m_1$ and $s^*((t_1^*, 1)) = m_2$. Likewise, the hearer strategy is defined by $r^*(m_1) = a_1^*$ and $r^*(m_2) = a_2^*$. We can determine the evolutionarily stable strategies by jointly maximizing speaker and hearer expected utility. That is, we solve a system of partial differential equations for t_1^*, a_1^* and a_2^*. For any amount of bias, the maximizing values are shown in Figure 2.[4] Along the horizontal axis is the amount of speaker bias, the vertical axis represents the point at which speakers partition the state space and the actions of hearers in response to the forms. For any value of the speaker bias b, the solid black line represents the point at which speakers partition the state space t_1^* and the dashed lines represent hearer responses to the different messages, a_1^* and a_2^*.

We are now in a position to answer our first question regarding the relationship between speaker bias and the use of different forms at equilibrium. Namely, if speakers are too biased when it comes to keeping track of common versus private knowledge then only a single message can be used in equilibrium. For example, we can read off of Figure 2 that if $b > \frac{1}{6}$ then only m_2 will be used in equilibrium.[5] When speaker bias is sufficiently large this form carries no information about the activation of the proposition being negated. The *information gain*, or Kullback-Leibler divergence, from receiving the message is zero exactly when the message fails to shift the hearer's beliefs from the prior probability.

[3] We treat each of the discrete categories as equal portions of the unit interval and find values of α and β such that $\int_0^{\frac{1}{3}} \mathcal{B}(\alpha,\beta)(t)dt \approx p(\text{NON-ACTIVATED})$, $\int_{\frac{1}{3}}^{\frac{2}{3}} \mathcal{B}(\alpha,\beta)(t)dt \approx p(\text{INDIRECTLY ACTIVATED})$, and $\int_{\frac{2}{3}}^{1} \mathcal{B}(\alpha,\beta)(t)dt \approx p(\text{DIRECTLY ACTIVATED})$, where the probabilities are estimated from the totals in Table 1. Obtaining a better empirical estimate of the prior from contemporary data and intuitions is something we leave for future research.

[4] See Appendix A for the full calculations of the solution.

[5] In fact, for any number of messages n there exists a maximum amount of bias b_n such that all messages are used in equilibrium. For any number of messages Crawford and Sobel (1982) show that $b_n > b_{n+1}$. That is, for a given number of messages, there is a maximum amount of bias that allows for all messages to be used in equilibrium. As speakers' bias decreases, $b \to 0$, the number of messages that can be used in equilibrium goes to infinity. The closer the incentives of senders and receivers, the finer and more detailed the information senders want to signal. The number of messages available is the only limit on the amount of information conveyed when the interests of both parties are perfectly aligned.

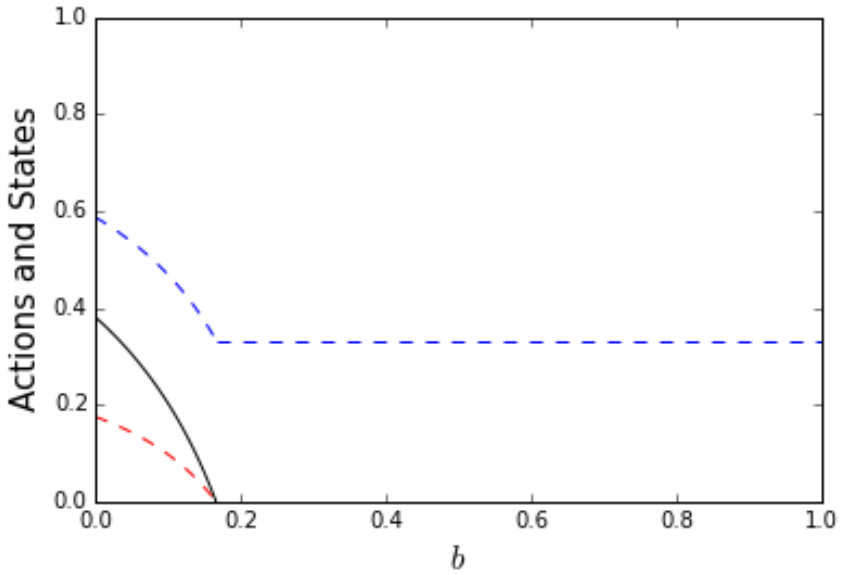

Figure 2: Equilibrium solution for two messages for values of bias

$$KL(m_2) = \int_0^1 log\left(\frac{p(t \mid m_2)}{p(t)}\right) p(t \mid m_2)dt \tag{9}$$

If the posterior $p(t \mid m_2)$ is the same as the prior $p(t)$, as is necessarily the case when a single message is used, then the information gained is zero. This follows directly from the definition of information gain, the logarithm of one is zero. This offers a precise definition of bleaching as the loss of information as a signal spreads across states.

There are two important implications for the functional cycle. First, if speakers are sufficiently biased towards their own perspective when estimating the degree of activation for a proposition, then only a single form is stable. In particular, *ne* and *ne...not* cannot coexist, in fact, only *ne...not* can be used in equilibrium. This leads to our second question. Where speaker bias is sufficiently large, are equilibria where only a single form is used evolutionarily stable? In fact, we can show that a signaling equilibria is evolutionarily stable only if all available signals are used.

To see why this is the case, suppose that the amount of bias only allows for a single form to be used in equilibrium, call it m_1. Now, suppose further that there is an additional form that is not used in equilibrium, m_2. The action that hearers would take in response to this unused message can vary without affecting the expected utility of speakers or hearers. This means that the equilibrium is not strict, and therefore is not evolutionarily stable. In particular, a single-form equilibrium can be disturbed by the introduction of a new message, it is not *neologism-proof* (Farrell, 1993). For example, suppose that a new form is only used for high degrees of activation, and that hearers respond by inferring a high degree of activation. Given speakers' bias, there will be additional

states that speakers think warrant using the new form. In response to this increase, hearers will infer a lower degree of activation, meaning additional states will be used by speakers, and so on. This holds for any pair of messages m_i and m_{i+1}.

Importantly, this is just the functional cycle as we have been describing it. So, even though two forms may not be stable for a sufficiently large degree of speaker bias, a single form is never evolutionarily stable. The functional cycle can always be set in motion by the introduction of the appropriately conditioned form. This means that just as *ne* can be pushed out by *ne...not*, so too could *ne...not* be pushed out with the introduction of the right new form, one that is initially restricted to high degrees of activation.

5 Dynamics

While we can reason about how speakers and hearers might react to the introduction of a new form, this kind of equilibrium reasoning is essentially static. That is, it allows us to reason about what would happen if we started at a particular state, but not whether we will ever reach that state in the first place. More importantly, it does not allow us to examine how a population evolves from any starting state in general. To understand how speakers and hearers change over time, we must posit a process that underlies how speakers and hearers interact and respond to each other. Doing so will allow us to examine how different degrees of speaker bias impact the trajectory of meaning. First, we discuss the replicator dynamics as an appropriate evolutionary game dynamics for studying changes in meaning. Then, we simulate trajectories of a population interacting over time.

The replicator dynamics were originally introduced as an explicitly dynamic model of biological replication, but have since been shown to have deep connections with some of the most widely-studied models of learning. In particular, Börgers and Sarin (1997) prove that the expected behavior of agents playing an asymmetric game while learning according to a *linear reward-inaction* scheme (Bush and Mosteller, 1955) is equivalent to the asymmetric replicator dynamics if the agents interact frequently and change their behavior slowly. That is, if individuals tend to do things that are more successful, then their expected behavior can be modeled by the replicator dynamics.

If we assume that speakers and hearers learn in this manner, there are a few conceptual clarifications to be made to justify the use of the replicator dynamics in modeling the functional cycle. First, we need to be sure that speakers and hearers interact frequently and change their behavior slowly. Both of these would seem to follow from the overall frequency of negation. Given that negation is one of the most frequently used forms in any language, it is safe to assume that speakers and hearers interact frequently and do not dramatically alter their use or interpretation of negation from one sentence to the next. Second, we assume that each individual acts as a speaker and hearer, but cannot introspectively reason about the impact of one on the other. That is, individuals cannot use their behavior as speakers to change their own behavior as hearers, nor vice versa. Third, while the replicator dynamics can be used as a model of individual learning, we are interested in how the population as a whole changes. However, given that the *expected* behavior of individuals is

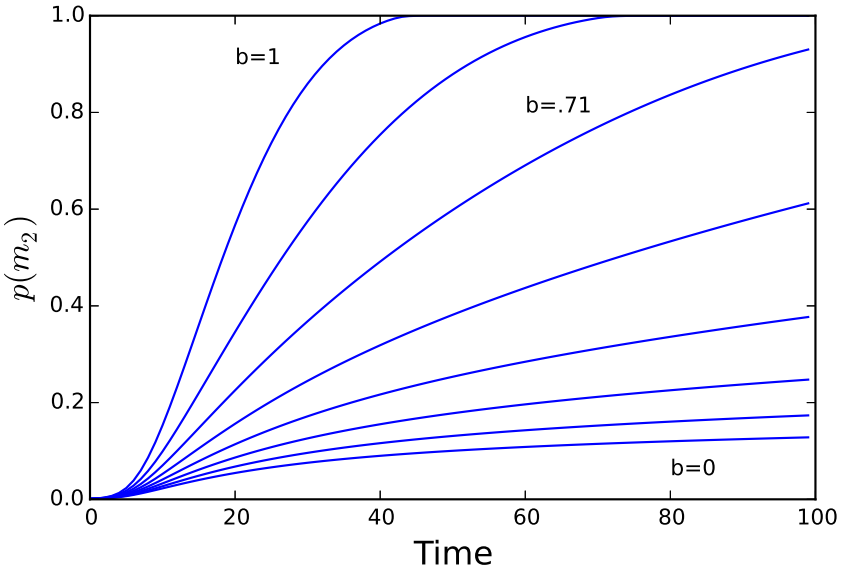

Figure 3: Proportion of m_2 under the discrete-time replicator dynamics for varying amounts of speaker bias

equivalent to the replicator dynamics under this kind of learning, then the expected behavior of a population of individuals should be as well. That is, if averaging at the individual level yields the replicator dynamics, then so should averaging over averages. This is akin to treating the populations of speakers and hearers *as if* they were individuals.

We simulate the change in the proportion of the different forms over time for a population evolving according to the replicator dynamics from the same starting conditions.[6] In this case, the population starts off from a state where speakers only use m_2 for high degrees of activation and hearers respond by inferring a high degree of activation. Varying the bias parameter b yields the trajectories shown in Figure 3. Along the horizontal axis we have time, and along the vertical axis we have the proportion of m_2 used at any point in time. For sufficiently large amounts of speaker bias, the incoming form is guaranteed to replace the incumbent form, the amount of bias controls the rate at which this happens. For sufficiently small amounts of speaker bias, both forms are guaranteed to persist.

So, our numerical simulations agree with the predictions from the static equilibrium analysis. If speaker bias is too large, then only a single form is used. More specifically, the form that is used

[6]Here we use the discrete-time replicator dynamics for computational tractability, whereas the results presented by Börgers and Sarin (1997) hold for the continuous-time replicator dynamics. While the two dynamics are substantially similar for our purposes, we leave a comparison for future work. Note that we also discretize the set of states and actions for speakers and hearers, respectively. That is, for some n, we treat the set of states $T : \{t_0, ..., t_n\}$ and actions $A : \{a_o, ..., a_n\}$, where $t_i = a_i = \frac{i}{n}$. See Appendix B for full details of the numerical simulations.

at equilibrium is the form that started off restricted to higher degrees of activation. Now, while these simulations yield qualitative information about the dynamics of the functional cycle, we are interested in how this model can be used to understand the details of the functional cycle in the history of English. We now turn to fitting the model to historical trajectory of negation in English.

6 Modeling

Defining the stage game and the evolutionary dynamics allowed us to investigate the effect of speaker bias on the functional cycle in the abstract, but we are really interested in how the resulting model can be used to explain the actual historical trajectory of negation in a concrete case. In particular, we are interested in what happens when we fit the model to data from the history of negation in English. First, we describe the data that we fit the model to. Second, we define the parameters of the dynamics that we fit. Finally, we evaluate these parameters in light of the experimental data presented above.

The data we use are drawn from the PPCME2 (Kroch and Taylor, 2000). All tokens used are negative declaratives, excluding cases of contracted negation as well as cases that appear to be constituent negation.[7] Each circle represents tokens in a given year. The size of the circle represents the number of tokens. The height of the circle represents the proportion of those instances that are a particular form. Locally-weighted regression lines are fit to these proportions. We see the transition from *ne* to *ne...not* starting around the 12th century, followed closely by the transition from *ne...not* to *not* in the 14th century.

Now, given that we are interested in the functional cycle, we care about the transition from *ne* to *ne...not*. That is, we care about an incoming emphatic form displacing an incumbent form. So, the subsequent rise of *not* is the second transition in the formal cycle, but not a part of the functional cycle. How should we deal with *not* in our analysis? There are two possibilities. First, we could ignore *not* and simply fit the model to the proportions of *ne* and *ne...not*. The problem with doing so is that this attributes too much to small fluctuations in the proportions of *ne* and *ne...not* even if those fluctuations are not meaningful in any sense relevant to the model. It is unlikely that small changes in the 15th century are something that we want to model.

Second, we could ignore the distinction between *ne...not* and *not*, and treat them as if they were the same form. This alleviates the potential problem of attributing too much meaning to small fluctuations past a certain date. More importantly, it captures the contingency of the second transition of the formal cycle to purely post-verbal negation. That is, the rise of *not* is not a part of the functional cycle, nor is it a necessary and immediate consequence of the functional cycle. We only need to compare the history of negation in French where the embracing form goes to completion before being eventually replaced by the post-verbal form. Taking this route allows us apply the same model across languages without regard to subsequent contingent developments.

[7] We model the data used here after Wallage (2008), which makes a compelling argument for treating contracted negation, among other cases, separately. Many thanks go to Aaron Ecay for sharing the code for generating the queries.

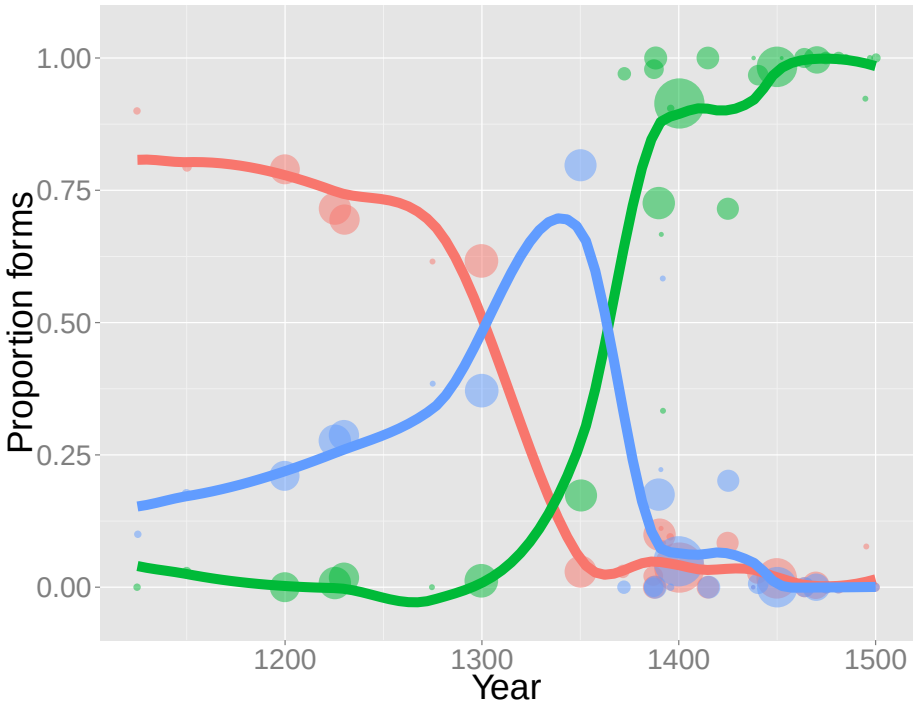

Figure 4: Proportion of *ne*, *ne...not*, and *not* in Negative Declaratives

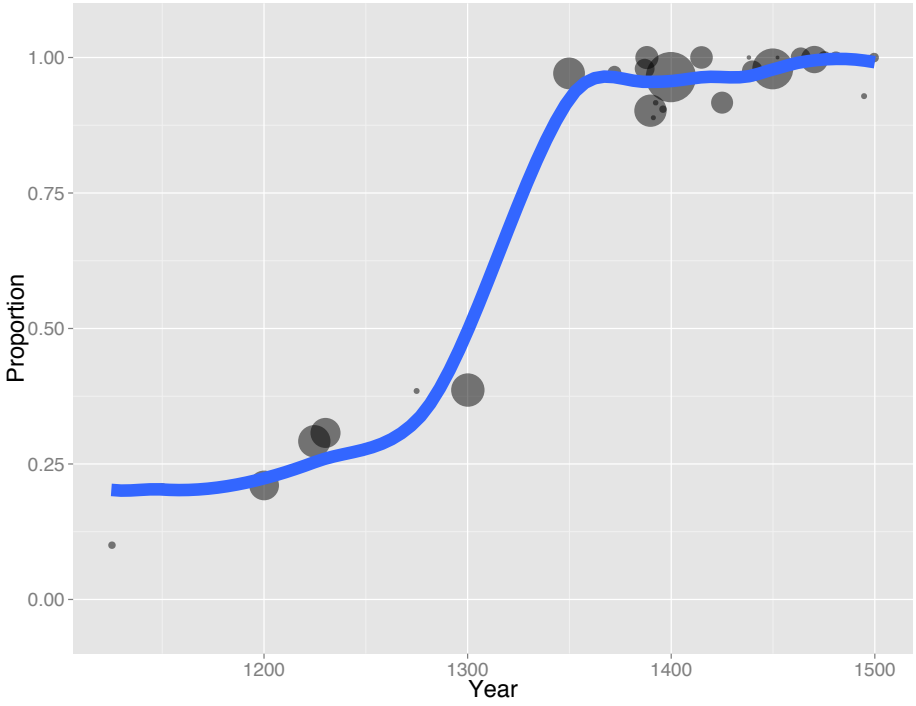

Figure 5: Proportion of *ne...not* and *not* versus *ne* over time

The results of doing so are shown in Figure 5.

Taking the trajectory of forms in Figure 5 as the data we want to fit our model to, we need to specify the parameters of the model to be fit. In particular, we need to define the initial state of how speakers use the different forms and how hearers respond to them. In fact, we have quite a bit of information regarding what the initial state of the functional cycle actually is. That is, we know that *ne...not* is fairly infrequent and largely restricted to high degrees of activation. Likewise, we know that hearers' response to *ne...not* is also largely restricted to actions corresponding to high degrees of activation. We can translate this information into conditions on the initial states of the speaker and hearer populations.

Regarding speakers we assume that both forms have a particular meaning, which is captured by conditional probability of states given a form. Namely, *ne* is the default form and does not carry any information above and beyond the prior, it roughly satisfies the following conditional distribution $p(t \mid ne) \sim \mathcal{B}(1, 2)$. In contrast, *ne...not* is overwhelmingly used in states with high degrees of activation, that it satisfies the following conditional distribution $p(t \mid ne...not) \sim \mathcal{B}(\alpha, 1)$. The larger α is, the more skewed towards high degrees of activation is *ne...not*. Note that these two distributions along with the prior determine the initial proportion of *ne...not*. So, we only have a single parameter α to fit for the initial state of speakers.

Regarding hearers, we assume that the expected value of the responses to both forms corre-

spond to the expected value of the conditional probability of states given the form. Intuitively, this corresponds to hearers starting off with a fairly accurate responses to the two forms. For *ne* this is satisfied by any distribution $\mathcal{B}(\alpha, \beta)$ such that $\alpha = \frac{1}{2}\beta$, which has an expected value $\frac{\frac{1}{2}\beta}{\frac{1}{2}\beta+\beta} = \frac{1}{3}$. Note that is the same as the expected value of the conditional probability of the state given the message $p(t \mid ne) \sim \mathcal{B}(1, 2)$. So we take the conditional probability of actions given *ne* to be $p(a \mid ne) \sim \mathcal{B}(\frac{1}{2}\beta_1, \beta_1)$. All that β_1 does is to determine how concentrated the action is around the expected value. For *ne...not* let $\gamma = E[t \mid ne...not]$, then this is satisfied by any distribution $\mathcal{B}(\alpha, \beta)$ such that $\alpha = \left(\frac{\gamma}{1-\gamma}\right)\beta$, so we take the conditional probability of an action to be $p(a \mid ne...not) \sim \mathcal{B}(\left(\frac{\gamma}{1-\gamma}\right)\beta_2, \beta_2)$. Again β_2 determines how concentrated the action is around the expected value. So, we have two parameters β_1 and β_2 to fit for the initial state of hearers.

The last thing to note before fitting the model is the notion of time. That is, the replicator dynamics specify how populations change from one point in time to the next, but how these abstract units correspond to days or years is unspecified. In what follows we treat each of these abstract time units as a year, but leave open the possibility that another proportion may be more appropriate. One option would be to treat the ratio between years and abstract time units as another parameter to be fit in the model, but for now we leave this as an avenue for future research.

We fit the initial state parameters and bias parameter to the data.[8] We begin by visualizing the overall trajectory of the incoming form, then turn to the change in the meaning of the two forms over time. Figure 6 shows the predicted proportion of *ne...not* over historical time for the fitted model with the bias parameter $\hat{b} = 0.49132877$. Perhaps more importantly, we can actually inspect the inner workings of the model as they relate to the functional cycle.

In particular, we can examine how the information carried by the two forms changes over time. We gain insight into the functional cycle by considering how the meaning of *ne...not* changes over time as in Figure 7. The horizontal axis represents states and vertical axis represents the conditional probability of states given that *ne...not* was used. We show this conditional probability at various points as the functional cycle proceeds. The dashed line indicates the prior probability distribution over states. The initial meaning of the incoming emphatic form is represented by the curve with the most rightwards skew. This indicates the point at which the incoming emphatic form carries the most information and is thus the most emphatic. But, as time goes on, *ne...not* spreads to more and more degrees of activation as it the form increases in frequency. We represent this with subsequent distributions that move more and more towards the prior distribution. As they do so, the form loses its emphasis, as indicated by the thickness of the line. When *ne...not* is the only form, it carries no information about activation beyond the prior. Visually speaking, at this point its emphasis has faded entirely.

We also gain insight in to the functional cycle by comparing the relative meaning of both forms. Comparing the meaning of *ne* at the outset of the cycle and *ne...not* at the end of the cycle is particularly informative. Figure 8 emphatically demonstrates the dynamics of the push-

[8] See Appendix B for the full details of the starting states and the resulting fit.

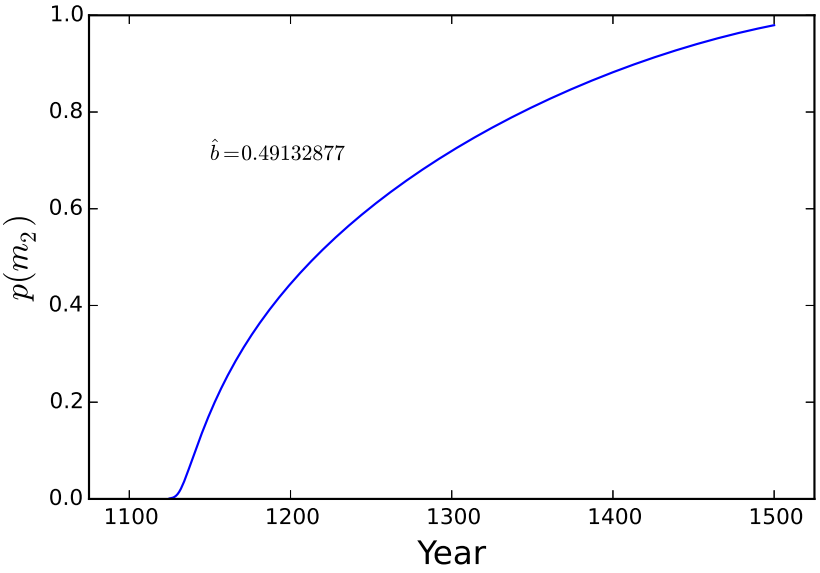

Figure 6: Predicted probability of *ne...not* over time for fitted model of functional cycle.

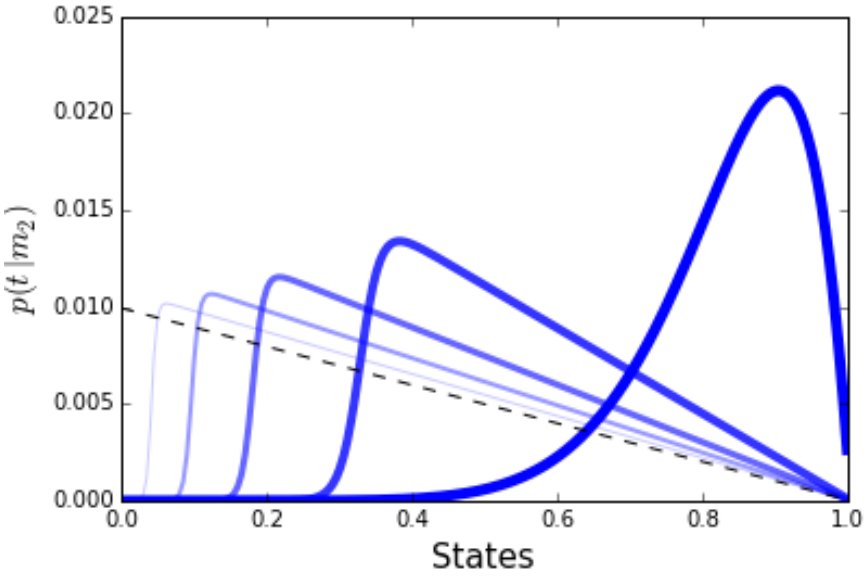

Figure 7: The emphatic form over time as given by the conditional probability of states given *ne...not*, where dashed line indicates prior probability distribution.

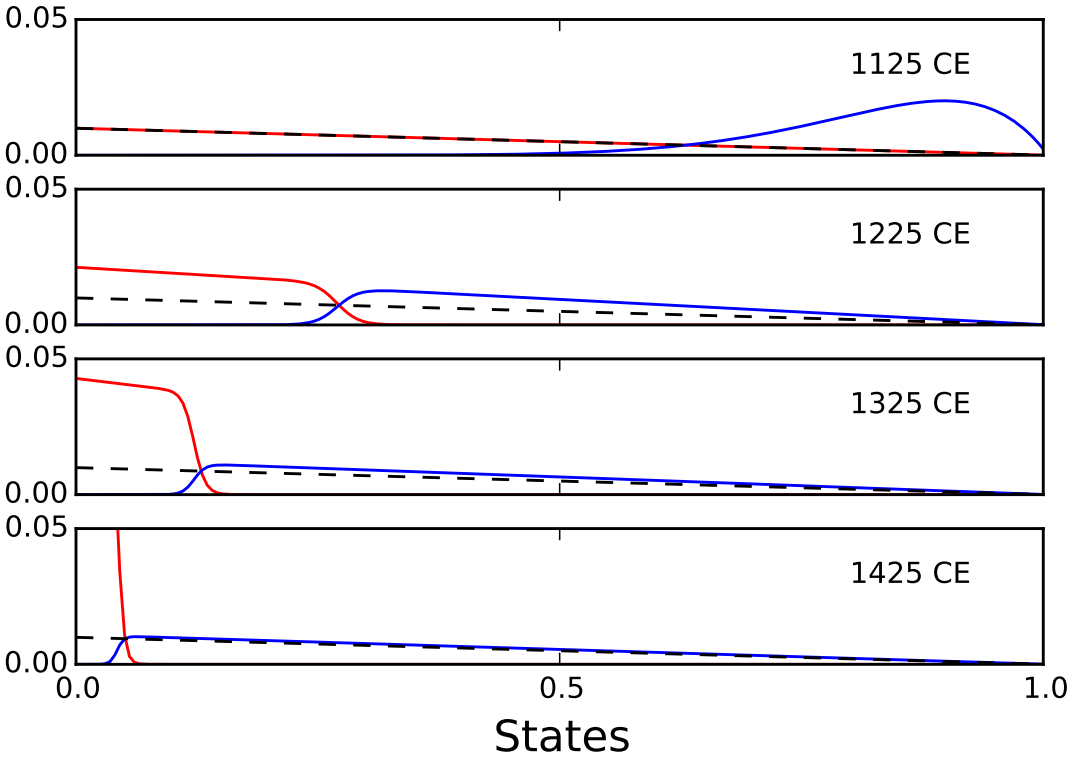

Figure 8: The push-chain of the functional cycle given by the conditional probability of states given form at various points in time, where dashed line indicates prior probability.

chain scenario. At the beginning of the cycle in 1125 CE, *ne* carries no information about the degree of activation, it coincides with the prior as indicated by the dashed line. In contrast, *ne...not* is overwhelmingly restricted to cases where the proposition being negated has a high degree of activation. Both of these facts are shown in the top panel of Figure 8. But, one hundred years later, *ne...not* has expanded to more states as it increases in frequency and *ne* is pushed to lower and lower degrees of activation. This is shown in the second panel of Figure 8. As the functional cycle proceeds, the old form is pushed lower and lower down the scale. Eventually, the incoming form has displaced the incumbent form and ceases to carry any information about the degree of activation.

So, the dynamics of the fitted model match our theoretical conceptions of the functional cycle as a kind of push chain. The incoming form pushes the incumbent form out, eventually taking its place and losing its emphasis. Given that the driving force behind this change was posited to be speakers' bias to overestimate activation, it is important to take a moment to evaluate the value of the fitted bias parameter, $\hat{b}$. Given that the incoming form replaces the incumbent form, we would

expect from our equilibrium analysis that at the very least $\hat{b} > \frac{1}{6}$, but this still leaves a fair amount of room for the parameter to vary. In fact the fitted value $\hat{b} = .4913287$ is well above this minimum.

Now, given this value, we might ask whether it is reasonable in light of the experimental results discussed above. To evaluate the parameter, we return to the results reported in Wu and Keysar (2007). Namely, when playing a communication game, speakers relied on private knowledge of the names of shapes in a proportion of trials. We can use this information to estimate the expected value of the bias parameter exhibited by speakers in the experiment.

To see this, first suppose that there are only two states corresponding to whether or not the name for a particular shape was learned privately or jointly. That is, knowing whether a name was learned privately or jointly is categorical. As Heller et al. (2012) found, this is a reasonable assumption given that participants were incredibly accurate at recalling the context of learning for shapes. Second, suppose that the utility functions for both speakers and hearers are the same form as above. Third, suppose that hearers take one action in response to names and another for descriptions that correspond to initial guesses about the status of the target shape. Then a speaker would only prefer the action taken in response to a name for a privately learned shape if $b > \frac{1}{2}$.

But, this preferences is not categorical. In fact, from the experimental results we only know the probability that $b > \frac{1}{2}$. However, we can estimate the expected value of the bias parameter. Let $p(b) \sim \mathcal{B}(\alpha, \beta)$ be a distribution over the unit interval. We find the parameters such that $\int_{\frac{1}{2}}^{1} p(b)db = p(b > \frac{1}{2})$, which in turn give the expected value of b. For example, where speakers use a privately known name in 5% of trials $E[b] = .1398$, and where speakers use a privately known name in 28% of trials $E[b] = .3549$.

So, there are a range of potential values of speaker bias that we can estimate from the experimental evidence. In both cases these are smaller than the fitted value of the bias parameter for the functional cycle. However, there are good reasons to treat these experimental estimates as lower bounds. First, the fitted parameters deal with different domains. Where the experiments deal with the referential domain, the functional cycle deals with the propositional domain. It is certainly possible that speaker bias varies across these domains. In fact, we might even expect this. For example, referents often come along with some externally observable entity in the real world, whereas propositions often do not. The fact that propositions are in this sense more abstract may lead speakers to rely on their own perspective more. Second, the experiments found speaker bias even between strangers. These kinds of communicative biases are even more pronounced between people who know each other well (Savitsky et al., 2011).[9] Thus the degree of speaker bias in everyday life may be significantly larger than these experimental estimates suggest.

Careful experimentation will be needed to nail down how private and common knowledge are

[9]For example, one day when I got home the first thing my wife said to me was, "I DID make an appointment." This struck me as out of the blue, but she said that she told me that she was feeling a little under the weather and debating whether her cold symptoms warranted a trip to the doctor. This conversation had happened several days prior and I had completely forgotten about it, but it was on her mind. In other words, her own subjective estimate of $p =$"I made a doctor's appointment." was greater than the actual degree of activation. In this case, I would say that the bias was fairly high $b \approx 1$.

tracked in the propositional domain, and how this plays out in everyday life. However, the results are largely compatible with both the mechanics of the dynamic model of the functional cycle we have defined here.

Summary

Separating out the formal and functional cycles lightens the explanatory burden. By isolating the functional cycle we were able to identify what conditions the incoming form and reason about why those conditioning factors change over time. In particular, we argued that speakers have difficulty in keeping track of private versus common knowledge, which biases them towards overestimating the activation of propositions being negated. The tools used to model the functional cycle allow us to offer the first explanatory model of the dynamics of how meaning changes over time. Importantly, they also highlight the fact that while the driving force of the functional cycle is a byproduct of our cognitive limitations in tracking common knowledge, change comes about through the social interactions between individuals in a population. Thus explaining the functional cycle requires a model of how pragmatic competence shapes signaling over time.

Before moving on, we pause to consider two potential lines of research related to the model we have discussed here. The first deals with the alternative definition of emphasis offered at the outset. That is, emphatic negation widens and strengthens negation to preclude exceptions. This interpretation is appealing insofar as negative polarity items are often recruited to create emphatic forms and have exactly this effect (Kadmon and Landman, 1993; Eckardt, 2006). However, this approach to the functional cycle would have to do two things. First, it would have to specify what serves the role of speaker bias in driving the increase of an incoming emphatic form. Second, it would have to address the problem of over-prediction. That is, if new signals can be formed with the addition of any negative polarity item, then there will always be new forms available, and thus the functional cycle should always be occurring. The fact that we do not observe Jespersen's treadmill means there must be some kind of restriction on what can serve as a new emphatic form. One potential restriction is that the new form must be free from sortal restrictions. For example, both "I didn't move a crumb" and "I didn't eat a crumb" must be equally acceptable.

The other line of research has to do with the implications of this model of the functional cycle for referring expressions that are also sensitive to degrees of activation. Gundel et al. (1993) refer to the scale of sensitivity as the *giveness hierarchy*, which is roughly ordered by pronouns, demonstratives, definites, and indefinites. Pronouns are restricted to referring to entities that are directly activated, whereas indefinites can be used with any entity. Interestingly, similar diachronic patterns are observed as forms spread to lower degrees of activation. For example, the Modern English definite *the* comes from the Old English demonstrative *se*.[10] However, there are at least two

[10]I cannot help but note discussions along this line with Jon Stevens (p.c. March 26, 2010): "One long term goal of this sort of research could be to connect it up with facts about language learning and pragmatics (perhaps using game theoretic tools) to paint a larger picture of why grammaticalization phenomena are so pervasive across languages.

interesting implications of the model of the functional cycle for the stability within the givenness hierarchy. First, we would predict greater stability in these referential terms given the prior distribution over degrees of activation. If propositions are largely skewed towards being non-activated, then referents are largely skewed towards being activated. This change in the distribution largely counteracts any amount of speaker bias. Second, the generation of new pronouns, demonstratives, or definites is arguably a rare event. At least, it would seem rarer than a form of negation becoming associated with activation. We leave exploring both these lines of research for the future.

References

Austin, John. 1962. *How to do things with words*. Clarendon Press.

Börgers, Tilman, and Rajiv Sarin. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77(1):1–14.

Bush, Robert, and Frederick Mosteller. 1955. *Stochastic models for learning*. Wiley.

Chafe, Wallace L. 1974. Language and consciousness. *Language* 111–133.

Clark, Herbert H, and Catherine R Marshall. 1981. Definite reference and mutual knowledge. In *Psycholinguistics: Critical concepts in psychology*. Cambridge University Press.

Crawford, Vincent P, and Joel Sobel. 1982. Strategic information transmission. *Econometrica* 1431–1451.

Detges, Ulrich, and Richard Waltereit. 2002. Grammaticalization vs. reanalysis: A semantic-pragmatic account of functional change in grammar. *Zeitschrift f"ur Sprachwissenschaft* 21(2): 151–195.

Dryer, Matthew S. 1996. Focus, pragmatic presupposition, and activated propositions. *Journal of pragmatics* 26(4):475–523.

Eckardt, Regine. 2006. *Meaning change in grammaticalization: an enquiry into semantic reanalysis*. Oxford University Press.

Farrell, Joseph. 1993. Meaning and Credibility in Cheap-Talk Games. *Games and Economic Behavior* 5:514–531.

Grice, H.P. 1989. *Studies in the way of words*.

As I alluded to in my vignette yesterday, a good model of semantic learning will likely interact with pragmatics in an interesting way; if such modeling techniques become sophisticated enough so as to model the acquisition of grammatical forms as well as content forms, then predictions will be made about the actuation and spread of bleaching, which could serve as a nice test of a model's plausibility."

Grieve-Smith, Angus. 2009. The spread of change in french negation. Ph.D. thesis, University of New Mexico.

Gundel, Jeanette K, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 274–307.

Hansen, Maj-Britt Mosegaard. 2009. The grammaticalization of negative reinforcers in old and middle french: a discourse-functional approach. *Current trends in diachronic semantics and pragmatics* 227–251.

Hansen, Maj-Britt Mosegaard, and Jacqueline Visconti. 2009. On the diachrony of" reinforced" negation in french and italian. *Grammaticalization and pragmatics: Facts, approaches, theoretical issues* 137–171.

———. 2012. The evolution of negation in French and Italian: Similarities and differences. *Folia Linguistica* 46(2).

Heller, Daphna, Kristen S Gorman, and Michael K Tanenhaus. 2012. To name or to describe: Shared knowledge affects referential form. *Topics in cognitive science* 4(2):290–305.

Kadmon, Nirit, and Fred Landman. 1993. Any. *Linguistics and philosophy* 16(4):353–422.

Kahneman, Daniel. 2011. *Thinking, fast and slow*. Macmillan.

Keysar, Boaz, Shuhong Lin, and Dale J Barr. 2003. Limits on theory of mind use in adults. *Cognition* 89(1):25–41.

Kiparsky, Paul, and Cleo Condoravdi. 2006. Tracking Jespersen's Cycle. In *Proceedings of the 2nd International Conference of Modern Greek Dialects and Linguistic Theory*, ed. Mark Janse. University of Patras.

Krifka, Manfred. 1995. The semantics and pragmatics of polarity items. *Linguistic analysis* 25(3-4):209–257.

Kroch, Anthony, and Ann Taylor, eds. 2000. *Penn-Helsinki Parsed Corpus of Middle English Second Edition (PPCME2)*.

Landman, Fred. 1991. *Structures for semantics*. Springer.

Lewis, David. 1969. *Convention*. Cambridge: Harvard University Press.

———. 1970. General semantics. *Synthese* 22(1):18–67.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In *Radical pragmatics*, ed. Peter Cole, 223–255. Academic Press.

Savitsky, Kenneth, Boaz Keysar, Nicholas Epley, Travis Carter, and Ashley Swanson. 2011. The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology* 47(1):269–273.

Schwenter, Scott A. 2005. The pragmatics of negation in Brazilian Portuguese. *Lingua* 115(10): 1427–1456.

———. 2006. Fine-tuning Jespersen's cycle. In *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honour of Laurence R. Horn*, ed. Betty T. Birner and Gregory Ward, 327–344. John Benjamins.

Stalnaker, Robert. 1978. Assertion. In *Syntax and Semantics*, ed. Peter Cole, 315–32. New York: New York University Press.

———. 2002. Common ground. *Linguistics and philosophy* 25(5):701–721.

Thompson, Sandra A. 1998. A discourse explanation for the cross-linguistic differences in the grammar of interrogation and negation. In *Case, typology and grammar: In honor of barry j. blake*, 309–341.

Tottie, Gunnel. 1991. *Negation in English speech and writing: a study in variation*. Academic Press.

Wallage, Phillip. 2008. Jespersen's cycle in Middle English: parametric variation and grammatical competition. *Lingua* 118(5):643–674.

———. 2013. Functional differentiation and grammatical competition in the English Jespersen Cycle. *Journal of Historical Syntax* 2(1).

Wardlow Lane, Liane, Michelle Groisman, and Victor S Ferreira. 2006. Don't talk about pink elephants! Speakers' control over leaking private information during language production. *Psychological science* 17(4):273–277.

Wu, Shali, and Boaz Keysar. 2007. The effect of information overlap on communication effectiveness. *Cognitive Science* 31(1):169–181.