2

The role of entropy and surprisal in phonologization and language change

ELIZABETH HUME AND FRÉDÉRIC MAILHOT*

"What are the laws of motion but the expectations of reason concerning the position of bodies in space? We are thus justified, not only in saying that all complete knowledge involves anticipation, but also in affirming that all rational expectation is knowledge." (Hitchcock 1903: 673)

2.1 Introduction

Traditionally, the term phonologization has been used to describe a diachronic change within a given language system from a state of phonetic variation to that of phonological generalization (Hyman 1976). More specifically, we take this to mean a diachronic shift from variation across a large number of uncorrelated dimensions to correlated variation of lower dimensionality. Such transitions are relevant both to the creation of new categories and patterns (e.g. phoneme, stress pattern), as well as to the change from one existing category into another. Many factors external to a language's grammatical system have been shown to play an influential role in this process. Some of these external factors are listed below (for relevant discussion see Archangeli and Pulleyblank 1994; Blevins 2004; Bybee 2001; Culicover and Nowak 2002; Davidson 2007; Guion 1998; Hayes and Londe 2006; Hume and Johnson 2001a; Hyman 1976; Joseph and Janda 2003; Jeffers and Lehiste 1979; Lindblom 1990; Moreton and

* We owe a debt of gratitude to Kathleen Currie Hall, Dahee Kim, Adam Ussishkin and Andrew Wedel for much lively discussion regarding the ideas in this chapter. We would also like to thank the following people for their input on aspects of this research: Paul Boersma, Chris Brew, Joan Bybee, Jennifer Cole, Peter Culicover, Alex Francis, John Goldsmith, John Hale, Ilana Heintz, Robert Kirchner, Kate Kokhan, Jeff Mielke, William Schuler, Andrea Sims, Rory Turnbull, Mike White, Alan Yu, members of the Ohio State phonetics/phonology and socio-historical linguistics discussion groups, and two anonymous reviewers.

Thomas 2007; Ohala 1981, 1993c, 2003; Peperkamp, Vendelin and Nakamura 2008; Yu 2007, *inter alia*).

Grammar-external factors influencing phonologization include:

a. phonetic factors, e.g. perceptual distinctiveness, articulatory difficulty;
b. usage factors, e.g. familiarity, frequency;
c. processing factors due to, e.g., structural complexity.

While there is ample evidence showing the impact of these diverse forces on language systems, they are often treated independently of one another (though see Blevins and Wedel 2009). As such, the literature on language change is replete with arguments for why one factor, as opposed to another, underlies a particular modification. In this chapter we propose that a unified account of the influence of these and other factors is possible when we view the phenomena of phonologization through the lens of *information theory* (Shannon 1948), in particular making use of the concepts of *surprisal* and *entropy*. Not only do the tools of information theory allow us a deeper understanding of why these factors influence language systems in the way that they do, they also provide insight into the process of phonologization.

In the current context, entropy models a cognitive state of the language user associated with the amount of uncertainty regarding the outcome of identifying or producing some linguistic event, e.g. the next word in a sentence (Townsend and Bever 2001; Hale 2003; Levy 2008), or the vowel that is epenthesized or deleted (Hume and Broomberg 2005; Hume et al. 2011). All linguistic elements have an associated (context-dependent) surprisal, and contribute individually to an overall measure of uncertainty in selecting among outcomes in a system (the entropy) associated with the outcome of some event, e.g. which vowel will be epenthesized. As we show, each element can contribute to entropy as a function of factors such as those discussed above, e.g. perceptual distinctiveness, usage frequency.

Entropy and surprisal are of particular relevance to phonologization for a number of reasons. The first is linked to learning. The mind's attentional focus is drawn to contextually informative, or higher surprisal, elements, e.g. auditory cues (Grossberg 2003; Baldi and Itti 2010), and attentional focus is known to be a crucial component of learning (McKinley and Nosofsky 1996; Kruschke 2003). Given that speaker-hearers must learn to associate phonological meaning to particular phonetic details in order for phonologization to occur, surprisal (and by extension system entropy) is likely to play a key role. The second reason, and the main focus of this chapter, is that the approach advocated here brings clarity to phonologization by making strong predictions about both the likely *targets* of change, as well as the nature of the *resultant* change.

Surprisal is a continuous measure, taking values in the interval $[0, \infty]$, with increasing surprisal being a function of decreasing probability. As elaborated below, elements falling toward each pole of the range of surprisal are unstable, making them more

prone to change than elements occurring away from the extremes. Phonologization is thus predicted to preferentially affect elements linked to extreme degrees of surprisal, i.e. that have a small entropic contribution. Interestingly, while the mechanisms that affect elements with very low or very high surprisal may differ, they pattern together in being prone to change given their low contribution to predicting outcomes in a system.

The current approach also speaks to the nature of change. Unstable elements with high surprisal are biased to change in the direction of a similar element or pattern with lower surprisal, consistent with observations regarding analogical change (see e.g. Phillips 2006; Wedel 2007). In other words, change affecting high surprisal elements is predicted to preserve structures that the speaker-hearer is already familiar with. Conversely, as developed below, change in patterns with low surprisal need not be structure preserving, and such patterns are typically prone to production-based reduction processes (Bybee 2001), which can introduce novel patterns into a speaker-hearer's linguistic system.

Before delving into these points in more detail, we define the information-theoretic concepts of surprisal and entropy more rigorously, then briefly discuss the cognitive state modeled by surprisal, which we call 'expectedness'. With this groundwork in place, we turn to the heart of the chapter: the relevance of entropic contribution and surprisal for phonologization and language change. Section 2.3 outlines in general terms the effects of surprisal on language systems. The section also focuses on the linguistic consequences of two key properties of our approach: instability and bias. In doing so, we take a closer look at the potential for a given element to undergo change or be the outcome of change given the degrees of surprisal associated with it.

2.2 Information, surprisal, and entropy

In this section we introduce the basic notions of information theory that we shall make use of in the approach to phonologization and language change developed further below. While information-theoretic concepts are foundational to the field of computational linguistics, they are less familiar to linguistics more generally, though see Cherry, Halle and Jakobson (1953); Hockett (1955); Broe (1996); Hale (2003); Goldsmith (1998, 2002); Aylett and Turk (2004); Hume (2006); Hall (2009); Jaeger (2010); Jaeger and Tily (2011); Levy and Jaeger (2007); Goldsmith and Riggle (to appear). For further coverage of information-theoretic concepts, the reader is referred to Shannon (1948), the founding document of the field, which remains an excellent introduction, or to Cover and Thomas (2006), the currently standard text, for extensive and mathematically rigorous coverage of information-theoretic concepts.

2.2.1 *Entropy and surprisal*

Information theory is concerned with representing mathematically how much information is needed to convey a message given the constraints imposed on a

communication system. Entropy, H, can be understood in terms of making a decision over a range of outcomes related to the message, e.g. identifying the quality of an epenthetic vowel in context $C\_\_\_C$. It is a probabilistic measure of the amount of uncertainty associated with selecting among outcomes, e.g. a set of vowels. Higher uncertainty correlates with higher entropy. Studying system entropy is useful for determining mathematically how much an element in the system contributes to uncertainty in predicting probabilistic outcomes. As such, it can provide a measure of the element's contribution to the language's effectiveness as a system of communication. Elements that contribute more to predicting an outcome are more crucial for successful communication.

In information-theoretic terms, an element's contribution to system entropy is its probability multiplied by its *surprisal* (also referred to as *information content*). Every element in a system has an associated surprisal, S, which is the negative logarithm[1] of its probability:

$$S(x_i) = -\log_2 P(X = x_i) \tag{1}$$

where X is an event[2] ranging over a set of possible outcomes $\{x_1, x_2, \ldots, x_i, \ldots\}$ each with an associated probability, $P(X = x_i)$. In the general case, these probabilities are defined contextually, e.g. phonologically, morphologically, etc.

Figure 2.1 illustrates the relation between probability and surprisal. Surprisal varies continuously between zero and positive infinity; the occurrence of a highly likely event (e.g. observing some vowel in a context where it is the only permissible one) has low surprisal, while a highly unlikely event (e.g. observing some phonotactically prohibited sequence of segments) has high surprisal. This reflects the intuition that the occurrence of improbable events is highly surprising, while the occurrence of highly likely events is not surprising.

As noted above, an element's contribution to the uncertainty (i.e. entropy) associated with predicting the outcome of an event is its probability multiplied by its surprisal, as given in Equation 2,

$$\begin{aligned} H_c(x_i) &= P(X = x_i) \cdot S(x_i) \\ &= -P(X = x_i) \cdot \log_2 P(X = x_i) \end{aligned} \tag{2}$$

where X, as above, is an event whose outcome can take one of several values in the vocabulary set V_X (e.g. outcomes of X could be any vowel in a language under consideration), $P(X = x_i)$ is the probability that outcome x_i will be observed, and the quantity $-\log_2 P(X = x_i)$, as discussed above, is the surprisal of outcome $X = x_i$. We label $H_c(x)$ the *entropic contribution* of x.

[1] We follow convention here and use a logarithmic base of 2, which allows us to express surprisal and entropy in units of *bits*. Using a different logarithmic base is equivalent to a multiplicative scaling.

[2] Formally, a random variable.

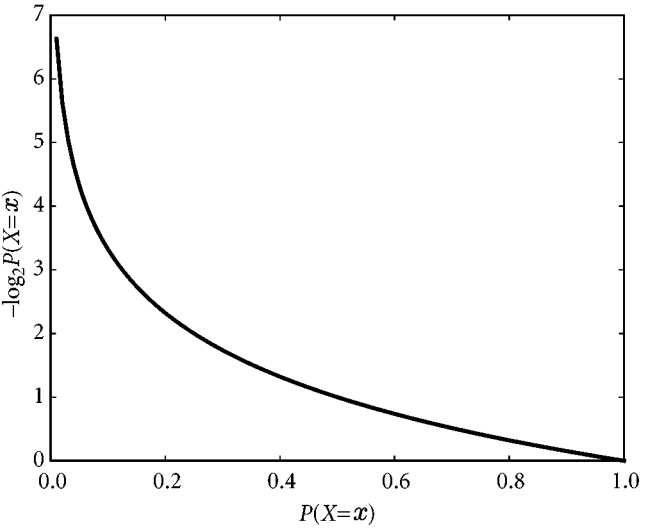

FIGURE 2.1 Plot of probability vs. surprisal

The *entropy* of a system is the sum of its elements' entropic contributions, as in Equation 3. Thus, it is a measure of the average surprisal of the system.

$$\begin{aligned} H(X) &= \sum_{x \in V_X} H_c(x) \\ &= -\sum_{x \in V_X} P(X = x) \cdot \log_2 P(X = x) \end{aligned} \tag{3}$$

Probabilistic notions are clearly relevant to the study of language acquisition, use, change and representation, as discussed in works such as: Bod, Hay, and Jannedy (2003); Boersma and Hayes (2001); Bybee (1985, 2001); Coleman and Pierrehumbert (1997); Frisch, Pierrehumbert and Broe (2004); Goldsmith (2007); Greenberg (1966); Hooper (1976b); Hume (2004a, b); Jurafsky et al. (2001); Phillips (1984, 2006); Luce and Pisoni (1998); Pitt and McQueen (1998); Trubetzkoy (1969); Vitevitch and Luce (1999); Zipf (1932), *inter alia*. Hence, the cognitive state modeled by surprisal correlates with probability. The notion 'probability' here may be approximately equated to subjective degree of belief, as in a Bayesian approach to cognition (Pearl 1988; Chater, Tenenbaum and Yuille 2006), in which prior states of knowledge are taken into consideration when computing the probability of some future event or state.

To illustrate, consider a hypothetical language, $\mathscr{L}$, with the following vowels: $V_{\mathscr{L}}$ = {i, e, a, o, u, ə}. We wish to compute the entropy of $\mathscr{L}$'s system of vowels; more specifically, we want a measure of the amount of uncertainty associated with e.g. predicting the observation of some vowel in a given phonological context, an event

we label L. First we take the case where each vowel is assumed to be, *ceteris paribus*, equiprobable; then each $v \in V_{\mathscr{L}}$ has a probability of observation $P(L = v) = \frac{1}{6}$. The entropy computation is then as follows:

$$\begin{aligned} H(L) &= -\sum_{v \in V_{\mathscr{L}}} P(L = v) \cdot \log_2 P(L = v) \\ &= -1 \cdot 6 \cdot \left[\frac{1}{6} \cdot \log_2 \frac{1}{6}\right] \\ &\approx 2.585 \end{aligned} \tag{4}$$

Of course, since the entropy of a system is its average surprisal, and each vowel in this case has the same surprisal value (since they are equiprobable), the entropy of this system is equal to each vowel's surprisal. To illuminate the relationship between surprisal and entropy more clearly, we can examine how the entropy of this system changes as we alter the probability estimates for particular vowels. As a simple initial case, assume that one vowel, e.g. {ə}, is more probable in some context than the others, which are all equiprobable. For concreteness let us assume that the probability of observing a schwa, $P(L = \text{ə})$, is $\frac{3}{8}$, hence the surprisal $S(L = \text{ə}) = -\log_2 \frac{3}{8} \approx 1.4$. Then the surprisal of observing any of the remaining vowels is $S(L = v \neq \text{ə}) = -\log_2 \frac{1}{8} = 3$. The entropy of the system under this distribution is then

$$\begin{aligned} H(L) &= -\sum_{v \in V_{\mathscr{L}}} P(L = v) \cdot \log_2 P(L = v) \\ &= -1 \cdot \left[\frac{3}{8} \cdot \log_2 \frac{3}{8}\right] - 5 \cdot \left[\frac{1}{8} \cdot \log_2 \frac{1}{8}\right] \\ &\approx 2.406 \end{aligned} \tag{5}$$

Note that the entropy in this case is lower than when all vowels are equiprobable. This is because there is now less uncertainty about which vowel will occur in the context under consideration, due to schwa's higher probability of observation. We state here without proof the theorem that the entropy of a system is maximized when all of its outcomes are equally probable (Shannon 1948: 11).

Consider finally a slight generalization of the previous case, where we examine all possible values for the probability of schwa occurring in some context, assuming the remaining vowels are equiprobable. In lieu of additional calculations of entropy, consider the graphs in Figure 2.2 and Figure 2.3: the first is of the entropy of $\mathscr{L}$'s vowel system versus the probability of observing schwa, the second is of schwa's contribution to the entropy of $\mathscr{L}$ versus its probability of observation.

Note that entropic contribution goes to zero in Figure 2.3 for both low and high probabilities. That is, outcomes known to be either (near) certain or (near) impossible contribute little to the entropy of the system. As will be discussed further below, the fact that surprisal extremes contribute little to system entropy is crucial to our model

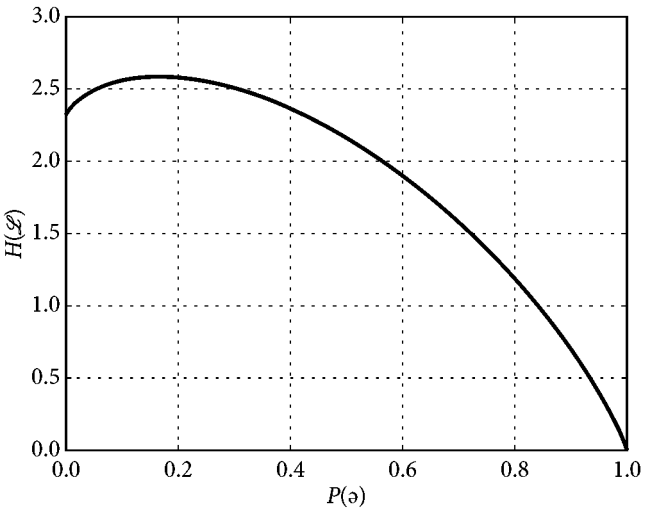

FIGURE 2.2 Entropy of $\mathscr{L}$'s vowel system, as a function of the probability of observing {ə}, assuming equiprobability of other vowels

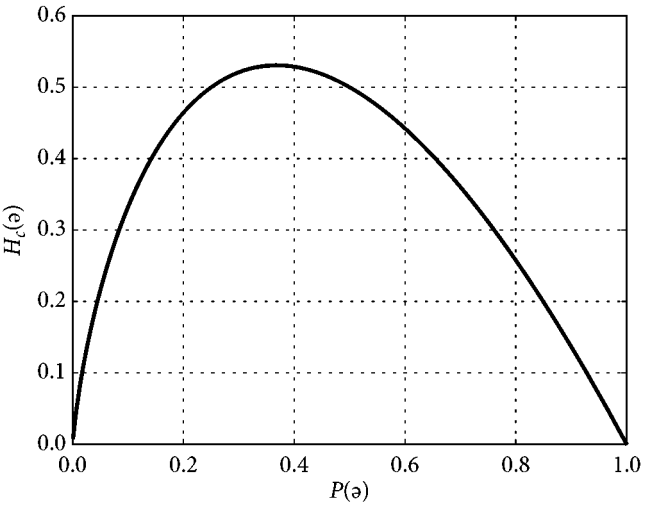

FIGURE 2.3 Contribution of {ə} to the entropy of $\mathscr{L}$, as a function of its probability of observation, assuming equiprobability of other vowels

of phonologization. In Figure 2.2, the entropy of the system does not go to zero for $P(L = ə)$, since there is still maximal uncertainty about which of the remaining five vowels will be observed. Before turning to the details of our model, we discuss more specifically the measures relevant to the calculation of surprisal.

2.2.2 *Bases of phonological surprisal (and entropy)*

Our discussion of surprisal thus far is compatible with the use of maximum likelihood estimates of probability. If we use such estimates, probability is calculated in terms of the frequency of occurrence of some element; a more frequent element has lower surprisal than a less frequent one. In this manner, frequency can be viewed as conditioning the outcome of some linguistic event which, as noted above, is strongly supported by evidence showing that frequency impacts the learning, use and representation of sound patterns. Yet frequency is not the only factor that conditions phonological patterns. As stated in the introduction, it is well-established that other factors are also relevant, including a pattern's perceptual distinctiveness and the precision with which a sound sequence is produced (e.g. Blevins 2004; Davidson 2007; Guion 1998; Hume and Johnson 2001a; Joseph and Janda 2003; Jeffers and Lehiste 1979; Lindblom 1990; Ohala 1981, 1993c, 2003). An adequate model of phonologization and language change must then also provide a means of integrating these factors.

The concepts of surprisal and entropy allow for precisely this. While both concepts are formulated probabilistically, it is important to bear in mind that on the view adopted here, probability is simply an arbitrary mathematical measure of the subjective degree of belief ascribed to some outcome on the basis of a set of observations. Probability says nothing about *which* observations are relevant to phonologization and sound change. For this, we must draw on the results of linguistic study, such as those expressed by taking into account, for example, phonetic as well as statistical information. As we sketch just below, expressing results relating to these factors in terms of a combined measure of surprisal allows for the development of a unified model of language change.

We begin by considering how to incorporate perceptual distinctiveness into the measure of surprisal. For this, we follow the information-theoretic account of French epenthetic and deleted vowels in Hume et al. (2011). Of interest is the observation that the vowels in question are non-back and rounded [ø, œ], an apparent anomaly in the world's languages given that deleted/epenthetic vowels are typically front or central unrounded vowels. Hume et al. (2011) show that the patterning of the French vowels is consistent with universal patterns when we take seriously the view of language as a system shaped to meet the competing demands of efficiency and robustness in communication. In this approach, both deletion and epenthesis contribute to communicative effectiveness. Deleting a vowel enhances system efficiency by removing elements that contribute little to conveying the message. Conversely, epenthesis enhances system

robustness by helping to disambiguate low frequency structures, those with otherwise perceptually-masked cues, and/or those with a low probability of being accurately produced. As with deletion, the epenthetic sound contributes little to system entropy.

Perceptual distinctiveness is modeled as a function of miscategorization probability: the more a vowel's acoustic space overlaps with those of other vowels in the system, the higher the probability that the vowel will be miscategorized. Put another way, a high degree of overlap is correlated with poor perceptual distinctiveness and high confusability. A modified version of Nosofsky's 1986 *Generalized Context Model*, with frequency information factored out, was used for deriving categorization probabilities from a set of vowel tokens.

The result of applying the modified GCM is a ranking of sounds in a given context in terms of confusability. This is reminiscent of the P-map (Steriade 2008), though note crucially that we express an element's confusability in probabilistic terms and define confusability on a language-specific basis. By taking the negative logarithm of the resultant probabilities, we derive a surprisal value for each segment in question: an element with a high probability of being confused is associated with low surprisal, while an element with extremely noticeable cues, is associated with high surprisal. In terms of entropic contribution, elements with extremely high or extremely low surprisal contribute little to system entropy.

We can take a similar approach to production. Consider a scenario in which we are interested in evaluating the stability of word-final consonants $C_{\mathscr{L}}$ in a language $\mathscr{L}$ which includes the set of sounds {t, s, !}. Assuming, perhaps non-trivially, the availability of an independent measure of articulatory complexity, an 'A-map' of sorts, on the basis of which the members of $C_{\mathscr{L}}$ may be ranked in terms of probability of accurate production from least to most probable, ! $\prec$ s $\prec$ t, we predict that those elements with very complex or very simple articulations will be less stable than mid-range elements. Very simple elements will have very low surprisal associated with accurate production, and very complex elements will have high surprisal; in both cases they have small entropic contributions.

While our discussion above has briefly sketched out how some factors that condition phonological patterns can be recast within a model of surprisal, it seems reasonable to assume that other factors could also be defined in probabilistic terms. Moreover, we can go one step further and combine the various factors to create a unified model. In fact, in Hume et al.'s (2011) study of French epenthesis and deletion, it is only when the factors of frequency and perceptual distinctiveness are combined that the model correctly predicts the non-back rounded vowels to contribute least to the entropy of the system. When calculated independently, frequency and perceptual distinctiveness were only weakly predictive.

A unified model may take the following form. Let $V_{\mathscr{L}} = \{v_1, \ldots, v_i, \ldots, v_n\}$ represent the set of vowels from a language user's experience, $\mathscr{L}$, and e_i represent a context in which v_i may be observed (i.e. produced or perceived), an event we label

X. We assume that e_i is defined by grammatical factors relevant to v_i (e.g. 'between obstruents', 'in coda position', etc.) as well as statistically (e.g. n-gram frequencies of the grammatical elements). As described in Equation 6, the surprisal associated with v_i being accurately produced or identified is determined by the set of conditioning factors noted above, perhaps among others: confusability k, articulatory precision a, contextual frequency f, conditioning context e_i. Hence, as a first approximation, the surprisal $S(X = v_i)$ associated with the observation of a given element v_i is:

$$S(X = v_i) = -\log_2 P(X = v_i | k, a, f, e_i) \tag{6}$$

A segment's entropic contribution, $H_c(v_i)$, provides a measure of the degree to which that element is a factor in $\mathscr{L}$'s effectiveness as a system of communication.

How the various factors interact and contribute to the overall surprisal associated with a particular system is an important line of research yet beyond the scope of this chapter (though see Hume et al. 2011). As we discuss below, however, it is surprisal extremes that are of particular relevance to the present discussion, since elements at these ends are least stable and thus good candidates for phonologization. In this regard, it is reasonable to assume that extreme degrees of surprisal typically arise when the impact of several factors point to a common end of the continuum, although a single factor could potentially contribute sufficiently to determine the surprisal on its own.

One might ask why we need to talk about surprisal and entropic contribution, rather than simply limiting our discussion to probability itself. We can think of at least three reasons. First, although it is a formal measure, the quasi-metaphoric term 'surprisal' helps to evoke and preserve the intuition that we are discussing human cognition, and the impact of (socio)cognitive factors on phonologization and language change. Second, surprisal is a key component of the entropy of a set of possible outcomes (e.g. in a linguistic system), and it is the notion of entropy that allows us to provide a unified account of those elements that are prone to change. Third, Hume et al. (2011) show that probabilities based on confusability and frequency alone cannot predict the quality of the epenthetic vowel or deleted vowel in French. Rather, it is the entropic contribution based on these combined measures that correctly predicts the observed patterns.

2.2.3 *Surprisal and expectedness*

To the extent that we are correct in using surprisal to model a cognitive state, we might call this state (inverse) expectedness.[3] That is, a low degree of surprisal associated with some linguistic outcome in production, perception, and/or processing correlates with a high degree of expectedness. For example, a sound sequence that has a high

[3] We previously (cf. Hume and Broomberg 2005) used the term 'expectation' for this notion, but since this term overlaps with a concept from probability theory relevant to our discussion, we adopt the neologism *expectedness* in its stead.

probability of occurring, of having an articulation that is easy to produce accurately, and weak perceptual distinctiveness will, all else being equal, be associated with a low degree of surprisal (whether in production or perception) and greater expectedness. Conversely, high surprisal sequences (e.g. due to extreme perceptual distinctiveness, low frequency, complex articulation, etc.) will have weaker expectedness. These points are developed in greater detail below.

Expectedness has been studied (under a variety of names) extensively in fields such as psychology (e.g. Feather 1982; Hitchcock 1903; Kirsch 1999; Reading 2004), music cognition (e.g. Huron 2006; Jones, Johnston and Puente 2006), vision (e.g. Haith, Hazan, and Goodman 1988; Puri and Wojciulik 2008), and on language topics relating to sentence processing (e.g. Kutas and Hillyard 1984), computational modeling of language (e.g. Hale 2003; Jurafsky 2003; Levy 2008), and markedness (Hume 2004a, 2008).

Huron (2006) describes the biological roots of this notion as follows:

> Expectation refers to the cognitive function that helps fine-tune our minds and bodies to upcoming events . . . The biological purpose of expectation is to prepare an organism for the future . . . The capacity for forming accurate expectations about future events confers significant biological advantages. Those who can predict the future are better prepared to take advantage of opportunities and sidestep dangers. Over the past 500 million years or so, natural selection has favored the development of perceptual and cognitive systems that help organisms to anticipate future events . . . Accurate expectations are adaptive mental functions that allow organisms to prepare for appropriate action and perception.

Grossberg (2003) represents expectation in a neural network model as a resonant state of the brain:

> Such a resonance develops when bottom-up signals that are activated by environmental events interact with top-down expectations, or prototypes, that have been learned from prior experiences. The top-down expectations carry out a matching process that selects those combinations of bottom-up features that are consistent with the learned prototype while inhibiting those that are not. In this way, an attentional focus starts to develop that concentrates processing on those feature clusters that are deemed important on the basis of past experience. The attended feature clusters, in turn, reactivate the cycle of bottom-up and top-down signal exchange. This reciprocal exchange of signals eventually equilibrates in a resonant state that binds the attended features together into a coherent brain state. Such resonant states, rather than the activations that are due to bottom-up processing alone are proposed to be the brain events that represent conscious behavior.

Expectedness and thus, surprisal, have considerable explanatory force when it comes to understanding how phonetically variable material is transformed into phonologically meaningful units, an explanation that lies in the connection between expectedness/surprisal and attentional focus. As expressed in the quote from Grossberg (2003) above, expected outcomes yield an attentional focus that concentrates

on those elements (e.g. auditory cues) considered important on the basis of past experience (cf. Kirby (this volume) for a model of a diachronic shift in the weights given to various acoustic cues). Given that attentional focus is a crucial component of learning (e.g. Kruschke 2003; McKinley and Nosofsky 1996), it is directly relevant to phonologization, since for change to take place, the user must learn to associate phonological meaning with some phonetic detail. Further, since the resonant states that result from the interaction of expected outcomes and perceptual input are 'the brain events that represent conscious behavior', it is instrumental in shaping the form that behavior takes. This is of particular relevance for our understanding of phonologization, since although we often refer to the way that *languages* behave, it is in fact the behavior of the *language user* that is at issue. It is the individual who, for example, perceives the auditory cues that are subsequently phonologized as an epenthetic vowel, or fails to produce the gestures involved in making one sound as opposed to another.

It is perhaps worthwhile pointing out that while the discussion above has focused on phonetic, processing and usage factors, an additional advantage of the approach developed here is that it can be easily expanded to take into account other factors including e.g. sociolinguistic attributes and attitudes. For example, if a language variable, such as the pronunciation of [ŋ] in e.g. *running*, has a specific social meaning (Campbell-Kibler 2005), there are expectations associated with when and by whom the variable is used which can influence behavior including an individual's attitudes regarding its usage. We leave this topic open for future consideration.

2.3 Phonological effects of surprisal

We turn now to discuss more specifically why we believe surprisal is fundamental to phonologization and language change. Two properties of the current approach are particularly important: the relation between surprisal and instability, which provides insight into which elements are likely to be the targets of change, and the relation between surprisal and direction of change.

2.3.1 *Instability associated with the target of change*

An important prediction of the current approach is that change preferentially affects elements associated with extreme degrees of surprisal. The core insight here is that such extremes create phonological instability, as elaborated on just below. As is clear from Figures 2.1 and 2.3, what unifies these seemingly divergent cases is that elements with extreme degrees of surprisal, whether high or low, *contribute little to system entropy*. So the key prediction we derive is that elements that contribute little to predicting an outcome are less crucial for effective communication. As a result, they are more likely to be unstable, and thus prone to be the targets of diachronic change. They are, in a sense, more expendable.

In order to answer the question of why this might be so, we take any token of language use (i.e. any speaker-hearer interaction) to be an instantiation of a communication system striving (perhaps implicitly) to meet the competing demands of *efficiency* and *reliability*. The reliability of a communication system is a function of the degree of redundancy in transmitted elements. If symbols are on average highly redundant (i.e. recapitulating information available elsewhere), then they are more predictable/probable, and hence less informative (i.e. lower surprisal). Efficiency, conversely, is a function of a communication system's rate of transmission of information; increasing efficiency corresponds to transmitting more informative (i.e. higher surprisal) items on average. Consider now the effects of noise; a reliable system will in general be able to recover from an error in transmission, as the built-in redundancy ensures that the information lost is likely to be predictable from context, whereas a maximally efficient system, being non-redundant, makes no such guarantees, and hence is more adversely affected by transmission errors. The net result of striking a balance between the demands of reliability (maximal redundancy/predictability) and efficiency (minimal redundancy/predictability) is that elements that contribute significantly to the entropy of the system, those that are neither too surprising, nor too expected, are most important for effective or successful communication (see Lindblom 1990; Aylett and Turk 2004; Levy and Jaeger 2007; Jaeger 2010, for related discussion). Interestingly, while elements at opposite ends of the continuum pattern together in terms of being unstable, the cause of the instability differs, as discussed below.

2.3.1.1 *Low surprisal* Low surprisal elements are associated with high frequency, weak perceptual distinctiveness and simple articulations, among other properties. As is well documented, elements associated with these properties tend to be unstable. We acknowledge that isolating the effects of these properties may be a non-trivial enterprise.

In terms of perception, elements with poor perceptual distinctiveness can result in a failure to correctly parse the signal, which may result in assimilation or deletion (Jun 1995) and subsequent sound change. This is consistent with Ohala's (1981) thesis that an ambiguous signal can cause misperception giving rise to language change. In fact, the present account subsumes Ohala's proposal as a special case, given that low surprisal, on our account, can result not only from confusability, but from any of the factors listed immediately above, presumably among others. Production-related instability in cases of low surprisal may lead to, for example, reduction, deletion, or assimilation, a claim supported by the phonetic, phonological and psycholinguistic literature.

For example, words that occur frequently tend to be reduced, and high frequency sounds and sequences are prone to processes such as lenition, deletion, and assimilation, among others (cf. Bybee 2001, 2002; Bybee and Hopper 2001; Fosler-Lussier

and Morgan 1999; Frank and Jaeger 2008; Hooper 1976b; Jurafsky, Bell, Gregory, and Raymond 2001; Jurafsky 2003; Munson 2001; Neu 1980; Patterson and Connine 2001; Phillips 1984, 2001, 2006; Pierrehumbert 2001a; Raymond, Dautricourt, and Hume 2006; Tabor 1994; Zuraw 2003). Further, high frequency function words in English such as *just* and *and* have been found to undergo deletion of /t, d/ at significantly higher rates than less frequent words containing alveolar stops in comparable contexts (cf. Bybee 2001, 2002; Guy 1992; Jurafsky et al. 2001; Raymond et al. 2006). The result of phonological processes such as metathesis are also conditioned by frequency (Hume 2004b). Consistent with the current approach, changes often have their start in high frequency forms, subsequently spreading to other similar forms (see, e.g., Bybee 2001; Phillips 2006, *inter alia*).

It is worth pointing out that this approach is consistent with the observation that the more a routine is used, the more fluent it becomes (Bybee 2001, 2002; Phillips 2006; Zipf 1932). However, in the current approach changes are viewed as more than a practice effect. On our view, production, perception, and processing are guided by surprisal and expectedness, and we hypothesize that this grounds the physiological reflexes of practice in a cognitive explanation.

2.3.1.2 *High surprisal* High surprisal is associated with elements that occur with very low frequency, have complex articulations, and/or have extremely noticeable perceptual cues, among other factors. Given the link between surprisal and expectedness, when an element has high surprisal, its realization will correspondingly be only weakly expected by the language user. This, we suggest, gives rise to instability from both the speaker's and hearer's perspectives.

From a production perspective, it is well established that articulatory complexity can create instability, with phonological consequences taking the form of deletion, metathesis, assimilation, or other repairs to the unstable form. We provide an example from metathesis further below.

Very low frequency sequences are also unstable. Treiman et al. (2000), for example, found that English speakers made more errors in pronouncing syllables with less common rimes than those with more common rimes. Similarly, Dell (1990) reports that low frequency words are more vulnerable to errors in production than high frequency ones. Interestingly, when a form is unstable because aspects of its realization are unexpected, a speaker may also 'choose' to compensate by producing it more slowly and carefully. In this regard, Whalen (1991) found that infrequent words were longer in duration than frequent ones. The current approach is also consistent with the observation that low frequency is a factor associated with forms that undergo analogical change.[4] Phillips (2001, 2006), for example, presents numerous examples of change affecting low frequency items such as the case of [h] deletion in Old English

[4] In her study of analogical change in Croatian morphology, Sims (2005) shows frequency as well as social salience to be contributing factors, findings that are consistent with the current approach.

(Toon 1978): low frequency words underwent deletion first giving rise to *nut, ring, loaf*, from OE *hnutu, hring, hlaf*.

With respect to frequency, an interesting consequence of the current approach is that it provides a unified account of the observation that high and low frequency elements tend to lead language change (Bybee 2001; Phillips 1984, 2000). As discussed in subsection 2.2.1, frequency is a determinant of, and in direct proportion to, the probability assigned to a linguistic outcome, hence to its surprisal. To the extent that, all else being equal, low frequency correlates with high surprisal and high frequency corresponds to low surprisal (recall Figure 2.1), the current theory makes the strong and apparently correct prediction that high and low frequency elements will both be prone to change.

Metathesis provides an apt example showing low frequency and articulatory complexity contributing to instability, thus promoting change. In Hume's (2004b) study of 37 cases of consonant/consonant metathesis, low frequency of occurrence and similarity emerged as significant predictors of metathesis. In all cases, a consonant sequence that underwent metathesis was a non-occurring or infrequent structure in the language. In some cases, the word in which the sequence occurred was also uncommon, contributing an additional layer of surprisal to the sequence. Further, in over a third of the cases, the sounds involved were similar. Some shared the same manner or place of articulation, or agreed in sonorancy, differing only in place and/or manner, as attested in Georgian (Hewitt 1995; Butskhrikidze and Van de Weijer 2001), Chawchila (Newman 1944), and Aymara and Turkana (Dimmendal 1983), among other languages. The significance of similarity in the present context relates to the probability of accurate production. To the extent that sounds in a sequence are articulatorily similar, it is reasonable to expect an increase in the effort required to accurately produce and thus render each sound distinct.

A further prediction of the current approach is that elements with extremely distinctive cues will also be unstable. Clicks would seem to be an example of this type. The observation that clicks are typologically rare and do not seem to be spreading among language communities may provide some evidence for this prediction (A. Miller, p.c.).[5] However, our understanding of variable processes involving clicks and other high surprisal elements is incomplete at this time and thus, we leave this issue for future consideration. It is worth noting, however, that the patterning of sequences that are neither overly noticeable or unnoticeable lend support for the present approach in that they are predicted to be more stable than sounds/sequences at the extreme ends of the noticeability pole. We thus hypothesize that common sound sequences such as stop+vowel, sC, and other perceptually well-formed sequences, would be situated away from surprisal extremes.

[5] It is likely that articulatory complexity is also a factor, meaning that both articulatory and perceptual factors contribute to their high surprisal.

To summarize, in this section we have suggested that an approach drawing on considerations of communicative effectiveness provides a unified account of the patterning of elements with very high and very low degrees of surprisal. In both cases, they are predicted to contribute little to the entropy of the system and thus be less crucial for effectively communicating the message in question. In the following section, we focus on the role that surprisal plays in biasing the outcomes of phonological change.

2.3.2 *The output of change*

The current approach also speaks to the nature of the change affecting unstable language patterns. As stated above, the degree to which particular linguistic elements are expected guides processing, perception, and production. As a result, to the extent that these expectations are biased in one direction or another, we would expect there to be linguistic consequences (for related discussion see Pierrehumbert 2001a; Wedel 2007). For example, if a linguistic item has properties that are strongly expected in a given context, processing should be faster since the listener will be biased toward perceiving the item. This is supported by findings that high frequency words and words containing frequent sound sequences are processed more rapidly than infrequent ones (see Oldfield and Wingfield 1965; Jescheniak and Levelt 1994; Vitevitch, Luce, Charles-Luce, and Kemmerer 1997, among others).

The observation that expectations bias perception is not limited to language. Kirsch (1999) presents an amusing case relating to visual perception.

> When stimuli are ambiguous enough, sets of expectancies can lead to their being misperceived, even when they are examined slowly and carefully. For example, when 17th- and 18th-century biologists who believed in preformation examined sperm under the microscope, they reported seeing fully formed miniature beings. They saw miniature horses in the sperm of a horse, tiny chickens in the sperm of a rooster, and minuscule human babies in human sperm. The ambiguity of the stimulus allowed them to see whatever they expected to see. (Kirsch, 1999: 6)

As in the vision example above, the influence of bias is particularly strong in contexts of ambiguity, such as low surprisal sequences with weak perceptual distinctiveness. Bias also influences the outcome of high surprisal sequences, such as those associated with very low frequency or considerable articulatory complexity. In both cases, bias drives the sequences away from the surprisal extremes. That is, a high surprisal sequence due to, for example, articulatory complexity will be realized as one with less complexity. Conversely, a low surprisal sequence due to weak perceptual distinctiveness, will generally be replaced by one with more distinct cues. In each case, the sequences in question end up contributing more to system entropy and thus, to communicative effectiveness.

Pitt and McQueen (1998), for example, found that the transitional probabilities of voiceless alveolar and postalveolar fricatives at the end of nonwords influenced listeners' identification of an ambiguous fricative as well as that of the following stop

consonant; subjects were biased toward the fricative with the highest transitional probability. This is also consistent with the findings of Vitevitch and Luce (1999), which reveal segment and sound sequence probabilities to be most influential when listeners are presented with unfamiliar words; that is, high surprisal words. The observation that bias is especially strong in cases of high surprisal is of particular relevance to understanding phonologization. It predicts that if an item is unstable because of high surprisal, it will be prone to subsequent change to a pattern with lower surprisal; that is, it will be biased in the direction of a more expected pattern. This is exactly the pattern of change observed in cases of analogical change.

The study of metathesis once again provides an appropriate example. As noted above, sequences prone to metathesis are those associated with high surprisal due to a low probability of accurate production, and the user's limited experience or lack of experience with the sequence (and perhaps the word it occurs in as well). As predicted, the direction of change is biased toward a more expected structure with lower surprisal. As the study of metathesis shows, the resultant structure is not only more common than the form that undergoes metathesis, but it has a higher probability of being accurately produced, resulting in better perceptual cues. Building on Hume (2004b), the reason why improved perceptual salience is a characteristic of so many results of metathesis is thus simply an artifact of the nature of sequences that undergo metathesis (those associated with high surprisal) and those that influence how the speech signal is parsed (those associated with low surprisal); in short, unstable sequences that undergo metathesis are biased toward phonologically similar patterns with lower surprisal. Variable pronunciations of the word *chipotle* provide a simple illustration:

> The influence of native language patterns on metathesis can also be heard in some varieties of American English in the variable pronunciation of *t-l* in the word, *chipotle*, the [Nahuatl] name for a particular kind of pepper and, recently, for a chain of Mexican restaurants. Both orders of the final two consonants can be heard, even in the speech of the same individual: *chipotle* (the original order) or *chipolte* (the innovative order) [...] The two sounds involved are archetypical 'metathesis sounds' and thus contribute to indeterminacy: /t/ with perceptually vulnerable cues and /l/ with stretched out features [...] Another factor [...] is unfamiliarity with the borrowed word [...] With indeterminacy, the order of sounds is inferred based on experience, with the bias towards the most robust order. As predicted, although both /tl/ and /lt/ occur intervocalically in English [...] /tl/, in the original form, occurs in 67 words, while the innovative /lt/ sequence occurs in 356 words. (Hume, 2004b: 223)

An interesting corollary of the influence of bias on the outcome of change concerns the notion of *structure preservation* (Kiparsky 1985, 1995). When change occurs in the direction of a low surprisal pattern, as it does with unstable high surprisal elements, such changes will, *ceteris paribus*, be structure preserving; for a pattern to have relatively low surprisal, i.e. to be relatively more expected, a user must already be familiar with it, that is, it must already be part of the user's linguistic experience. Cases of

analogical change and the observation that the output of metathesis is an existing structure in the relevant language support this view.

Conversely, the result of change involving unstable patterns with low surprisal need not be structure preserving. In such cases, the linguistic consequence of high expectedness is under-realization; that is, a pattern contributes little to the entropy of the system and is thus less crucial to the message. As discussed above, these elements can thus be reduced in the interests of communicative efficiency without sacrificing reliability.

An example of non-structure-preservation comes from the observation that reduction processes involving low surprisal segments, such as English schwa, can create syllable structures not otherwise occurring in the language. Schwa can be considered a low surprisal element given its simple articulation, its poor distinctiveness, its predictability in unstressed syllables, and its overall high frequency of occurrence in the language (Hume and Broomberg 2005). As such, a native speaker will have strong expectations concerning the occurrence of schwa in the initial unstressed syllable of a word such as *telepathy*, thus licensing its omission, i.e. [tlɛpəθi]. While schwa deletion can result in phonotactically licit syllable onsets (e.g. *police* [plis]), it can also create onsets such as [tl], which do not otherwise occur word-initially in the language.

2.3.3 *Summary*

The ideas presented above are summarized in Table 2.1. It is proposed that a language pattern is prone to change when, as listed in column I, it has a very low or very high degree of surprisal and thus contributes little to the entropy of the linguistic system. Column II identifies some of the factors that can give rise to the relevant level of surprisal. The rightmost column summarizes the discussion above concerning bias and the nature of the outcome of language change. For patterns that are unstable due to

Table 2.1 Overview of relations between surprisal, conditioning factors, and change

I: Surprisal	II: Influencing factors	III: Outcome of change
high	low familiarity, low frequency, strong perceptual distinctiveness, complex articulation	change biased toward similar low-surprisal pattern (structure preserving)
low	high familiarity, high frequency, weak perceptual distinctiveness, simple articulation	change can be unbiased (need not be structure preserving)

high surprisal, bias influences the direction of change, while for unstable low surprisal elements, the outcome of change takes some form of reduction which can result in an increase in entropic contribution.

2.4 Conclusion

As we hope to have shown in the preceding pages, taking into account communicative effectiveness, as formally expressed in terms of surprisal and entropy, allows us a deeper understanding of phonologization and language change. To the extent that this approach is on the right track, it has the potential to provide a unified model of the factors conditioning an individual's language system. Given that the preceding pages offer only a sketch of the current theory, many important aspects remain unresolved. These include at least the following fundamental issues: (a) understanding how the diverse factors interact and contribute to cognitively and linguistically plausible estimates of an element's surprisal, and (b) identifying the consequences of differing degrees of surprisal and entropy for language systems, at the segmental level and beyond.