Linguistic Society of America Annual Meeting, January 7, 2006
Colloquium on Information Theory & Phonology

Language Specific and Universal Markedness: An Information-theoretic Approach

Elizabeth Hume (ehume@ling.osu.edu)
Ohio State University

I. Introduction

(1) The goal of this talk is to show how the tools of Information Theory enable us to resolve some key problems with the concept of markedness. While this paper focuses largely on markedness as it relates to sound patterns, the core ideas can easily be extended to other aspects of language as well.

(2) Some common markedness claims regarding speech sounds and sound patterns.
Compared to its marked counterpart, an unmarked sound is generally assumed to be:
a. more frequent across languages
b. the target of processes such as reduction, deletion, assimilation, metathesis; the undergoer of sound change
c. the output of processes such as epenthesis, neutralization; the result of sound change
d. more phonetically variable
e. more widely distributed
f. acquired earlier
g. preserved in the formation of creoles

These claims are generally considered some of the diagnostics of markedness.

(3) Most explanations of these diagnostics have focused on phonetics. Other explanations such as functional load and frequency have received less attention (though see, e.g. Trubetzkoy 1939, Greenberg 1966, Battistella 1990).
a. Articulatory ease. Sounds that are easier to articulate are generally assumed to be:
- more frequent cross-linguistically, e.g. [t] vs. [ʧ]
- less stable phonetically, e.g. more variable
- more likely to undergo reduction, deletion, assimilation
- the result of neutralization (Korean: aspirated, tense and plain stop are neutralized in coda position. Only the plain stop occurs.)
- acquired earlier
- etc.

Acknowledgements: I am grateful to the many people who have given me input on the ideas expressed in this paper. This includes Ilana Bromberg, Nick Clements, Kathleen Currie Hall, Mary Beckman, Chris Brew, John Goldsmith, Keith Johnson, Patrick Lee, Anton Rytting, members of the OSU phonetics/phonology discussion group (aka Phonies), and participants at LabPhon 2004, MOT 2005, and the 7th Annual Meeting of the French Network of Phonology 2005. I would also like to express my great appreciation to Kathleen Currie Hall for her willingness to present this paper on my behalf.
Many, many linguists have worked on markedness and I apologize in advance for not listing all relevant references in this handout; a more complete list will accompany the written version. I alone am responsible for any and all errors or omissions. Comments welcome.

b. Perceptual salience.
Low salience: Sounds/sequences with poor phonetic cues are generally assumed to be:
- less stable phonetically and thus subject to, e.g. assimilation, deletion, metathesis

High salience: Sounds/sequences with good phonetic cues are generally assumed to be:
- more frequent across languages, e.g. CV syllables vs. VC syllables; [s] vs. other fricatives

II. Some Problems

(4) **Problem 1.** Markedness diagnostics can be paradoxical.
unmarked = most likely to be deleted = **least** preferred in system
unmarked = most likely to be inserted = **most** preferred in system

(5) **Problem 2.** Patterns contradicting the markedness diagnostics in (1) are commonly observed.
• Target of assimilation (Hume & Tserdanelis 2002; Hume 2003)
• Passive neutralization (Hume 2003)
• Underspecification and default values (Steriade 1987, 1995; Clements 1988, 1993; Mohanan 1991, Odden 1992, Calabrese 1995)
• Neutralization (Odden, p.c.)
• Epenthesis (e.g., Vaux 200x, Hume & Bromberg 2005)
• Cross-language phoneme frequency (Battistella 1990, Rice 2000)
• Sound change (e.g. Lass 1976, Ladefoged 1984)
• Creole genesis (Thomason 1993) e.g. Chinook Jargon (a pidgin) contains a stable phonemic inventory which includes glottalized, labialized and uvular (vs. velar) stops.
• Child language acquisition (e.g. Menn 1984; Beckman et al. 2003) e.g. Quiché-learning children master /ʧ/ at an earlier age than English-learning children.

(6) **Problem 3.** Markedness makes predictions about universal patterns but is silent about what is predicted at the level of the individual language.

(7) **Problem 4.** Explanations for the diagnostics can seem contradictory.
i. **Low perceptual salience** is commonly used to explain why certain sounds or sequences of sounds are more apt than other sounds to undergo phonological processes such as deletion, dissimilation, metathesis, and assimilation: sounds with poor phonetic cues tends to be phonetically unstable and thus undergo change. In this case, unmarkedness is associated with poor phonetic cues (weak salience).

ii. **High perceptual salience** is used to explain why certain sounds or sequences are more apt to recur across languages: Sounds/sequences with good phonetic cues (high salience):
- are more stable and occur frequently in languages, e.g. CV vs. VC syllable type; [s]; sC-, -Cs clusters

The problem is that unmarked is defined in terms of both high salience and low salience. It is therefore impossible to tell *a priori* whether a sound should be considered marked or unmarked given its salience.

(8) **Problem 5.** Lack of an objective definition.
There is no objective definition of what would validate or falsify markedness. Is markedness falsified if the evidence from different diagnostics or their explanations conflict? If not, how do we determine the markedness status of the element? For example, a sound that is perceptually salient may be relatively simple from a production perspective. Does this make it marked in a system or unmarked? What if certain sounds are frequent in one context cross-linguistically but less frequent in another?

Given the lack of an objective statement of what markedness is, the concept tends to be used as is convenient; patterns that support a particular view of markedness are used as evidence for that view while those that do not, tend to be labeled 'irregular', 'exceptional', 'irrelevant'. The result is that markedness is not predictive; it is not a scientific concept.

(9) **Problem 6.** What is the basis of markedness?
Markedness is a descriptive label, rather than a scientific concept. As such, it does not explain anything; it is something that needs to be explained (Menn 1984). If our goal is to develop a predictive model, we first need to understand what is at the basis of markedness and its diagnostics.

III. Proposal

(10) I suggest that **information content** and **entropy** are better predictors of so-called markedness patterns.

Information content is a tool of INFORMATION THEORY (Shannon 1949) and well-established in the field of computational linguistics, though less familiar to theoretical linguistics (though see, e.g., Hockett 1955, Broe 1996, Goldsmith 1999, 2002, Hume & Bromberg 2005 for discussion concerning the role of Information Theory in phonology.)
• It does not refer to the concept of knowledge, as in the popular use of the word (Shannon 1949; see also Zipf 1932, Pierce 1980, Applebaum 1996).
• Rather, information content is a measure of the surprise value or probability that a particular element has within a given communication system.
The greater the probability of some element in a system, the lower its information value.

(11) To illustrate, consider a language with 5 vowels (i, e, a, o, u). When we say that one vowel (e.g. [o]) has lower information content than another (e.g. [a]), we mean that the probability of the vowel [o] occurring in some context in the language is greater than it is for [a]. That is, we would be less surprised to see [o] than we would be to see [a] in that context.

Higher Probability → Lower Information Content

(12) Calculating Information Content and Entropy:

a. Given a particular linguistic element ψ_0 (word, sound, feature, etc.) in some language (L), the information content (I) of ψ_0 is equal to the negative log probability of ψ_0. Following Shannon (1949), this is formalized as:

$$I(\psi_0) = - \log_2 (p(\psi_0))$$

b. The entropy (H) of a system, such as language L, consisting of *n* distinct elements (ψ_1, ψ_2, … , ψ_n), is the sum of the information content of each element in the system, defined as

$$H(L) = - \sum_{i=1}^{n} p(\psi_i) \log_2 p(\psi_i)$$

(13) Why use Information Theory?

a. Information theory is a well-established theory of communication.
b. Information content and entropy are quantifiable; markedness is not.
c. Using the log value instead of simple probability gives us more manageable values and facilitates the combination of different determining factors, as discussed below.
d. Probability is at the heart of information content and entropy. There is considerable psycholinguistic evidence showing that language users are sensitive to probabilistic information about the words, sounds and sound combinations of their language (e.g. Bybee 1985, 2001, Dell et al. 2000, Frisch 1996, Frisch et al. 2000, Luce 1986, Lindblom 1990, Pierrehumbert 1994, Pitt and McQueen 1998, Saffran, Aslin and Newport 1996, Saffran, Newport and Aslin 1996, Vitevitch and Luce 1999, Makashay 2001).
e. Information theory gives us the mathematical tools to develop a rigorous theory of markedness which, importantly, makes the right predictions.

(14) Making use of information content and entropy allows us to talk about both the markedness of **individual elements within a system** (the information content of a given element), as well as the markedness of **language systems** (the information content of the entire system, better known as the entropy of the system). The former allows the comparison of elements within a system (e.g. is [a] more marked than [o]? …) while the latter allows the comparison of entire language systems (e.g. is one language less marked than another? easier to learn? etc.). In this talk, the discussion is limited to the comparison of individual elements and thus focuses largely on information content.

Recall Problem 1: The seemingly paradoxical nature of the unmarked, e.g. the unmarked sound is both the most likely to delete but also the most likely to be inserted.

(15) Probability: These effects are precisely what is predicted within an information theoretic approach. Recall that probability is at the heart of information content: $I = - \log_2 (p(\psi))$

(16) I suggest that high probability elements, e.g. sounds, share two properties:

a. *Instability*: An element with high probability in a language system (or a given context in that system) is less stable, e.g. more variable, than one with lower probability. This is because there is a greater degree of expectation by the language user that the more highly

probable element will occur. That is, its presence is more predictable, regardless of whether it is clearly present or not.

b. *Bias*: There is a greater (perceptual/articulatory) bias towards an element with higher probability than there is towards one with lower probability. Again, this is because there is a greater degree of expectation by the language user that the more highly probable element will occur in the relevant context.

(17) **Illustration:**
Compare the words 'for' and 'four' in the following sentences:

a. *I was so hungry when I got home from work last night. I ate* [*f or*] *hours.*

b. *I was so hungry when I got home from work last night. I ate* [*f our*] *pizzas.*

Instability: 'for' is less stable than 'four', i.e. it will tend to be reduced, and may even delete.
Bias: There is a greater perceptual/articulatory bias towards 'for'. For example, if the words 'for' and 'four' were masked in the sentences above, it is likely that listeners would be able to recover 'for' more easily than 'four'.

Why?
- 'for' has a **high probability** of occurring in this context, higher than most words in English including 'big, an, apple, four', etc.
- 'four', on the other hand, has a **lower probability** of occurring in this context since many other words could just as easily occur, e.g. 'ten, more, five', etc.

High probability is the cause. Reduction, deletion, perceptual recovery, etc. are effects.

(18) Resolving the deletion/epenthesis paradox: How can the same sound, (e.g. schwa in English, [ø/œ] in French) both be most likely to be deleted and most likely to be inserted (Hume & Bromberg 2005)?
- Deletion is a result of instability.
- Epenthesis is a result of bias.

The factor unifying both is high probability: the sound in question has the highest probability in the language (or in a particular context).

Determining Probability: What makes one sound more probable than another? To determine this, we must draw on our knowledge as linguists.

(19) As a starting point, I draw on the following assumptions as a basis for determining the factors relevant to information content:
a. Human physiological constraints relating to:
- the auditory/visual perception of the speech signal, and
- the production of the linguistic elements in question

b. A person's knowledge of language includes information about the language's:
- syllable structure
- sounds/features
- phonotactics
- contrast
- patterns of usage (frequency of sounds, sequences of sounds, syllable types, words, neighborhood effects, etc.)

c. A person's knowledge of language includes information about the social value of elements in the language system(s) that they are familiar with.

The Model: An overview

(20) Based on these assumptions, I hypothesize that there are at least five factors involved in determining the probability and hence, the information content, of a sound or sequence of sounds in a language. (The first four factors make use of Bayesian probabilities, based on our knowledge as linguists, while the fifth uses traditional frequentist probability.)

a. **Precision:** The probability that a sound will be produced correctly. Precision is influenced by the complexity and nature of the articulation (stops require less precision than fricatives), the context in which it occurs, as well as how similar the sound is to other sounds in the inventory (dissimilarity with other sounds will typically require less precision).

Less precision required → higher probability of being produced correctly

b. **Cue quality:** The probability that a sound will be confused with another sound. Cue quality is influenced by the nature of the sound, the context in which it occurs, and the acoustic/auditory similarity that the sound has with others in the inventory.

Poorer cue quality → higher probability of being confused with another sound

c. **Contrast:** The probability that a sound has a low impact on distinguishing meaning in the language, as determined by the number of contrasts in which it is involved (cf. functional load, e.g. see Hockett 1955, Surendran & Levow 2004, Surendran & Niyogi 2003).

Lower degree of contrast → higher probability of having low impact on meaning

d. **Contextual frequency:** The probability that a segment appears in a given context (or system), as determined by frequency count (type, token, mutual information, etc.).

Higher frequency → higher probability of occurring

A number of traditional markedness diagnostics can be thought of as providing evidence for the frequency of an element within a system (or context), which in turn contributes to the element's probability and total information value:

- broader distribution (a more widely distributed element will tend to be have a higher frequency – type frequency, at least)
- high salience (such sounds are more resistant to change and so tend to recur in language, which in turn may make them more frequent)

e. **Social value:** The probability that a sound (or some other ling. element) is a marker of social identity (age, social class, geographical region, gender, etc.).

Lower social value → higher probability

(21) Each of the five factors in (20) is relevant in determining the probability of a sound occurring in a language, and thus, that sound's information content. The total value of a sound's information content is the sum of the information content contributed by each factor, where I = information content and ψ represents a speech sound in some language (L). (For simplicity, I assume that all factors are equally weighted and independent though further research will probably show that these assumptions need to be modified.)

$$I_{total}(\psi) = I_{prec}(\psi) + I_{cues}(\psi) + I_{contrast}(\psi) + I_{soc.value}(\psi) + I_{freq}(\psi).$$

Returning to the problems in section II.

(22) **Problem 2.** Patterns contradicting the markedness diagnostics in (1) are commonly observed.
Problem 3. Markedness predicts universal patterns but typically is silent about what is predicted on a language specific basis.

Solution: The proposed model predicts cross-linguistic variability in the patterning of sounds. That is, a sound may be marked cross-linguistically, but unmarked within an individual language. The reasons:

Language systems may differ in a way that has a bearing on any of the five factors due to differences in, for example, the inventory of sounds, the phonetic properties of phonologically similar sounds, the social value of the various sounds, to what extent the elements are used, etc.

This means that the values of each of the five factors contributing to information content can differ from language to language. E.g., if sounds in two different languages have similar values when it comes to degree of contrast, frequency, social value and precision, it will be potential differences in the cue quality of the relevant sounds that will influence their phonological patterning. Or, if they are basically the same with regards to all factors except precision, it will be this last property that will have the greatest influence. To the extent that the value of a particular factor differs across languages, cross-linguistic variability is predicted.

(23) **Example: Epenthesis**

Observations regarding the quality of epenthetic vowels:

a. There are certain vowel types that recurrently surface as epenthetic across languages, e.g. [ə] (e.g. English, German), [i] (e.g. Maltese, Japanese), [ɨ] (Korean, Japanese), [e] (e.g. Spanish). Other types are less common, e.g. low vowels (outside of a pharyngeal context), rounded vowels (outside of a labial context).

b. However, at the level of the individual language, so-called marked vowels can be epenthetic, e.g. mid front rounded vowel (French), low central vowel (Axininca Campa).

Hume & Bromberg (2005) show that the proposed model is able to correctly predict differing vowel qualities in English (schwa) and French (mid front rounded vowel). In both cases, the epenthetic vowel is the vowel with the lowest information content in the language.

Unlike a traditional markedness approach, the model also predicts that two languages with the same phoneme inventory and the same patterns of contrast could have different epenthetic vowels. The difference would be determined by frequency or social value. Spanish and Maltese are currently being compared as a potential test case. The vowel systems of both languages include [i, e, a, o, u] yet the epenthetic vowel in Spanish is [e], while in Maltese it is [i]. All else being equal, we predict [e] and [i] to be the vowels with the lowest information content in Spanish and Maltese, respectively.

c. The observation that certain vowel types recur as epenthetic across languages also falls out from the model. Vowels that require less precision or are more confusable will have low information values for these factors, as compared to other vowels in the system. Vowels like schwa, and short front and central non-low vowels are good examples. Thus, all else being equal, of all vowels in a system, these vowels are favored to be unmarked in language. This would then explain the observation that they are among the most common epenthetic vowels across languages. But a universally marked vowel, like the mid, front rounded vowel, can also emerge as unmarked in a language if the combined values of all five factors give it a lower information value than that of other vowels in the system. For example, the exceptional high frequency of an otherwise marked vowel or the vowel's similarity to other vowels in the system may offset the contribution of other factors. As shown in Hume & Bromberg, this is precisely the case in French where the epenthetic mid front rounded vowels are not only among the most frequent in the language, they share features such as rounding with many other vowels in the language which contributes to their confusability.

(24) **Problem 4:** Explanations for the markedness diagnostics in (1) can be contradictory. With respect to perceptual salience, **unmarked** is defined simultaneously in terms of **low salience** and **high salience**.

Probability provides the explanation:

Low salience, or poor cue quality, is one of the factors used to calculate the probability of an element in a system. Poorer cue quality correlates with higher probability so these sounds will have the properties of instability and bias and show all the effects associated with these properties (e.g. *instability*: the undergoer of phonological processes like deletion, reduction,

assimilation, metathesis; *bias*: the outcome of phonological processes like epenthesis, neutralization)

High salience has been proposed as an explanation for the pervasive presence of certain sounds and structures, e.g. CV syllables, clusters with [s], etc. This is because sounds and sequences with richer cues tend to be more stable in a language and consequently, typically occur in more words than those with weaker cues (Makashay 2001). All else being equal, sounds/sequences with higher salience will be more frequent in a system and so contribute to the element's probability.

Thus, the apparent contradiction is because "salience" can contribute to two DIFFERENT information content factors and so results in different findings: it contributes to both cue quality and frequency, in opposite directions.

(25) **Problem 5.** Lack of an objective definition: There is no objective definition of what would validate or falsify markedness.

Solution: Information theory provides us with a predictive, falsifiable model of markedness. The unmarked element is the element with the highest probability, and thus lowest information content, within the system in question, with probability defined in terms of the five proposed factors.

$$I(\psi) = I(\psi)_{prec.} + I(\psi)_{cues} + I(\psi)_{contrast} + I(\psi)_{soc.value} + I(\psi)_{freq.}$$

(26) **Problem 6:** What is the basis of markedness?
Answer: Information content (probability) is the unifying concept that underlies the diagnostics in (2) and is the basis of markedness. Traditional markedness diagnostics either provide evidence for the effect of probability on an element and/or they contribute to the element's probability within the system.

a. The following diagnostics show the effects of high probability (low information content) and thus give evidence for the high probability of the element in question.

instability

- more phonetically unstable than the marked one
- the target of phonological processes, e.g. assimilation, deletion, reduction, metathesis
- the undergoer of sound change

bias

- the output of phonological processes epenthesis, neutralization, metathesis
- the result of sound change
- acquired earlier than a marked sound
- preserved in the formation of creoles
- more frequent across languages than the marked counterpart

b. Some diagnostics (or explanations of the diagnostics) also are influential in determining the probability of a particular factor.
   - Precision: articulatory simplicity
   - Contextual frequency: broader distribution; high salience

(27) Other advantages:

a. Information content is quantifiable, thus moving us closer to a scientifically rigorous theory of the observations. Markedness is not quantifiable.

b. Markedness observations are predicted directly using a single metric, information content. Markedness is a descriptive label.

c. Information content, predicts both the **instability** of an element as well as a **bias** towards that same element. In both cases, the element in question has **low information content**.

d. While the discussion here has been limited to the comparison of individual elements (e.g. speech sounds), Information Theory also provides the tools (entropy) to evaluate different language systems.

(29) To conclude, this paper has provided the outline of an information-theoretic theory of markedness. It has been proposed that by taking into account probabilistic information, we are able to develop a predictive model of markedness, able to account for both universal and language specific patterns.

References:
Please contact the author for references.