Phonological Relationships: A Probabilistic Model*

Kathleen Currie Hall
CUNY: College of Staten Island and the Graduate Center

SUMMARY

This paper presents a probabilistic model of phonological relationships, recasting the traditional relationships of "contrast" and "allophony" in terms of a gradient scale of predictability, based on the Information-Theoretic concept of entropy (uncertainty). The model is applied to the case of [ɑɪ] and [ʌɪ] in Canadian English to demonstrate its utility in describing relationships that are neither perfectly allophonic nor fully contrastive.

RÉSUMÉ

Cet article présente un modèle probabiliste de relations phonologiques, en reformulant les relations traditionnels de "contraste" et "allophonie" en termes d'une échelle de prévisibilité gradiente, basée sur le concept de la théorie de l'information d'entropie (l'incertitude). Le modèle est appliqué au cas de [ɑɪ] et [ʌɪ] en anglais canadien pour démontrer son utilité dans la description des relations qui ne sont ni tout à fait allophoniques ni entièrement contrastives.

1 INTRODUCTION

The twentieth century saw quite a number of developments in phonology, with the focus early in the century on structural phonemic analysis giving way to a focus on phonology as a generative system of representations and rules and later to a system of constraints on possible outputs. Despite this evolution in the way in which phonology was conceptualized, one of the key foundations on which phonology rests has remained fairly constant: within the large amount of phonetic variability that

* Thanks go especially to Mary Beckman, Chris Brew, Cynthia Clopper, Beth Hume, Becca Morley, and audiences at the 2008 MOT Phonology Workshop, the 2008 LabPhon conference, SUNY-Stony Brook, the CUNY Graduate Center, and the 2011 conference on Phonology in the 21^{st} Century for comments on earlier versions of this work.

McGill Working Papers in Linguistics, 22.1, Winter 2012

exists in language use, phonology is concerned with the categories of sounds that play a meaningful role in organizing the structure of language. Two particularly important insights of phonology in the twentieth century were that (1) the ability to define phonological relationships (contrast, allophony) is crucial to the determination of phonological patterns in language (see, e.g., Goldsmith (1995)) and (2) the notion of predictability of distribution is one of the key tools phonologists should use in determining phonological relationships (see, e.g., Steriade (2007) for a review of much of the previous literature on contrast).

There are, however, a number of cases that are problematic for the usual classifications of units as being either "predictably distributed" (allophonic) or "unpredictably distributed" (contrastive). One example is the long-standing debate about the status of the vowels [ɑɪ] and [ʌɪ] in Canadian English: Should they be considered allophonic because they are predictably distributed in most environments, or contrastive because they are unpredictably distributed before a flap in (near) minimal pairs such as *idol* [ɑɪɾl̩] vs. *title* [tʌɪɾl̩]? This paper presents a model of phonological relationships, the Probabilistic Phonological Relationship Model (PPRM), that precisely quantifies the degree to which two phonological units are predictably distributed in a language. The model builds on insights from information theory, originally developed in the twentieth century but not entirely practical to apply to linguistic data before recent advances in corpus collection and statistical analysis. The model thus allows for a more insightful analysis of the actual relationship that might exist between two sound categories in a language.

2 A NEED FOR A PROBABILISTIC MODEL?

2.1 INTERMEDIATE PHONOLOGICAL RELATIONSHIPS

There are quite a number of different instances in which the traditional relationships of "contrast" and "allophony" do not quite seem to adequately account for actual phonological data. A number of researchers have come up with various terms for such intermediate cases, such as "quasi-allophonic" (e.g., Collins and Mees (1991); Rose and King (2007); Bye (2009)), "semi-phonemic" (e.g., Bloomfield (1939); Crowley (1998)), "marginally contrastive" (e.g., Gleason (1961); Goldsmith (1995); McMahon (2000); Zuraw (2002); Ash (2003); Blust (2003); Kiparsky (2003); Scharf and Hyman (2011)), or "fuzzy contrasts" (e.g., Scobbie and Stuart-Smith (2008)). Though the reasons for any particular relationship to be intermediate between contrast and allophony vary (see Hall, submitted, for a typology of intermediate relationships), one primary reason is that predictability of distribution – traditionally construed as a binary and categorical criterion for determining phonological relationships – can often appear to analysts as being better construed as gradient. The following section both describes the traditional interpretation of predictability of distribution and gives examples of how it can be probabilistic.

2.2 PREDICTABILITY OF DISTRIBUTION

One of the primary criteria for determining phonological relationships is that of predictability of distribution. It has been assumed that two sounds are allophonic if and only if they are predictably distributed in all environments–that is, if it is always possible to predict which of two sounds occurs

in any given environment of the language, based solely on knowing the environment (and not the lexical identity of the word). If there is any degree of unpredictability–for example, a single context in which it is not possible to make such a prediction–the two sounds are traditionally assumed to be contrastive.

There are instances, however, in which two sounds may be mostly predictable, with some degree of unpredictability, or mostly unpredictable, with some degree of predictability, and phonologists have often hesitated to label these cases with the broad name of "contrast." For example, if two sounds have historically been allophonically related, but are undergoing a phonemic split, they may be partially (even mostly) predictably distributed, and that predictability of distribution may be an active part of the generative grammar that speakers of the language make use of. One example of such only slight unpredictability is found with Canadian Raising, a phenomenon that has been reported for many dialects of English, both within and without Canada (e.g., Joos (1942); Chambers (1973, 1989); Trudgill (1985); Vance (1987); Allen (1989); Britain (1997); Trentman (2004); Fruehwald (2007)). The diphthongs [ɑɪ] and [ʌɪ] are generally predictably distributed, with [ʌɪ] occurring before tautosyllabic, tautomorphemic voiceless segments and [ɑɪ] occurring elsewhere (e.g., tight [tʌɪt] but tide [tɑɪd]). There are, however, surface (near) minimal pairs distinguished by the two vowels, such as *writing* [rʌɪrɪŋ] and *riding* [rɑɪrɪŋ] , or *title* [tʌɪɾl̩] and *idol* [ɑɪɾl̩] , in which the two systematically contrast before a flap [ɾ] . Given the presence of such (near) minimal pairs, it has been argued that [ɑɪ] and [ʌɪ] are contrastive in Canadian English (and other similar dialects) (see, e.g., Mielke et al. (2003); Kaye (2009)), but others have been reluctant to relinquish the status of the two as allophonic, largely because the pattern is actively productive in nonsense words (e.g., Bermúdez-Otero (2003); Boersma and Pater (2007); Idsardi (2006)) (or because the process of flapping itself is assumed to be predictable, thus allowing the vowel quality to be predicted from the underlying representation).

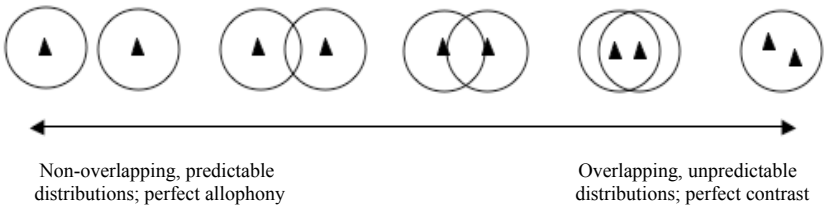
Other examples of relationships that seems to be intermediate between contrast and allophony on the basis of predictability of distribution include:

- Mandarin: Hume and Johnson (2003) provide experimental evidence from perception that the predictable neutralization of otherwise contrastive tones 214 and 35 in Mandarin leads to the perception of them as particularly similar. They term such neutralizable contrasts "partial contrasts."

- Spanish: Hualde (2004) describes a number of what he calls "quasi-phonemic" relationships in Spanish. For example, [r] and [ɾ] contrast between vowels but are predictably distributed elsewhere, so there is debate about whether they are allophonic or contrastive.

- Philadelphia English: Labov (1994) describes a relationship in certain American English dialects including Philadelphia English that Moren (2004) dubs "semi-allophonic." The general pattern is that the lax allophone of /æ/ occurs before voiced stops (e.g., in cad [kæd]) while tense allophone of /æ/ occurs before the voiceless fricatives [f, θ, s] (and some other contexts) (e.g., in calf [kɛəf]). There are lexical exceptions to this generally predictable pattern, however, such that certain words that end in a voiced stop, including mad, bad, and glad, have a tense /æ/ – [mɛəd], [bɛəd], [glɛəd] – creating near minimal pairs like cad [kæd] vs. bad [bɛəd].

Figure 1: Continuum of predictability of distribution on which the PPRM is based



Such examples are widespread. Hualde (2004: 20) says that "there are areas of fuzziness probably in every language"; Ladd (2006: 14) claims that "instances of these problems are widely attested in the phonology of virtually every well-studied language"; and Scobbie and Stuart-Smith (2006: 15) state that, "in [their] experience ... *every* language has a rump of potential/actual near-phonemes" (emphasis original). Given this widespread phenomenon, it would be useful to have an objective means of categorizing such intermediate relationships. The model proposed in this paper does precisely that.

3 THE MODEL

3.1 OVERVIEW

The foundation of the PPRM, depicted in Figure 1, is to reconceptualize predictability of distribution as a scalar measure instead of a binary one. In Figure 1, each circle represents the distribution of environments that a sound can appear in; the black triangle in each circle represents one realization of a phonological category. At the left-hand end of the continuum, the distributions of two sounds are entirely non-overlapping; a particular environment will occur in the distribution of only one of the two sounds, making it possible to predict which of two sounds will occur in that environment. At this end of the continuum, the sounds are perfectly predictably distributed and hence allophonic. At the other end, the distributions of two sounds are entirely overlapping; any given environment occurs in the distributions of both sounds, making it impossible to predict which of the two sounds will occur in that environment. At this end, the sounds are perfectly contrastive. Crucially, varying degrees of overlap between these two endpoints define different phonological relationships and can be precisely quantified.

More specifically, the notion of predictability of distribution can be recast in terms of the information-theoretic concept of entropy, a measure of the uncertainty that exists in making a selection of one element from a set of choices (where the set of choices is modelled as a random variable). If there is only one element to choose from, or if one element in a set of multiple elements has a probability of being chosen of 1, then there is no uncertainty about which element will be chosen, and the entropy is 0. Starting with a random variable whose outcomes are equiprobable, the entropy of that variable will increase as the number of equiprobable outcomes increases, and will decrease as the number of possible outcomes decreases or as the probabilities of outcomes move away from being equiprobable. When what is of interest is the relationship between two elements, as in two

sounds in a phonological relationship, the system we are dealing with is quite small–there are only two possible outcomes. The entropy of the system, however, will vary with the probabilities of those two elements. There will be the greatest amount of uncertainty if the two sounds are equiprobable in a given environment, and the least amount of uncertainty if one of the two sounds cannot occur in a given environment. The relationship between two sounds, then, is the amount of uncertainty that exists when we have to choose between them.[1]

Thus, the notion of phonological contrast as determined by predictability of distribution can be reformulated in terms of how uncertain a language user is about the choice between two sounds: complete certainty about which of two sounds will occur (based on phonological conditioning factors) is analogous to perfect allophony, where the relationship is entirely predictable, while complete uncertainty is analogous to perfect contrast, where the relationship is entirely unpredictable. Importantly, language users can be partially uncertain of the choice between two sounds: for example, there may be a bias toward one sound rather than another in certain environments. This could happen in the case where, for instance, one of the two sounds is more frequent than the other. For example, in Japanese, both [t] and [d] can occur in the context [#___o], which would make them simply "contrastive" in a traditional phonological analysis. It turns out, however, that when the NTT wordlist of Japanese (Amano and Kondo 1999, 2000) is examined for sequences of [#to] and [#do], [#to] occurs 2/3 of the time, and [#do] only 1/3 of the time. Thus, one could make an educated guess that [t] would occur in the environment [#___o] – one is more certain about which sound occurs in this context than one would be if they were equiprobable, and so the relationship is "less contrastive" than it would be if one simply had a categorical choice between "predictable" and "unpredictable." Language users do in fact seem to be aware of such phonological probabilities and indeed use them in linguistic processing; see, e.g., McQueen and Pitt (1996), Ernestus and Mak (2005), Ernestus (2006).

The PPRM makes use of three different probabilistic measures to create a picture of contrastiveness: bias, environment-specific contrastiveness, and systemic contrastiveness. The basic assumption of the model is that in any given environment *e*, a particular choice can be made between two sounds *a* and *b* –either *a* occurs in *e*, or *b* occurs in *e*. The bias and environment-specific contrastiveness measures focus on this choice within an environment and indicate which of *a* or *b* is more likely in that environment (bias) and how much uncertainty there is in the choice between a and b in that environment (environment-specific contrastiveness). The third measure, systemic contrastiveness, is a measure of how much uncertainty there is in the choice between *a* and *b* across all environments in the language where either *a* or *b* can occur. Each of these will be described in more detail below.

All three measures make use of probabilities. These probabilities can be approximated by using frequency counts from a corpus, keeping in mind the usual limitations on the accuracy of corpora as reflections of languages. A choice must be made, however, between using type frequencies and using token frequencies. Both can give useful insight into phonological patterns, so in the examples in this paper, I will demonstrate the measures using both type and token frequency calculations.

[1] There are of course additional criteria for determining phonological relationships besides predictability of distribution, e.g., lexical distinction. I do not contend that predictability of distribution is the only criterion that should be used in determining phonological relationships; rather, that, insofar as predictability of distribution is a useful measure, it can be better understood as a probabilistic rather than a categorical measure.

Type frequencies, however, are particularly useful if one is interested in the theoretical status of two sounds in a language, while token frequencies tend to reflect the status of the sounds in practice. For example, a pair of sounds that has traditionally been considered allophonic (predictable) may be undergoing a phonemic split in theory because of the occurrence of one or two word types in which they contrast, but still be quite clearly allophonic in practice because those word types are quite rare and have extremely low token frequencies.

3.2 BIAS

The first component of the model is the *bias*, p(a_e) or p(b_e), a measure of which of two sounds is more likely to occur in a given environment. This is calculated using simple probabilities: the bias toward one sound, X = a, is the probability that a will be the value of X in that environment (as opposed to the value of X being b).[2] The formula for bias is in (1).

(1) $$\mathrm{p}(a; a, b|e) = \mathrm{p}(a_e) = \mathrm{N}_{a/e}/(\mathrm{N}_{a/e} + \mathrm{N}_{b/e})$$

This formula indicates that the probability of *a* occurring when the choices are *a* and *b*, given environment *e*, is equal to the number of occurrences of *a* in that environment, divided by the total number of occurrences of *a* or *b* in that environment. For example, if the environment in question were word-initially, and the two sounds in question were [t] and [d], the bias toward [t] would be the number of word-initial [t]s divided by the total number of word-initial [t]s and word-initial [d]s. In addition to capturing the phonotactic distributions in a language in a more realistic manner than simply assuming that two sounds "are" or "are not" predictably distributed, bias can be connected with language users' *expectations* (cf. Hume 2009). That is, it is useful for understanding the biases or expectations of speakers of a language for or against particular sounds in various environments.

3.3 ENVIRONMENT-SPECIFIC CONTRASTIVENESS

The second component of the model is *environment-specific contrastiveness*, H(e), a more direct measure of the relationship between two sounds in a given environment. It is a measure of how much uncertainty (calculated in terms of entropy; cf. Shannon and Weaver 1949) there is between two sounds in that environment. Because there are exactly two choices, entropy ranges between 0 (complete certainty) and 1 (complete uncertainty). The formula for environment-specific contrastiveness is given in (2).

(2) $$\mathrm{H}(e) = -1 * (\mathrm{p}\ (a_e)\ \log_2\ \mathrm{p}(a_e) +\ \mathrm{p}(b_e)\ \log_2\ \mathrm{p}(b_e))$$

This formula indicates that the entropy between two sounds *a* and *b* in environment *e* is equal to the log probability of *a* in that environment, weighted by the probability of *a* in that environment, plus the log probability of *b* in that environment, weighted by the probability of *b* in that environment. This sum is multiplied by negative one simply to make it a positive number.

[2] It should be noted that X here is simply a random variable in the statistical sense; there is no claim that the two sounds *a* and *b* are different realizations of some *phonological* category X.

Notice that the bias toward each sound is integrated into this measure, as the bias is simply the term p(a_e) or p(b_e). On the one hand, it is important to have the measure of contrastiveness in addition to the measure of bias because the contrastiveness measure is truly a single measure of the relationship between two sounds rather than a measure of how likely a single segment is in a given context; on the other, it is important to pull out the measure of bias in addition to the measure of contrastiveness so that the direction of uncertainty can be determined. Because of the unified nature of the measure, I argue that the calculation of entropy is more directly a reflection of the concept of phonological contrast as defined by predictability of distribution (uncertainty as to which of a set of elements occurs) than the calculation of probability by itself; hence, I have termed this aspect of the model a measure of "contrastiveness."

3.4 SYSTEMIC CONTRASTIVENESS

The third measure, *systemic contrastiveness*, (H), is, however, the one that is perhaps of most interest from a traditional phonological perspective. It is the total degree of uncertainty (again calculated in terms of entropy) that exists between two sounds across all the different possible environments that either sound can occur in in a language. It is most analogous to the traditional concept of phonological relationship, which is not environment-specific. It is calculated according to the formula in (3); essentially, it is an average of the environment-specific contrastiveness scores across all environments for a pair of sounds, weighted by the frequency of occurrence of those environments.

(3) $\mathrm{H} = \Sigma(\mathrm{H}(e) * \mathrm{p}(e))$

Thus, environments that do not occur particularly frequently do not count very heavily toward the total average entropy of two sounds in a language. For example, if two sounds are highly unpredictable (there is a high degree of uncertainty as to which occurs) in most environments, but they are neutralized in some infrequent environment(s), then the overall entropy will still be relatively high. Again, calculating the weighted average entropies for each pair provides a more explicit understanding of how much uncertainty there is in the language about the distribution of two segments, as compared to the standard binary distinction between "predictable" and "not predictable"; the systemic measure evaluates this uncertainty across all possible environments in the language rather than treating each environment separately.

4 CANADIAN RAISING REVISITED

Consider how the three components of the PPRM – bias, environment-specific contrastiveness, and systemic contrastiveness – can illuminate the issue of Canadian Raising, described above in Section 2.2. Recall that there is disagreement as to whether the partial unpredictability of the vowels [ɑɪ] and [ʌɪ] before a flap should be sufficient to label the relationship between these two sounds as being contrastive.

To implement the model, the International Corpus of English for Canada (ICE-Can; Newman and Columbus (2010)) was used, in conjunction with the CELEX lexical database of English (Baayen et al. 1995). The ICE-Can corpus includes frequency information for Canadian English,

based on a 1-million word collection of spoken conversations and written texts by Canadians, but is not phonetically transcribed. The CELEX corpus, on the other hand, is phonetically transcribed, though not specific to Canadian English, and so was used to determine a list of possible English words that contain the vowel [ɑɪ], while the ICE-Can corpus was used to determine the type and token frequencies of these words in Canadian English.

There are reports that Canadian Raising has spread beyond the traditional domain of pre- tautosyllabic voiceless segments (e.g., Mielke, Armstrong, and Hume 2003; Hall 2005), an issue that will be returned to below. Given the dispute about which precise words exhibit raising (or non-raising) in unexpected domains, however, the current study simply used a traditional model of Canadian Raising (based on Paradis 1980 and Chambers 1989) to determine just how contrastive [ɑɪ] and [ʌɪ] are, without lexical exceptions. Thus, it was assumed that any instance of /ɑɪ/ that occurred before a tautosyllabic, tautomorphemic voiceless stop would be realized as [ʌɪ], as would any instance of /ɑɪ/ that occurred before a /t/ that might undergo flapping. Ambisyllabic segments (as defined by Kahn 1980) were assumed to be tautosyllabic with the preceding vowel (thus, raising was assumed to occur in words such as *hyper* but not in words such as *hyperbole*).

One question when applying the PPRM to linguistic data is how to count the environments included in the calculations. For example, one could calculate the probability of [ɑɪ] occuring followed by a voiceless segment (i.e., in the environment [___ [-voi]]), or one could calculate the probability of [ɑɪ] occurring followed by a tautosyllabic voiceless segment (i.e., in the environment [[-voi].]) separately from the probability of [ɑɪ] occurring followed by a non- tautosyllabic voiceless segment (i.e., in the environment [___.[-voi]]). While the total number of environments is the same in these two scenarios (the total number of [ɑɪ]s occuring before voiceless segments is the same), the different ways of dividing up the environments do result in different overall calculations of uncertainty.

I believe that these different choices are not a weakness of the model but rather an important insight into how phonology actually works. It is quite possible that different language users in fact make different generalizations about conditioning environments. In fact, I would argue that this is precisely why Canadian Raising has, as mentioned above, appeared in unexpected words– unexpected from the point of view of a traditional analysis, that is. For example, it may be obvious to an analyst that the "correct" (historical) conditioning environment for raising is before tautosyllabic, tautomorphemic, voiceless segments, but a language learner may overgeneralize to, say, all following voiceless segments and produce "anomalies" such as pipette or psychology with a raised [ʌɪ] (cf. Hall 2005).

The important point in this paper, however, is simply to illustrate how the PPRM can be applied to linguistic data to illuminate a question of phonological interest. Thus, for expository purposes, I will present the calculations assuming two different sets of relevant environments. The first set was chosen to emphasize the potential predictability of the two vowels, highlighting the allophonic side of their relationship, while the second set was chosen to emphasize the potential unpredictability of the two vowels, highlighting the contrastive side of their relationship. This is a useful comparison because it will show whether the two sounds can be considered truly allophonic (where allophony is defined as "completely predictable"), or alternatively, just how contrastive the two might be.

It should be noted that, regardless of the specific division of environments, the calculations

Table 1: Type and token frequency calculations, emphasizing the potential predictability of [ɑɪ] and [ʌɪ] in Canadian English

Environment (Before...)	Type Frequency					Token Frequency				
	p([ɑɪ])	p([ʌɪ])	Bias	p(e)	H(e)	p([ɑɪ])	p([ʌɪ])	Bias	p(e)	H(e)
Word Boundary	1.000	0.000	[ɑɪ]	0.069	0.000	1.000	0.000	[ɑɪ]	0.448	0.000
Voiced (Not Flap)	1.000	0.000	[ɑɪ]	0.650	0.000	1.000	0.000	[ɑɪ]	0.313	0.000
Flap	0.406	0.594	[ʌɪ]	0.049	0.974	0.375	0.625	[ʌɪ]	0.010	0.954
Voiceless (Tautosyllabic)	0.000	1.000	[ʌɪ]	0.205	0.000	0.000	1.000	[ʌɪ]	0.226	0.000
Voiceless (Not tautosyllabic)	1.000	0.000	[ɑɪ]	0.028	0.000	1.000	0.000	[ɑɪ]	0.003	0.000
Systemic Contrastiveness	$H = \sum p(e) * H(e)$				0.047	$H = \sum p(e) * H(e)$				0.010

should encompass all environments that either of the two sounds in question can occur in, and that no context should appear in more than one defined environment.

4.1 EMPHASIZING PREDICTABILITY

In order emphasize the predictable nature of the distribution of [ɑɪ] and [ʌɪ], environments need to be chosen such that in as many surface environments as possible, there is no uncertainty about which sound will occur. Given that the instances of [ʌɪ] were superimposed by rule anyway, not actually being transcribed in either corpus, this is a relatively easy task. The environments chosen were: before a word boundary, before a voiced segment other than [ɾ], before a non-tautosyllabic voiceless segment (all three of which contained exclusively [ɑɪ]); before a tautosyllabic voiceless segment (exclusively [ʌɪ]); and before [ɾ] (mixed occurrences of [ɑɪ] and [ʌɪ] depending on the presumed derivation of the [ɾ]). Table 1 summarizes the PPRM calculations for each of these environments; calculations based on type frequencies are on the left and token frequencies on the right.

As can be seen in this table, the overall systemic contrastiveness between [ɑɪ] and [ʌɪ] is quite low (0.05 based on type frequency, 0.01 based on token frequency), though it is not zero. Thus it is quite clear that it is still largely the case that these two vowels are predictably distributed in Canadian English, even considering the presence of the flap as an unpredictable context. Interestingly, however, it is also very clearly the case that "preceding flap" is a context of extremely high uncertainty ($H(e) > 0.95$), thus lending corroboration to the argument that this environment does considerably disrupt the predictability of these two vowels. Its impact on the overall relationship between the two sounds is limited, however, by the relatively low frequency of occurrence of that particular environment relative to other environments ($p(e) < 0.05$). It is far more likely that any given environment will be one in which it is possible to predict which of [ɑɪ] and [ʌɪ] will occur than that it will be one in which it is not possible to make such a prediction; thus, the overall uncertainty is quite low.

Table 2: Type and token frequency calculations, emphasizing the potential unpredictability of [ɑɪ] and [ʌɪ] in Canadian English

Environment (Before...)	Type Frequency					Token Frequency				
	p([ɑɪ])	p([ʌɪ])	Bias	p(e)	H(e)	p([ɑɪ])	p([ʌɪ])	Bias	p(e)	H(e)
Word Boundary	1.000	0.000	[ɑɪ]	0.069	0.000	1.000	0.000	[ɑɪ]	0.448	0.000
Voiced	0.959	0.041	[ɑɪ]	0.699	0.249	0.980	0.020	[ɑɪ]	0.323	0.141
Voiceless	0.119	0.881	[ʌɪ]	0.232	0.525	0.011	0.989	[ʌɪ]	0.229	0.089
Systemic Contrastiveness	H = ∑ p(e) * H(e)				0.296	H = ∑ p(e) * H(e)				0.066

4.2 EMPHASIZING CONTRASTIVENESS

At the other extreme, we can also apply the same PPRM calculations using environments that tend to minimize the predictability of distribution of [ɑɪ] and [ʌɪ] by choosing environments so that both segments can occur in those environments.[3] Specifically, collapsing all the voiceless environments or all the voiced environments from Section 4.1 leads to the divisions and calculations seen in Table 2.

It has been claimed that the influence of the flapping environment has disrupted the predictability of Canadian Raising such that the two vowels should be analyzed as being contrastive (e.g., Mielke, Armstrong, and Hume 2003; Kaye 2009). The above calculations show that while this is true from the perspective of assuming that any degree of unpredictability warrants an analysis of contrast, even when environments that tend to maximize contrastivity are chosen, the two vowels are not particularly unpredictably distributed (though cf. footnote 3). The flapping environments are such a small subset of the voiced contexts, and the non-tautosyllabic voiceless environments such a small subset of the voiceless contexts, that the measures of contrastiveness in these environments are simply fairly low, even though both [ɑɪ] and [ʌɪ] can occur in these contexts. Thus, while it may be the case that an analysis in which [ɑɪ] and [ʌɪ] are simply assumed to be allophonic is untenable, it is also not the case that the contrast created by their unpredictability in flapping environments has a particularly strong footing.

Nonetheless, the intermediacy of the relationship between [ɑɪ] and [ʌɪ] caused by flapping environments may open the door to further changes. As noted above, language users may make different generalizations about the environments in which [ɑɪ] and [ʌɪ] can occur, and those generalizations might lead to new distributions of the vowels, as has been documented (e.g., Hall 2005). In short, the PPRM predicts allophonic splits arising because sounds may be predictably distributed in some, but not all, of their environments, leading language users to be uncertain as to the correct generalizations to make about their distributions. This uncertainty can result in variability in the generalizations that

[3] Of course, to actually maximize unpredictability (contrastiveness), one could assume that there is no conditioning at all based on phonetic environment and simply calculate the overall probability and entropy of the choice between of [ɑɪ] and [ʌɪ] in the corpus. Doing so reveals that [ɑɪ] occurs about 78% of the time (with either type or token frequencies) and that the entropy between [ɑɪ] and [ʌɪ] is approximately 0.76. Given the fact that the choice between [ɑɪ] and [ʌɪ] is in fact at least somewhat productively predictable, however (e.g., Bermúdez-Otero (2003); Boersma and Pater (2007); Idsardi (2006)), this complete disregard for environmental conditioning seems unwarranted.

are made, and the variability among generalization can lead to change. The model thus allows us not only to calculate the precise degree of predictability of distribution of two sounds at a particular point in time but also to document the change in progress.

One particularly interesting observation in the Canadian Raising case is that the token- frequency calculations in both Table 1 and Table 2 indicate that the two vowels are less contrastive than do the type-frequency calculations. This indicates that the extent to which [ɑɪ] and [ʌɪ] contrast with one another is further advanced in theory than in practice, in that there are apparently a number of words that lead to a contrast between the two vowels that are not particularly frequently used. Thus, by comparing type and token based calculations, this model provides insight into just how much of an effect individual words have on the phonological relationship between two sounds.

5 CONCLUSION

This paper has proposed a model of phonological relationships, the Probabilistic Phonological Relationship Model, that quantifies how predictably distributed two sounds in a relationship are. The PPRM starts with one of the long-standing tools for determining phonological relationships, the notion of predictability of distribution. Building on insights from probability and information theory, the model provides a way of calculating the precise degree to which two sounds are predictably distributed. It includes a measure of the probability of each member of a pair in each environment the pair occurs in and two measures of contrastiveness: the uncertainty (entropy) of the choice between the members of the pair in a given environment and the overall uncertainty of choice between the members of the pair in a language. These numbers provide a way to formally describe and compare relationships, such as that found in Canadian Raising, that have heretofore been treated as exceptions, ignored, relegated to alternative grammars, or otherwise seen as problematic for traditional descriptions of phonology because of their intermediate status as neither "predictable" nor "unpredictable."

REFERENCES

Allen, H. (1989). Canadian raising in the upper midwest. *American Speech*, 64(1):74–75.

Amano, S. and Kono, T. (1999). *The properties of the Japanese lexicon*. Sanseido Co. Ltd.

Ash, S. (2003). A national survey of north american dialects. *Publication of the American Dialect Society*, 88(1):57–73.

Baayen, R., Piepenbrock, R., and Gulikers, L. (1995). The celex lexical database. *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*.

Bermúdez-Otero, R. (2003). The acquisition of phonological opacity. In *Variation within Optimality Theory: Proceedings of the Stockholm Workshop onVariation within Optimality Theory*, pages 25–36.

Bloomfield, L. (1939). Menomini morphophonemics. *Travaux du cercle linguistique de Prague*, 8(15):351–62.

Blust, R. (2003). *Thao dictionary*, volume 5. Institute of Linguistics (Preparatory Office), Academic Sinica.

Boersma, P. and Pater, J. (2007). Constructing constraints from language data: The case of canadian english diphthongs. *NELS*, 38.

Britain, D. (1997). Dialect contact and phonological reallocation:"canadian raising" in the english fens. *Language in Society*, 26(01):15–46.

Bye, P. (2009). Three types of marginal contrast. *Paper presented at the Toronto-Tromso Phonology Workshop*.

Chambers, J. (1973). Canadian raising. *Canadian Journal of Linguistics*, 18(2):113–135.

Chambers, J. (1989). Canadian raising: Blocking, fronting, etc. *American Speech*, 64(1):75–88.

Collins, B. and Mees, I. (1991). English through welsh ears: The 1857 pronunciation dictionary of robert ioan prys. In *Language usage and description: Studies presented to NE Osselton on the occasion of his retirement*, pages 47–58.

Crowley, T. (1998). The voiceless fricatives [s] and [h] in erromangan: One phoneme, two, or one and a bit? *Australian Journal of Linguistics*, 18(2):149–168.

Dinnsen, D., Chin, S., Elbert, M., and Powell, T. (1990). Some constraints on functionally disordered phonologies: Phonetic inventories and phonotactics. *Journal of Speech and Hearing Research*, 33(1):28.

Ernestus, M. (2006). Statistically gradient generalizations for contrastive phonological features. *Linguistic review*, 23(3):217.

Ernestus, M. and Mak, W. (2005). Analogical effects in reading dutch verb forms. *Memory & cognition*, 33(7):1160–1173.

Fruehwald, J. (2007). The spread of raising: Opacity, lexicalization, and diffusion. *CUREJ-College Undergraduate Research Electronic Journal*, page 73.

Gleason, H. (1961). Review of "african language studies i" (ed. malcolm guthrie) and "the role of tone in the structure of sukuma" (by i. richardson). *Language*, 37(2):294–308.

Goldsmith, J. (1995). Phonological theory. In Goldsmith, J., editor, *The Handbook of Phonological Theory*. Blackwell.

Hall, K. (2005). Defining phonological rules over lexical neighbourhoods: Evidence from canadian raising. In *Proceedings of the 24 th West Coast Conference on Formal Linguistics*, pages 191–199.

Hall, K. C. (submitted). A typology of intermediate phonological relationships. *Linguistic review*.

Hualde, J. (2004). Quasi-phonemic contrasts in spanish. In *WCCFL 23: Proceedings of the 23rd West Coast Conference on Formal Linguistics*, pages 374–98.

Hume, E. (2009). Certainty and expectation in phonologization and language change.

Hume, E. and Johnson, K. (2003). The impact of partial phonological contrast on speech perception. In *Proceedings of the XVth International Congress of Phonetic Sciences*.

Idsardi, W. (2006). Canadian raising, opacity and rephonemicization. *Canadian Journal of Linguistics*, 51(2-3):119–126.

Joos, M. (1942). A phonological dilemma in canadian english. *Language*, pages 141–144.

Kahn, D. (1980). Syllable-based generalizations in english phonology.

Kaye, J. (2009). Canadian raising, eh? Unpublished manuscript.

Kiparsky, P. (2003). Analogy as optimization:exceptions' to sievers' law in gothic. *Analogy, Levelling, Markedness: Principles of Change, Phonology and Morphology*, pages 15–46.

Labov, W. (1994). *Principles of linguistic change*, volume 1. Wiley-Blackwell.

Ladd, D. (2006). 'distinctive phones' in surface representation. *Laboratory Phonology*, 8:3–26.

McMahon, A. (2000). *Lexical phonology and the history of English*. Cambridge University Press.

McQueen, J. and Pitt, M. (1996). Transitional probability and phoneme monitoring. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2502–2505. IEEE.

Mielke, J., Armstrong, M., and Hume, E. (2003). Looking through opacity. *Theoretical linguistics*, 29(1-2):123–139.

Moren, B. (2004). The phonetics and phonology of front vowels in staten island english: When the traditional descriptions and the facts do not agree. In *9th Conference on Laboratory Phonology, University of Illinois, Urbana-Champaign*.

Newman, J. and Columbus, G. (2010). The ice canada corpus, version 1. Retrieved from http://ice-corpora.net/ice/download.htm.

Paradis, C. (1980). La règle de canadian raising et l'analyse en structure syllabique. *Canadian (The) Journal of Linguistics= La Revue Canadienne de Linguistique Toronto*, 25(1):35–45.

Rose, S. and King, L. (2007). Speech error elicitation and co-occurrence restrictions in two ethiopian semitic languages. *Language and Speech*, 50(4):451–504.

Scharf, P. and Hyman, M. (2011). *Linguistic Issues in Encoding Sanskrit*. The Sanskrit Library.

Scobbie, J. and Stuart-Smith, J. (2006). Quasi-phonemic contrast and the fuzzy inventory: Examples from scottish english. *QMU Speech Science Research Centre Working Papers*.

Scobbie, J. and Stuart-Smith, J. (2008). Quasi-phonemic contrast and the indeterminancy of the segmental inventory: Examples from scottish english. *Contrast in phonology: Perception and acquisition*.

Steriade, D. (2007). Contrast. *The Cambridge handbook of phonology*, pages 139–157.

Trentman, E. (2004). *Dialect death in Calvert County, Maryland*. NWAV, Detroit, MI.

Trudgill, P. (1985). New dialect formation and the analysis of colonial dialects: the case of canadian raising. In *Papers from the 5th International Conference on Methods in Dialectology. Victoria: University of Victoria*, pages 35–45.

Vance, T. (1987). " canadian raising" in some dialects of the northern united states. *American Speech*, 62(3):195–210.

Zuraw, K. (2002). Aggressive reduplication. *Phonology*, 19(3):395–439.