Probabilistic Phonology: Discrimination and Robustness

Janet B. Pierrehumbert
Northwestern University
Evanston, IL

January 20, 2002

1 Introduction

Phonology deals with the implicit knowledge of language sound structure as used contrastively to convey meaning. The ability of humans to use this knowledge productively shows the need for an abstract generative theory of phonology. A fluent adult speaker of a language can produce new words with native word-level allophony, and produce novel combinations of words with native phrase-level allophony. S/he can accommodate borrowings to native phonotactics, and create morphological neologisms involving phonological modifications of the component parts. Baayen (2001) presents calculations suggesting that new words are continually created. No matter how large of a corpus one has, a substantial increase in its size will uncover additional words. Thus phonology is productive for the same reason that syntax is: to express novel meanings, people construct new combinations of familar parts.

The productivity of phonology is widely believed to produce evidence for a theory in which the phonology of any particular language has the character of a formal grammar, and the universal theory of phonology delineates the types of grammars which are available for use in individual languages. For example, the phonology of Finnish includes a set of terminal elements, such as features or phones. It includes principles for combining these elements into well-formed syllables, metrical feet, and phonological words. These principles permit /h/ in coda position (as in the word /kahvi/, *coffee*. They permit the double attachment of a phoneme to coda and onset position (as in the word /hɛlp:o/, *easy*. They preclude the kind of complex onsets we see in English /strit/. They set up an alternating word pattern, described with metrical feet. They enforce initial word stress absolutely. Analyses of this sort are familiar to any reader. The assumption about such analyses that I want to emphasize is that they involve synoptic knowledge which exploits abstract variables, such as C (consonant), μ (mora) and σ (syllable). If the knowledge were less abstract and synoptic, it would not generalize to novel cases in the way it does.

This conception of phonology as a formal grammar (with abstract variables) is often assumed to stand in opposition to the idea that phonology involves statistical knowledge. However, this opposition is a spurious one, because probability theory requires us to assign probability distributions to variables. Without any variables, there would be no way for a statistical learning model to tabulate any statistics about anything. Once we have variables, they can be as abstract as we like;

in principle, we can assign probability distributions to any variable of any nature that we may care to define in response to scientific findings.

Introducing probability distributions into the model provides obvious and well-established tools for handling the variable data and gradient outcomes which I will discuss below. However, it in no way precludes a rigorous treatment of any phenomena that prove to be highly categorical. Such phenomena can be handled in probabilistic models by assigning extreme probabilities to particular outcomes, namely a probability of 0 to a nonoccurring event and a probability of 1 to an outcome which is certain. In short, within a probabilistic model, nonprobabilistic fragments of the grammar are readily treated as a limiting cases. The reader is referred to Carnap (1950) and Adams (1998) for complete formal development of the intrinsic connection between probability theory and logic. In the remainder of this paper I will assume that the ultimate and true theory of phonology will involve both probability distributions and abstract variables, because the abstract variables have probability distributions over other levels of representation. My discussion will focus on the questions of which distributions over which variables. I will argue that careful consideration of how statistical distributions are established and distinguished from each other enables us to reach important conclusions about the nature of human language, conclusions which would elude us in a nonprobabilistic framework.

In viable theories of phonetics/phonology, there is a ladder of abstraction, with each level having its own representational apparatus. Thus, the theory as a whole must delineate both the available representation at each level and the principles relating one level to another. In the remainder of this paper, it will be important to distinguish the following levels. This list represents a minimal rather than definitive list of levels of representation, and representational distinctions within each level on the list have been glossed over when they are not central to the goals of the paper.

1) **Parametric phonetics**.

The parametric phonetic representation is a quantitative map of the acoustic and articulatory space. In speech perception, it describes the perceptual encoding of the speech signal on each individual occasion. In speech production, it describes the articulatory gestures as they unfold in time and space.

2) **Phonetic encoding**.

The phonetic encoding system of a language abstracts over the parametric phonetic space, defining the inventory available in the language for encoding word forms (phonological representations of words). In traditional phonology, these categories are taken to be phonemes. Phonemes are minimal contrastive units and they can be equated across different structure positions; the same phoneme occurs at the beginning of *pat* and the end of *step*. However, much recent evidence (reviewed below) suggests that phonetic categories are considerably less abstract than phonemes. They are not minimal in the sense that they include redundant information. Nor are they equatable across context. Thus, they can be viewed as peaks in the total phonetic distribution of the language (e.g. areas of the parametric space which the language exploits preferentially), or as positional allophones. Phonetic encoding, as used here, includes not only speech segments but also aspects of prosody and intonation which are defined with respect to the speech signal.

3) **Word-forms in the lexicon**.

Each word in the speaker's lexicon has a representation of its sound structure which allows that word to be recognized despite variation in its phonetic form due to speaker differences and to context. The same representation presumably mediates between perception and production, making

it possible for a speaker to repeat words which s/he has acquired through perception. Given this description, it is clear that word-forms are also abstractions over the phonetic space. A given word is learned through repeated exposure to that word in speech. For any stretch of speech and any given word, the speech either is or is not an example of the word. Pervasive word frequency effects in psycholinguistics show that a word's frequency of occurrence impacts the long-term representation of that word. Connections between word frequency and detailed quantitative aspects of pronunciation are documented in Bybee (2001) and Jurafsky et al. (in press). Such phenomena strengthen the point that words are generalizations over speech.

4) **The phonological grammar**.

The phonological grammar, encompassing both prosodic structure and phonotactics, describes the set of possible words of a language. (Phrasal phonology is beyond the scope of the present paper) The grammar is revealed by well-formedness judgements as well as neologisms and borrowings. It is also revealed by the procrustean adjustments which morphologically complex forms can undergo if they become lexicalized as units.

Phonology represents generalizations over the word-forms in the lexicon, which are in turn generalizations over speech. Hence, phonology does not abstract over speech directly, but rather indirectly via the abstraction of word-forms. This jump in level has important consequences. First, we can find discrepancies between the strongest patterns in the lexicon and the strongest patterns in running speech. Discrepancies occur because a pattern which is common in running speech may be rare in the lexicon if its occurrence in running speech results from just a few different words. Secondly, the size of the lexicon is small in comparison to the total number of word tokens a person encounters. An adult speaker knows on the order of 10,000 monomorphemic words and 100,000 words total (see discussion below). Any generalization in the word-level phonology must be learnable from a data set of this size. In contrast, the 200 million word corpus used in Baayen's calculation (2001, as cited above) would correspond to about 18,000 hours of speech, about the amount of speech a native speaker has encountered by the time s/he reaches adulthood, assuming s/he hears speech two to three hours per day. The language learner has at his or her disposal a truly vast amount of data for learning generalizations over speech, such as native details of pronunciations. The consequences of this vast discrepancy will be discussed further below in connection with statistical robustness.

5) **Morphophonological correspondences**.

A given stem or affix can assume phonologically different forms in different, related words. Many of the differences are attributable to general constraints of the phonological grammar. For example, in a standard analysis the /n/ pronounced in the word *hymnal* (before a vowel-initial suffix) is not realized in the bare form of the stem *hymn*. This fact may be attributed to sequencing constraints within the syllable. The vowel-initial suffix rescues the consonant by parsing it into onset position. A general line of development in phonological theory, launched by Kisseberth's (1970) work on conspiracies, is to show how morphophonological alternations arise when contextual factors make a given stem subject to different surface constraints. This enterprise, insofar as it succeeds, minimizes the need for explicit statements about alternations. Optimality Theory is a recent manifestation of this trend. However, no current theory explains all morphophonological alternations on the basis of general phonological patterns, because many such alternations are not, in fact, general. Many alternations need to be learned on the basis of specific relations between words or within paradigms. I will refer to such alternations as correspondences.

Morphophonological correspondences will not be the primary topic of this paper, since they are treated independently in the contribution by Baayen. The cases that Baayen analyzes in depth are linking elements in Dutch nominal compounds and final voicing/devoicing in the past tense of Dutch verbs. Parallel cases in English include the irregular voicing in the English plural (as in *thief, thieves*, and (perhaps surprisingly), English flapping, as in the contrast between *capitalistic*, which shares the flap of *capital*, with *militaristic*, which maintains the aspiration of *military* (Withgott, 1983). Such patterns involve generalizations over pairs or sets of words. This point is brought home by an effort to estimate the plural of an unstable form. Is the plural of *roof rooves*, on the analogy *thief: thieves :: roof : rooves*? Or is it *roofs*, on the analogy *cuff: cuffs:: roof: roofs*? The AML model (Analogical Modelling of Language) developed in Skousen (1989) acknowledges this fact in its very name; this model is discussed in detail in Baayen's contribution to this volume. In Optimality Theory, Output-Output Correspondance constraints, or Sympathy constraints, acknowledge the same kind of relationships (see McCarthy and Prince, 1995 and McCarthy 1999). Morphophonological schemas in the usage-based theory of Bybee (2001) also involve generalizations over word relationships.

The levels of representation just introduced imply an organization of probabilities in the theory. Specifically, phonetic categories have probability distributions over the parametric phonetic space. Word-forms, viewed as sequences of phonetic categories, also have probability distributions over temporal sequences of events in the phonetic space. (These distributions may be deducible from the categories which comprise the word-form. Or nearly so.). The prosodic and phonotactic templates which define the phonological grammar have probability distributions over the word-forms in the lexicon. Morphophonological relationships also involve probability distributions over a universe of pairings or collections of word-forms.

In the following sections, I will first review empirical evidence that implicit knowledge at all levels of representation is probabilistic. Then, I will show how probabilistic reasoning predicts limits on the inferences that can be made about language, either by the scientist or by the language learner. These limits lead to level-specific effects, because the phonetic encoding system and word-forms are acquired from vast exposure to speech, whereas the phonological grammar is abstracted over the lexicon. The lexicon provides a much smaller number of data points than running speech. Language is statistically robust, and this robustness forms its character at all levels. Lastly, I will discusses correlations across levels and their implications for phonological theory.

2 Probability at various levels of representation

2.1 The phonetic space and categories over it

The phonetic inventory of a language is a set of labelled probability distributions over the phonetic space. By the phonetic space, I mean the acoustic and articulatory parameterization of speech as a physical event. For example, to a first approximation, vowels can be viewed as probability distributions over F1 - F2 space, as shown in Figure 1. Figure 1 (redrawn from the data of the famous Peterson and Barney, 1952) pools data for the same vowel across speakers. Each vowel occupies a continuous region of the space. The regions for different vowels are quite distinct, even ignoring F3 and any other characteristics which distinguish them. Each vowel is more frequently instantiated by values near the center of its distribution than by values near the edges of its distribution.

PUT FIGURE 1 ABOUT HERE

The F1-F2 space is continuous, because the parameters vary continuously. It is possible to define a vowel location which is *in between* any two particular vowel tokens, no matter how close these two may be to each other, and it is possible to define a distance metric between vowels which integrates the separation in the F1 and F2 directions. The same claims would apply to other physical parameters as well, such as spectral tilt, jaw opening, or activation of the cricothyroid muscle. If we conceive of the phonetic space in terms of all controllable or noticeable articulatory and acoustic dimensions, then it is a very high dimensional space indeed. A complex shape within this hyperspace describes the combinations of articulatory and acoustic values which human anatomy and physiology permit us to achieve.

Any individual language exploits as categories a reasonably small number of regions in hyperspace, with no language using all regions of the phonetic space equally. In fact, a rather sparse selection of regions is used to any individual language, as compared to the universal capabilities evidenced by the union of phonetic outcomes across all languages. The maximal phoneme inventories reported in Ladefoged and Maddieson's (1996) typological survey are still small compared to the complete IPA, viewed as an estimate of the total set of sound segments available for use in any language.

The claim that languages use regions of the phonetic space – as opposed to points in the space – is evidenced by the fact the phonetic realization of any given element is *always* variable. Even repeated recordings of the same speaker saying the same word in the same context will yield some variability in the measured values of physical parameters. When we consider the amount of variation relating to differences in vocal tract anatomy, speech style, and idiolect that the speech perception system copes with, it is clear that that the system has an impressive capability for coping with variation. Indeed, the whole point of abstract levels is to cope with variation. If productions of the same word were acoustically identical across speakers and situations, then the recognition system could work just by matching spectral templates, the mathematical equivalent of laying two spectrograms on top of each other and holding the pair up to the light to see whether they were identical or not. Lexical entries could link spectral templates directly to meanings, and no more abstract phonetic encoding would be necessary.

Acquiring the phonetic encoding system of a language involves acquiring probability distributions over the phonetic space. Evidence that these distributions are acquired comes both from language typology and from studies of language acquisition. As discussed at more length in Pierrehumbert (2000) and Pierrehumbert, Beckman, and Ladd (2001) there is no known case of a phoneme which has exactly the same phonetics in two different languages. Even the most closely analogous phonemes prove to be systematically different when examined quantitatively in analogous contexts, and patterns of variation across context reveal even more quantitative differences. For example Caramazza and Yeni-Komshian, (1974) show that voice onset times for stops in Canadian French differ systematically from those in both American English and Continental French. Experiments reported in Flege and Hillenbrand (1986) show that vowel lengthening before a voiced fricative is quantitatively different in English and French, and that English and French listeners are attuned to this difference as a perceptual cue for the voicing of the fricative. Beddor and Krakow (1999) explore language specific details of nasal coarticulation, and Beddor et al. (in press) present similar results for patterns of vowel-vowel coarticulation in different languages.

Studies of language acquisition show that phonetic knowledge is acquired gradually. Although

children show attunement to the encoding system of their language towards the end of the first year of life (Werker and Tees, 1994), it takes a long time to reach adult levels of competence. In production, elementary school children still lack adult levels of accuracy in durational and spectral patterns (Lee, Potamianos, Narayan,1999) and in control and perception of coarticulatory patterns (Nittrouer, 1992, 1993). Hazen and Barrett (2000) present similar results for phoneme perception, showing that categorization boundaries for minimal pairs such as *coat, goat* sharpen gradually from ages 6 to 12, but at age 12 they still differ from those of adults. Such findings are readily modeled by assuming that the probability distribution corresponding to any encoding unit is incrementally updated through experience. In contrast, the (still common) assumption that phonological acquisition involves selecting items from a universally available categorical inventory (along the lines of the IPA) provides no way of representing the fine phonetic differences which define a native accent. It also fails to explain why children take so very long between positing a category in their language and using it with adult levels of precision in perception and production.

As discussed in Johnson (1997b) and Pierrehumbert (2001a), the perceptual learning involved in the gradual acquisition of detailed phonetic categories is readily modelled using exemplar theory. In exemplar theory, labels are associated with a distribution of memory traces in a parametric space, in this case a cognitive representation of the parametric phonetic space. These traces are the exemplars which give the model its name. Empirical distributions of exemplars associated with each label are gradually built up as speech tokens are encountered and encoded. This concept is illustrated in Figure 2 (repeated from Pierrehumbert 2001a), in which a single dimension, F2 (the second formant value) is selected for the purposes of exposition. The two labels involved are /ɪ/ and /ɛ/, which are generally but not completely separated in F2. Vertical lines represent remembered instances of /ɪ/ and /ɛ/, with higher vertical lines representing the strong memory trace resulting from a pile-up of recent examples. The empirical distributions are not individually normalized, and so /ɪ/, being more frequent than /ɛ/, is more abundantly represented. An incoming token of an unknown vowel is indicated by the asterisk, located at a point on the F2 axis which reflects its actual F2 value. According to a standard model of perceptual classification, the labelling of this token is determined by a statistical choice rule which assigns the most probable labeling based on the location of the unknown stimulus in relation to the density of the competing labels in the neighborhood of the stimulus. (See R. D. Luce et al., 1963a and Krushke, 1992). The relevant neighborhood is indicated by the arrows around the stimulus. Since the distribution for /ɪ/ shows a higher density in this neighborhood, the stimulus is classified as /ɪ/. This classified speech event would induce an additional labelled memory trace at that location. Through continual updating in this fashion, the mental representation of the distribution for each label is gradually built up with experience.

PUT FIGURE 2 ABOUT HERE

For details of production to mirror perception, it is necessary to run the model backwards. Pierrehumbert (2001a, in press) provides an exact proposal for doing so. Once a label is selected (presumably, through a decision to say a word that involves that label), a production goal is established by sampling of the empirical distribution for that label. Specifically, an averaged neighborhood around a randomly sampled point in the distribution serves as a production goal. Bias terms on the initial sampling and on the computed production goal are available to model systematic shifts in the use of the phonetic space. Such shifts can come about through choices of speech style and historical changes in progress.

Johnson (1997b) and Pierrehumbert (2001a) leave open the question of what causes a phonetic category to be initiated in the first place. Categorization could be tied to functionality at another level (in particular, to functionality in meaning contrasts). However, increasing evidence that infants initiate categories of phonetic encoding before they know that words have meaning has encouraged researchers to look for bottom-up factors in categorization. A series of experiments reported in Maye and Gerken (2000) and Maye et al. (in press) contrasts encoding of the same phonetic continuum under two conditions, one in which the distribution of tokens over the continuum is bimodal and one in which it is unimodal. They found both adults and infants interpret the bimodal continuum as involving two categories, even in the absence of any information about meaning. These experimental results are substantially anticipated by calculations presented in Kornai (1998). Kornai carried out an unsupervised cluster analysis on vowel formant data from Peterson and Barney (1952). The clusters identified through this analysis are extremely close to the mean values for the ten vowels of Peterson and Barney. The success of this analysis reflects the fact that utilization of the vowel space is not uniform, as discussed above. Utilization is more intense in regions close to the prototypical values for vowels of the language, and it is sparse in between. Results of this character are not confined to speech segments. Mermelstein (1975) launched a line of research on bottom-up detection of syllables by obtaining strong results on the basis of the amplitude integral of the speech signal. Pausing, lengthening, and voice quality changes may effectively cue major intonational boundaries. Thus, it appears to be a hallmark of natural language that units of phonetic encoding units correspond to distributional peaks over the parametric phonetic space. I will return to the signifance of this finding in section 4 below.

2.2 Word-forms and lexical neighbors

People learn new words by encoding instances of those words encountered in speech. Recent results reviewed in Pierrehumbert (in press) indicate that these long term representations of individual words are surprisingly detailed. Evidence for long-term encoding of subphonemic detail includes morphological transfer effects, as discussed in Steriade (2000), historical entrenchment of sociostylistic features, as discussed in Yaeger-Dror and Kemp (1992) and Yaeger-Dror (1996), and unconscious imitation of specific voices, as revealed in Goldinger's (2000) experiments. It may also be noted that infants acquire their first words before they abstract phonemes, and that their phoneme inventory is only gradually developed on the basis of recurring patterns in a growing lexicon (see review in Vihman, 1996). Results such as these dictated the present treatment of word-forms as generalizations over the speech stream which use the resources of the phonetic encoding system.

Results from the psycholinguistic literature show that our total store of words is not an unstructured list. Instead, the words are organized in a lexical network in which words are connected to each other by virtue of phonological and semantic relationships. During language acquisition, the similarities and contrasts amongst words in the early vocabulary play an important role in the subsequent evolution of the system; see Beckman and Edwards (2000) for an incisive review. In adults, each word's competitiveness in perception and production is strongly affected by the number and frequency of word-forms which are similar to it. Luce et al. (1990, 1998) explored the consequences of such similarities for CVC syllables, using the term "lexical neighbor" for a CVC which differs minimally from a given CVC through the addition, deletion, or substitution of a single phoneme. They found that other things being equal, it takes longer to recognize words which have

many lexical neighbors than words which have few lexical neighbors. The delay is particularly great if the lexical neighbors have high average frequency relative to the target; the hardest words to recognize are infrequent words with many frequent neighbors, and the easiest are frequent words with few and infrequent lexical neighbors. However, it is not as widely acknowledged that these same factors affect both and allophonic outcomes and well-formedness judgments, phenomena which are traditionally viewed as the purview of phonology.

Wright (1997) measured formant values in a set of "hard" and "easy" words produced in citation form. He found that such "hard" words were more fully articulated in citation form than the "easy" words, as evidenced by a somewhat expanded vowel space. Already in (1964), Greenberg and Jenkins showed that words which are highly similar to existing words receive high ratings as possible words of a language. Consider a new word *canvle* . It is minimally different from several existing words, as shown in Table 1.

PUT TABLE 1 ABOUT HERE

The existence of these real words means that this *canvle* is a very good nonsense word. Bailey and Hahn (2001) undertook detailed experimental and computational study of the effects of phonotactics and lexical neighbors on wordlikeness judgments of nonsense words. They found that both of these factors have identifiable impacts, and together they explain wordlikeness judgments better than either one does alone. Notably, words with improbable phonotactics were rated better if they had more lexical neighbors than if they had fewer.

Both lexical neighborhoods and phonotactics reflect general knowledge of the lexicon. Conceptually, both involve searching the lexicon for words which match some phonological template. However, there are important differences between these two concepts, which relate to differences in the level of abstraction involved. The lexical neighborhood is found by identifying the few words in the lexicon which nearly match a given word in its entirety. The lexical neighbors of a word can mismatch it in various, unrelated, regards. In the example above, *canvle* mismatches *anvil* in initial position, whereas it mismatches *candle* in the onset of the second syllable. Phonotactic generalizations are made over all words in the lexicon, but involve far less detailed templates. For example, the fact that /nt/ is a phonotactically splendid nasal-obstruent cluster relates to the huge, diverse set of words containing /nt/ as either a heterosyllabic or a final cluster, including words such as *intrepid, pantry, encounter, poignant, tyrant* There are 650 such examples in the Celex monomorphemes. (The large on-line dictionary Celex, as described in Baayen et al. 1995, was the basis for the inventory of monomorphemic words used in Hay et al. in press and Pierrehumbert, 2001b). For /nf/, which is phonotactically words than /nt/, there are only 38 examples in the Celex monomorphemes. Examples include *influence* and *enforce*. For any of the words just cited, the nasal-obstruent template covers only a small part of the word. Lexical neighborhoods are understood to arise on-line during processing as words which are highly similar to the current word token are activated during perception or production. Thus, there is no evidence that they involve any more abstraction than the abstraction involved in encoding and storing the word-forms themselves. Phonotactic generalizations, in contrast, may plausibly involve abstraction of a phonological grammar based on knowledge of all the words. Thus, the Bailey and Hahn results may be understood as showing that wordlikeness judgments do not tap a single level of representation in the system. Instead, the decision about whether a neologism is a possible word of English is reached by combining information from two levels of representation. The input from the lexical level is the extent to which the target word activates highly similar existing words. The input from

the phonological level is the well-formedness of the word as determined by a stochastic phonological parse integrating its subparts.

Distinguishing lexical neighborhood from phonological effects is difficult, requiring the kind of exacting calculations and experiments that Bailey and Hahn have performed, because there is a high correlation between the size of a word's lexical neighborhood and its phonotactic likelihood as computed from its subparts. If a word has numerous neighbors, these neighbors will often have numerous shared subparts which all contribute to the counts for the subparts in a (stochastic) phonological grammar. However, the correlation is not perfect. A phonotactically probable word may have few neighbors if its various subparts are instantiated in nonoverlapping sets of other words. Even though each subpart would be found frequently in this case, no substantial fragment of the word would match a substantial fragment of any other word. Similarly, an improbable word may have fairly many neighbors if its rare subparts happen to cooccur in several other words.

Because of the difficulty of making this distinction in concrete cases, the existence of a phonological grammar as an explicit level of abstraction is in fact still controversial in the psycholinguistic literature. Bailey and Hahn (2001) provides one example of a study which provides evidence for a phonologically abstract level. Experiments reported in Vitevich and Luce (1998) and Vitevich et al. (1999) show evidence separating lexical and phonotactic effects. Though phonotactic goodness and lexical neighborhood density are correlated, they find that they have opposite effects in perception. Phonotactic goodness facilitates perception whereas dense lexical neighborhoods delay perception by creating competition. Furthermore, the role of the two levels is task-dependent. The influence of the lexicon is reduced in tasks which are carried out at high speed and for which only sound pattern information is functionally relevant. Such an outcome can only be modelled if the lexicon is distinguished from the phonotactics. Data on the psychological reality of OCP effects discussed in Berent and Shimron (1997) and Frisch and Zawaydeh (2001) also provide evidence for phonological abstraction of this particular constraint. In the light of such results, I will assume, following mainstream thought in linguistics, that an abstract phonological level is to be distinguished from the lexicon proper. I now turn to the properties of this level.

2.3 Words and word-level phonology

Words are made up of phonological units, and the way in which units combine differs across languages. Phonotactics deals with how these units combine to make words. Because the lexicon is an open set, as discussed in the introduction, the goal of phonological description is not the actual word list, but rather the larger set of possible words. A number of sources of evidence converge to indicate there is a scale of possibility, which relates to the perceived likelihood of the whole as a function of the frequency of the subparts and the specific way in which they were combined. All of the studies I will cite exemplify probabilistic knowledge of phonological constraints; however, not all authors eliminate lexical neighborhood effects as a possible artifactual source of their findings.

In an experimental study of nonsense CVCs, Treiman et al. (2000) constructed stimuli in which phoneme frequencies were controlled and the frequency of the rhyme (the VC) was varied. They found that both well-formedness judgments and outcomes in a blending task reflected the rhyme frequency. Frisch et al. (2000) collected data on the word-likeness of two to four syllable nonsense words comprised of either high frequency or low frequency CV syllables. The ratings were a cumulative function of the frequencies of subparts. Hay et al. (in press) investigated the

perception and judged acceptability of trochaic nonsense words containing medial nasal-obstruent clusters. They found that transcription errors disproportionately corrected infrequent clusters to (acoustically similar) more frequent ones. Acceptability was a gradient function of the likelihood of the (statistically) best morphosyntactic parse of the words. Nasal-obstruent constraints and an OCP constraint on strident fricatives both impacted acceptability, and these effects interacted cumulatively to determine the score of a form. Munson (2001) found that phonotactic likelihood affects judgments by both adults and elementary school children. It affects error rates in production by children but not by adults. Zamuner et al. (2001) replicates Munson's finding for adults and extends the production results to infants.

In the literature on the psychological reality of phonotactic regularities, effects by now have been found for a wide variety of phonological templates. Treiman et al. (2000) report probabilistic effects relating to the VC rhyme of CVC syllables. Hay et al. (in press) report effects relating to a C.C syllable juncture in English, and Cutler and Otake (1996, 1998) report effects relating to the same position in Japanese and Dutch. Frisch and Zawaydeh (2001) and Frisch et al. (2001) demonstrate the reality of OCP effects on the triconsonantal roots of Arabic, with the consonants in these roots being separated by various vowels in the various words in which the roots appear. (In these studies of Arabic, phonotactic effects and neighborhood effects are distinguished.) Cutler and Butterfield (1992) demonstrate effects in speech perception of the strong but not absolute constraint that English words begin in a stressed syllable. Frisch et al. (2000) report a cumulative effect of the likelihood of CV syllables when combined with each other in two-to-four syllable nonsense words.

All of these phonotactic patterns are readily described using the apparatus of autosegmental and metrical phonology. A combined autosegmental/metrical formalism (as is developed in Pierrehumbert and Beckman, 1988) permits all of them to be understood simply as fragments of phonological description. There is no single privileged level of analysis and the fragments cross-cut each other in the sense that they do not stand in any fixed hierarchical relationship. Taking the syllable as a kind of mental reference point, note that the list I have just given includes patterns which are bigger or smaller than a syllable (the word stress and the syllable rhyme), as well as syllables which happen to be diphones. They also include syllable junctures, containing just the end of one syllable and the beginning of the next, as well as consonantal projections which abstract across variations in syllable structure.

These observations are at odds with a thread in the psycholinguistic literature, which has sought a privileged unit of analysis. For example, in the (subsequently updated) speech production model of Levelt (1988), the syllable is the unit of production. This idea makes it hard to understand why some complex syllables such as /z ɪ m p/ are readily produced and judged as acceptable despite the fact that they do not occur in the lexicon at all, at least as indicated by a search of CELEX. However, the observations fit in readily with the framework of Data Oriented Parsing (DOP) as described in Bod (1998). In DOP, as developed for automatic grammar acquisition in syntax and semantics, all partial formal descriptions up to some threshhold of complexity are projected from any experienced utterance. Frequency counts for each descriptor are incremented as more and more utterances are encountered, and these frequency counts are exploited in scoring alternative parses of novel incoming forms. Compared to syntax, in fact, phonology seems like a very attractive forum for DOP because the phonological grammar lacks recursion and the known constraints appear to have relatively simple formal structure. Below, in the section on statistical robustness, I will return to the question of what constraints are possible in relation to the issue of what constraints are

statistically learnable from a lexicon of realistic size.

2.4 Morphophonological correspondences

The last level of representation introduced above was that of generalizations about relations between or amongst words. This is the level of morphophonological correspondences. Experimental studies of morphophonological generalizations indicate that their psychological reality and productivity has a dependence on type frequency, e.g the number of different word pairs which instantiate the generalization. Cena (1978) reports that the productivity of the English Vowel Shift (as in *serene: serenity, cone: conic*) depends on the specific vowel pair involved, only being productive for the more frequent vowel pairings. A study of conjugation of novel verbs in Spanish by Bybee and Pardo (1980) found that conjugation patterns exhibited only in a few high frequency verbs failed to generalize; those exhibited in more than six different mid-frequency verbs did generalize. Bybee (1995b) discusses the role of type frequency in the formation of German participles for verbs, and plurals for nouns. She shows that the default participial form of German is the most frequent one (contra Clahsen et al., 1992) and that type frequency also impacts the productivity of the default German plural ending /s/.

It is known that children need to have acquired a critical mass of examples before they project a morphophonological generalization. (Marchman and Bates, 1994). However, alternations are not necessarily acquired in order of frequency. Since morphophonological generalizations are generalizations over word pairs, the cognitive availability of the generalization depends not merely on the existence of the two words separately in a child's lexicon, but also on the perception that they are related. For example, the words *bicycle* and *biscotti* both contain (diachronically) a common morpheme meaning "two"; *biscotti* are twice-cooked. The relationship between these words goes unnoticed by many adult speakers. For a young child, the relationship between *imagine* and *imagination* might not be established, even if the child knew both words separately. Raimy and Vogel (2000) argue, in a similar vein, that the notoriously late acquisition of the distinction between compounds such as *BRICK warehouse* (a warehouse for bricks) and phrases such as *brick WAREhouse* (a warehouse made of bricks) is actually due to the complexity and irregularity of the semantic relationships expressed.

The scientist seeking to model frequency effects in morphophonology similar faces a challenge in determining what word relationships should figure in the universe over which probabilities are established. There is no comprehensive formal treatment of what word pairings or word sets are relevant. Many of the clearest results have been obtained with inflectional morphology, in which the paradigmatic organization of the forms is less controversial than for derivational morphology.

2.5 Interim summary

In summary, probabilistic effects are known to exist at all levels of representation of sound structure. These effects are perhaps most widely acknowledged in the area of phonetic encoding. Statistical classification models which originate with the foundational works of mathematical psychology can be adapted and extended to model how different languages utilize the phonetic space in different ways. A fairly large body of experimental work also reveals the existence of probabilistic implicit knowledge of word-forms in the lexicon. This knowledge is revealed both in speech processing and in tasks

closer to the traditional purview of phonology, such as well-formedness judgments and allophonic outcomes. Following results in psycholinguistics, we distinguish two aspects to this knowledge. One is the epiphenomenal knowledge of the lexical neighbors of a given stimulus, being the set of words which are so similar to the given word that they are activated in speech processing. The other is the long-term abstraction of the phonological grammar over the entire lexicon. Frequency effects are found at both levels. We also observe probabilistic effects of word-form relationships in the area of morphological alternations and productivity.

3 Expected and observed frequencies

The productivity of phonology indicates that humans have internalized an abstract system for making complex forms from simpler parts, and linguists have the scientific goal of characterizing this system. Comparing alternative characterizations of the system bring out an inverse relationship between grammatical complexity of the system and the productivity it can describe. At one extreme of this relationship lies the simplest possible grammar, an arbitrary cross-product of the phonological inventory (e.g. any combination of elements, in any order and of any length). This cross-product provides the maximum number of forms which could be generated with the inventory of elements. At the other extreme lies the list of word-forms actually observed in the lexicon. That is, we could take the words in the lexicon as a set of positive grammatical constraints which license all and only the stated combinations of elements. The lexical list would provide the minimum number of outputs consistent with our data set to date, and it supports zero productivity in the sense that any thus far unattested word is taken to be impossible.

Obviously, the true state of phonology is somewhere in between these extremes. We attempt to propose specific grammars whose productivity agrees well with the extensions that humans make, and whose restrictions agree well with the forms observed to be absent. It is just as important to explain the impossible forms as to explain the possible ones. This enterprise requires a way to determine what forms are systematically absent; systematically absent forms reflect complications of the grammar, in comparison to the simpler grammar which would generate them. Comparison of observed frequencies to expected ones provides the means for making deductions of this kind.

3.1 Learning, inference, and underrepresentation

The language learner is in many ways similar to a scientist, an analogy developed at length in the "theory theory" of Gopnik et al. (1999). The learner is attempting to construct a synopsis of the forms s/he has encountered which is general enough to be productive while being restrictive enough to rule out impossible forms. If the grammar is not general enough, the learner will not be able to process novel forms generated by other people or to express new meanings. If it is too general, the learner will generate novel forms which nobody else can understand. Although the learner may have specific perceptual and cognitive characteristics which differ from those of the scientist, this broad analogy means that general mathematical bounds on inferring grammars from data constrain the scientist and the learner in the same way.

A notorious issue for the language learner is the lack of negative evidence. Linguistic scientists are in exactly the same difficulty. Although we can obtain negative judgments about well-formedness,

these are a very limited tool. As discussed above, the cognitive status of well-formedness judgments is under dispute, since they are known to conflate effects at different levels. As sociolinguistic studies have shown, well-formedness judgments are vulnerable to perceived social norms and do not always conform to more naturalistic data. Even if the judgments were valid, there is no way we could obtain enough of them. No matter how many such judgments we obtain, they are a sparse sampling of the set of forms the grammar should rule out. Lastly, linguists can't reliably produce forms which are impossible in their own language, and synthesizing the forms using a computer leaves no way to ensure that they exhibit the allophony that they would have had if they were possible. Therefore, there is no reliable way to present impossible words to subjects in a way which guarantees they are encoded with the phonological representation that the researcher has in view.

This situation has the result that statistical underrepresentation must do the job of negative evidence. Recent studies indeed show that children are quite sensitive to the statistics of sound patterns. By using statistical tools, linguists can also surmount the problem of negative evidence. It is also the case that underrepresentation has limits as a tool for deciding amongst alternative formal theories. We can turn these limits back on the theory, and infer that there are some questions which the language learner must not be asking, because they could not in principle be answered.

A phonological configuration which is systematically underrepresented is one which appears less than one would expect if the grammar had no constraint disfavoring it. Thus, any claim about underrepresentation must be made within the context of a comparison between two grammars: a simpler grammar which provides the null hypothesis for the calculation, and the more complex and restrictive grammar whose status one wishes to evaluate. A choice for a simpler grammar which is always available is the generalized cross- product described above. Under this grammar, each element has a frequency (estimated by its count relative to the size of some corpus) and elements combine at random. The expected frequency E of any combination of elements P1 and P2 is accordingly the product of their two frequencies:

$$E(P1P2) = P(P1) * P(P2) \tag{1}$$

In practice, phonologists computing expected frequencies acknowledge that phoneme frequencies differ with position (position in the syllable and/or position in the word). That is, a partial phonological grammar is presupposed, and competition is between this presupposed grammar and a possible complication of it.

A major application of this reasoning is work on the Obligatory Contour Principle as it affects homorganic consonants in Arabic and other languages. Combinations of consonants at the same place of articulation are disfavored, with the degree of underrepresentation being a function of the similarity and proximity of the consonants as well as of the specific language. McCarthy (1988) and Pierrehumbert (1992) both present results on this effect as it applies to the triconsonantal roots of the Arabic verbs. In their calculations of E, positionally correct phoneme frequencies are used (C1, C2, or C3 position). Berkley (1994, 2000) presents similar results on English, French, and Latin. She overturns the claim of Davis (1991) that /t/ is somehow exempted from OCP effects which apply to other stops in /sCVC/ configurations. He notes that forms like /spɛp/ are bad, whereas words such as *stats* actually exist. However, Berkley is able to show that the *stats* case is also underrepresented, suggesting a more uniform grammar than Davis proposed. Frisch (1996) also presents further calculations on Arabic which relate the degree of underrepresentation

of consonant pairs in Arabic to the degree of similarity between the consonants as computed from a psychologically motivated metric. A gradient relationship is found, in which the proscription against totally identical consonants emerges at one end of a continuum on similarity. These results obviate the need to posit two different grammatical mechanisms in order to capture the fact that the proscription against identical consonants is statistically stronger than the constraint against similar but nonidentical consonants.

Another application of the idea of expected value is found in Pierrehumbert (1994). Pierrehumbert computed expected frequencies of occurrence for all medial clusters of three or more consonants in monomorphemic words (such as the /lfr/ in the word *palfry*, by assuming that these arise as random combinations of codas and onsets. The positional frequencies for codas were roughly approximated by the frequencies of the various consonants in word-final position once appendices were stripped off; the positional frequencies for onsets were approximated by frequencies of onsets in word-initial position. Pierrehumbert showed that the vast preponderance of long medial clusters are expected to occur less than once in any lexicon of realistic size. For example, a medial cluster such as /lskr/ in the nonsense word *pelskra* is syllabifiable and violates no sequential constraints (c.f. *dell, else, screw*). It is unattested in any monomorphemic words of English found in the Collins on-line dictionary distributed by the Linguistic Data Consortium. Due to its relatively rare subparts, its expected count in a dictionary of this size is less than one. Thus, its absence is expected under a syllable grammar which includes positional probabilities for phonemes.

In a probabilistic model, no complication of the grammar is necessary to explain such cases. If the grammar lacked positional probabilities, then the absence of these forms would need to be modeled by complicating the grammar with additional constraints. The same argument can be made for more than 8400 additional long clusters which are unattested in monomorphemic words. The article also identifies a number of syllable contact constraints, with observed counts less than the values expected from random combination of codas and onsets. These include an OCP effect leading to underrepresentation of forms with a C1C2C1 cluster (e.g. the hypothetical *palfly* and a constraint disfavoring coronal obstruents in word-internal coda position (e.g *pirtfy.) An experiment involving two different judgment tasks showed that these systematically underrepresented patterns relate to psychologically real constraints.

3.2 Sample size

Both the statistical analyses of the Arabic OCP and Pierrehumbert's analysis of the long medial clusters of English are made using large on-line dictionaries yielding data sets which are realistically related to the total mental lexicon of an adult speaker. An individual Arabic speaker probably does not know a great many more triconsonantal verb stems than the 2674 found in the Cowan (1979). The Collins on-line dictionary contains 69737 phonologically distinct entries, and is similar in its coverage of monomorphemic words to the Celex dictionary used later by Hay et. al (in press), and discussed below. Data sets such as these can be viewed as random samplings of the larger, hypothetical, data set in which all the latent potential for lexical additions is actually instantiated. The constraints governing the larger data set must, however, be inferrable from a sampling of the potential data set. This is the case because adult speakers have learned the phonology of their language, and learning the phonology cannot require more word-forms than individual adult speakers know.

In general, the bigger the sample, the more confidently can patterns can be established, and in more complexity can patterns be established with confidence. The sample sizes for monomorphemic words (at around 10,000) are relatively small in comparison to the complexity of hypotheses phonologists might entertain. This is even more true for constraints such as the Arabic OCP which are active on only a subpart of the vocabulary (e.g. the verbs, which participate in nonconcatenative morphology and hence have strong projection of the consonantal tier.) To appreciate the force of this observation, let us consider how much data are required to be confident about two different constraints of English, differing in their statistical force.

The first is the constraint that a triphonemic monosyllable must contain a vowel or vocoid in English (unlike in Berber, well known for permitting even obstruents in nuclear position.) For the sake of exposition, I make the simplifying assumption that these monosyllables are all (in fact) of the form CVC, VCC, CCV. There is probability 1/3 that a phoneme is a vowel/vocoid in this universe, the probability 2/3 that is not. (Probabilities sum to one and in this case there are only two alternatives). Our null hypothesis will be that phonemes combine at random, and so positional probabilities will not figure in the null hypothesis. Under the null hypothesis, the expected probability of a CCC form is $(2/3)^3 = 0.296$ Under the null hypothesis we would need to look at 3 or 4 forms in order to expect to find one of form CCC. But if such a small item set fails to have any CCC forms, that provides very weak evidence that this outcome is impossible; on the average, one expects to get a 6 once on six roll of a die, but rolling a die 6 times without getting a 6 is not good evidence that the die has no six. Specifically the probability under the null hypothesis of obtaining no CCC forms in a sample of size n is;

$$(1 - (2/3)^3)^n = (19/27)^n \tag{2}$$

The bigger n gets, the smaller this probability is and the more confident we become that the null hypothesis can be rejected. In this particular case, the null hypothesis becomes rapidly disfavored as n increases. For n=9 (a sample of 9 triphonemic words, of which none prove to have the form CCC), the probability of the null hypothesis being true is already less than 0.05. For $n = 14, P < 0.01$ and for $n = 20, P < 0.001$.

Figure 3 provides graphical understanding of the comparison between the null hypothesis and the actual state of affairs, for the case of a 14-word sample. There are two probability distributions on this figure, the one on the right having the common lumpy shape and the one on the left being a spike located on top of the y-axis. The distribution on the right shows the distribution of counts of CCC words in a 14-word sample under the null hypothesis. Obviously, this distribution peaks around 4. In a 14 word sample, we expect to find $(0.296 * 14) = 4.144$ CCC words. However, we could find more or less by the luck of the draw, as the figure shows. The "spike" distribution, to the left, represents the distribution of counts of CCC words under the hypothesis that the probability is 0. This is a degenerate situation in which the distribution has zero variance, because luck of the draw will never provide more or fewer than zero CCC words. The two distributions in this figure are well separated, exhibiting only a barely visible overlap at zero, where the left-hand distribution has a value of 1.0 (because all the probability is at a single location), and the right-hand distribution has a value of .007. The nearly perfect separation of the two hypotheses is what allows them to be distinguished in practice with such a small data set.

PUT FIGURE 3 ABOUT HERE

In this example, a very small amount of data causes confidence in the null hypothesis to deteriorate rapidly. That is because there was a huge difference between the probability of a CCC according to the null hypothesis, and its actual probability. When the quantitative difference between the two hypotheses is smaller, the rate at which increasing the sample size nails the point is corresponding less.

For example, let us consider how large a sample is needed to infer that the probability of /s/ in word onset position differs systematically from its likelihood in word final position. I will assume (idealizing somewhat) that the positional frequencies of /s/ in the Celex monomorphemes are the true underlying frequencies for this comparison. (Counts in morphologically complex words could be misleading, because some words and word-level affixes, such as *-less* and *-ness*, show up repeatedly in complex words). The probability of /s/ initially is 0.1153 in this set, and the probability of /s/ finally is 0.0763. Note first that for a word sampling of size 9, we expect to find 0.1153 * 9 = 1.06 cases of initial /s/, and $0.0763 * 9 = 0.6867$ cases of final /s/. Rounding to the nearest integer, a sample of size 9 yields the same expected count of one initial /s/ and one final /s/. Thus, a sample size that already was decisive (on a typical significance threshhold of $P <= 0.05$) for the case of the *CCC constraint is too small for one to expect any difference for the positional /s/ frequencies. The situation is explored further in the three panels of Figure 4, constructed along the same lines as Figure 3.

PUT FIGURE 4 ABOUT HERE

The left hand distribution in Figure 4 is the probability distribution for the counts of final /s/s, and the right hand distribution is that for the initial /s/s. Due to the fact that we are now varying the counts by two orders of magnitude across the panels of the figure, the x-axis has now been normalized by the sample size n. In the first panel, the sample size is 14 (the sample size for achieving significance level $P <= 0.01$ for the *CCC constraint). The heavily overlaid distributions reveal how poorly discriminable the two cases are with this small of a sample. In the next panel, the sample size is 140, and the overlap of the distributions is reduced. In the last panel, corresponding to a sample size of 1400, the distributions are well separated, but still show some degree of overlap.

Thus, it requires at least two orders of magnitude more data to evaluate positional effects on /s/ than to discover the underrepresentation of CCC monosyllables. This situation comes about through two factors. The *CCC constraint is very general, since C is a broad category whose overall probability in the system is 2/3. /s/ is a more specific category which is instantiated less frequently. In general, the more specific a phonological description is, the fewer forms it will be true over and the more data will be needed to distinguish its probability from the probability of any of its close formal relatives. Also, in the *CCC example, the null hypothesis of P = 0.296 differs greatly from the actual state of affairs to which we compare it. The positional effect on /s/ was relatively slight, with the probability for final position being about 2/3 of the probability for initial position. A slight difference is harder to evaluate than a massive one.

It is important to keep in mind that the probability of an anomalous run of events under the null hypothesis will never be zero, it will only get smaller and smaller. The probability of throwing all Heads on n throws of a fair coin is $(1/2)^n$, so one could get million heads in a row with a probability of one half to the millionth power, which though small is greater than zero. If we imagined that people had vocabularies of infinite size, then no amount of data would suffice to prove that CCC monosyllables are truly impossible, we could only show that they are vanishing rare.

3.3 Vocabulary size

Since people do not have vocabularies of infinite size, the sizes of actual vocabularies can be used to place a bound on the statistical resolution which is either possible or useful. Estimation of vocabulary size is a difficult issue. Productive and receptive vocabulary differ. It is unclear which morphologically complex words are stored and which are derived on the fly. Also, undersampling is a problem even with large data sets. However, some useful reference points can be deduced. One important count is the number of monomorphemic (or root) words in an individual's vocabulary. This set provides the basis for phonological generalizations about within-word phonotactics and prosodic patterns. Lexical items containing internal word boundaries, such as *peppermint* or *matchless* show word-boundary phonotactics even if they have idiosyncratic semantics. Data from Anglin (1993), as graphed in Vihman (1996) indicates that first graders (age 6 to 7) have a receptive vocabularies with about 2500 root words, and by fifth grade (ages 10 to 11) this number has grown to aproximately 6000. The number of monomorphemic words in CELEX, according to a tabulation made by the OSU phonetics lab, is 11,382. As discussed in Pierrehumbert (2001b), this is a quite comprehensive list which probably exceeds the monomorphemic inventory of any single individual. For phonological constraints over monomorphemic words, probabilities of less than 0.0005 are effectively zero (since the expected count in a sample of size 10,000 would be zero). Any phonological constraints which are learned early in life need to be learnable on much smaller data sets, a point to which I return below.

For other phonological processes, such as stress patterns and morphophonological alternations, a fuller vocabulary is relevant. Anglin (1993) estimates that fifth graders know 25,000 root words plus derived words and idioms. (Productive and unproductive derivation do not appear to be distinguished in this estimate). A detailed study of printed school English by Nagy and Anderson (1984) yields the estimate that printed school English includes approximately 89,000 word "families". This number includes as distinct entries words which are morphologically complex but which have moderately to extremely opaque semantics. For example, in this study, *collarbone* is distinct from *collar*. However, the authors assume that *senselessly* can be productively derived from *senseless* and *stringy* can be derived from *string*. The existence of 89,000 word families in the entire corpus of printed school English does not mean that any individual child knows all the words. Example words provided in the discussion include words such as *solenoid*, *hornswoggle* and *ammeter*, and it appears improbable that any single school child knows all them; discussion draws attention to the ability to read materials containing some unknown words. The same article gives a range of 25,000 to 50,000 as the total number of words known by a high school senior. From figures such as these, we can infer that the total vocabulary of an adult is an order of magnitude larger than the vocabulary of monomorphemic words, but probably less than two orders of magnitude larger. If a statistical inference is so delicate that it cannot confidently be made with a lexicon of around 100,000 words, there is no way for learners to make it and its cognitive status is highly dubious. If a probability over this set is < 0.00005, it is effectively zero.

4 Discrimination and robustness

We have just seen that a full blown adult lexicon, though large, is still not large enough for grammatical constraints of arbitrary subtlety to be deduced from it. In practice, the limits on grammatical subtlety are much more severe. First, children learn much of phonology at a young age, before they have large vocabularies. Secondly, different children have different vocabularies, and adults likewise. The commonplace observation that the grammar can be learned from poor levels of language exposure amounts to saying that it can be learned from a severe downsampling of the maximal data set, and it can be learned equally well from diffferent downsamples of the data set. The fact that people with different language exposure can end up with essentially the same grammar also means that language acquisition is resistant to outliers (or statistically anomalous events). A statistically anomalous hurricane may destroy a large city, but early exposure to the word *Ladefoged* does not destroy a child's stress system. These marvels of the acquisition system amount to the claim that language is statistically robust. This section develops the idea of robustness further and shows how it constrains the nature of the phonological system.

4.1 Robustness in categorization

Statistical robustness has been most explored in relation to categorization of the phonetic space. Vowel inventories have provided a particularly fruitful subtopic due to the ease with which the relevant phonetic dimensions can be conceptualized and manipulated. Accordingly, I resume discussion of the perception and production of phonetic categories in connection with the /ɪ/-/ɛ/ example of Figure 2.

In Figure 5, idealized distributions for /ɪ/ and /ɛ/ have been plotted on the assumption that the underlying distributions are Gaussian and that both categories have been so abundantly experienced that the mental representations approach the true underlying distributions. Thus, the distributions in Figure 5 are smooth instead of showing individual spikes for memory traces. Three cases are compared. In panel A, the means of the distributions differ by 200 Hz and there is slight overlap between the two vowels. In panel B, the distributions have been shifted so they differ by only 100 Hz. In panel C, the means are as just separate as in panel A, but each category has more variability so there is more overlap.

PUT FIGURE 5 ABOUT HERE

Since the distributions are smooth, the classification of an incoming token at any point can be read off of the graph by seeing which distibution line lies above the other at that location on the x-axis. Figure 6 shows how this works by reproducing panel 5C with additional annotations. The listener is attempting to classify a stimulus at the F2 location indicated by the up-arrow. The line for the label /ɪ/ is above that for /ɛ/, so /ɪ/ is the most probable label. It is immediately obvious that given the distribution type under consideration, there is an some F2 value such that any stimulus with an F2 above this value will be classified as /ɪ/, and any stimulus with an F2 value below this will be classified as /ɛ/. This cutoff is defined by the intersection of the two distributions, and is shown with a vertical dashed line in the figure. R.D. Luce et al. (1963a) provide a mathematical treatment of classification situations in which a threshhold or cutoff is well-defined.

PUT FIGURE 6 ABOUT HERE

For the situation shown in Figure 6, the result is that a sizable region of the /ɛ/ production space will be perceived as /ɪ/, and a certain region of the /ɪ/ production space will be heard as /ɛ/. The extent of such confusions would be exactly the same for Figure 5B. Although the separation of the means is half as much in 5B as in Figure 6, the distributions are also narrower and the degree to which the distributions overlap is (by design) exactly the same. To achieve reliable discrimination, in which the intention of the speaker can be recovered by the listener with very little chance of error, we need the situation displayed in Figure 5A. Here, the separation of the means of the distribution is large in **relation to their width.**

Discrimination is not the same as statistical significance. As the quantity of data contributing to the distributions in Figure 5B or 5C increases towards infinity, the difference in the means of these distributions can become as statistically significant as one could please. That is, if the two labels are given, we can become more and more sure, scientifically speaking, that the distributions for these labels are not the same. But no amount of certainty on this point improves the situation with respect to discrimination of the categories when classifying an unknown incoming token. In a similar vein, thanks to millions of test-takers, it is extremely certain that the distribution of verbal SAT scores is slightly higher for women than for men. (The SAT is a standardized test administered to high school students applying to university in the United States). However, since this difference is slight compared to the variation of scores for each gender, knowing someone's verbal SAT score does not permit one to deduce their gender with any reliability.

Statistical discrimination plays a central role in theories of the phonetic foundations of phonology, in particular pioneering work by Lindblom and colleagues on adaptive dispersion theory for vowels. (See Liljencrants and Lindblom, 1972, Lindblom 1986, Lindblom, 1990). The model is predicated on the understanding that the language system requires robust discrimation of categories in order to guarantee the integrity of communication. Given that speech production and perception are intrinsically variable, this characteristic places bounds on the number of categories that can be maintained over the same space, as well as constraining their optimal arrangement with respect to each other. The effects are seen both in segmental inventories and in the way speakers manipulate the vowel space in continuous speech, making the minimal effort needed to maintain sufficient contrast. One consequence is that total phonetic space will be underutilized in any particular language, since accuracy of discrimination depends on statistical troughs like that in Figure 5A. Application of adaptive dispersion theory by other researchers include Engstrand and Krull (1994), who demonstrate reduced durational variance for vowels in languages in which vowel length is distinctive, and Miller-Ockhuizen and Sands (2000), who apply the model to explain phonetic details of clicks in two related languages with different click inventories.

The equations of adaptive dispersion theory treat perceptual discriminability (contrastiveness) synoptically as a direct pressure on productions. The same effects can also be approached in terms of how category systems evolve over time, because discrimination is related to the stability of the systems as they evolve. Assuming that distributions are used in a perception-production loop, as discussed in section 1, the classification of stimuli in perception provides data for the probability distributions controlling production. (The productions of one individual provide perceptual data for another individual, so that the perception-production loop goes through members of the speech community.) Situations such as Figures 5B and 5C are intrinsically unstable over time, because in any region of the phonetic space where the distributions overlap, productions of the less likely label will be misclassified in perception.

Assuming only that the probability distributions are updated with perceptual data (an assumption needed to model how phonetic learning can occur in the first place), there are two major outcomes, which depend on the degree of distributional overlap in relation to the random variation in the perception and production processes. One outcome is that the two distributions sharpen up, each on its own side of the discrimination cutoff. The other major outcome is that the distribution for the less frequent label gets entirely eaten up by the distribution for the more frequent label, as each fresh misclassification enhances the frequency advantage of the more frequent label. This results in a collapse of the category system. Pierrehumbert (2001a) presents in detail the case of loss of contrast through a historical lenition process. A small but systematic lenition bias in production of the marked (less frequent) member of a contrast eventually results in its collapse with the more frequent member. Note that both of the major outcomes are characteristized by a distinct distributional peak for each category label. Thus, consideration of the stability and robustness of categories over time provides a basis for the observations of adaptive dispersion theory, as well as Kornai (1998) and Maye et al. (2000, in press). Well-defined clusters or peaks in phonetic distributions support stable categories, and poor peaks do not.

Such observations also have important implications for our understanding of the abstractness of phonetic encoding. The calculations presented by Lindblom, Miller-Ockhuizen, Kornai, and Maye et al. all deal with phonemes in a fixed context. Thus, they do not clearly differentiate phonemes (which are equivalenced across different contexts) from positional allophones. However, distributions of phonetic outcomes in continuous speech have revealed many cases in which the realization of one phoneme in one context is extremely similar or even identical to that of some other phoneme in another context. For example, Pierrehumbert and Talkin (1992) show the distribution of vocal fold abduction (or breathiness) for /h/ overlaps that of vowels when both are tabulated out of context; when tabulated relative to context, there are just a few sporadic cases of overlap. Overlap of devoiced allophones of /z/ with /s/ is discussed in Pierrehumbert (1993). Cases such as these mean that parametric distributions for phonemes are not always well distinguished from each other, if they are tabulated without regard to context. Within each given context, the distributions are much better distinguished. Thus, positional allophones appear to be a more viable level of abstraction for the phonetic encoding system than phonemes in the classic sense. Further implications follow from cases in which the phonemes are well distinguished in one context but not in another. These situations favor positional neutralizations. Steriade (1993) explores this concept in connection with stop consonants, demonstrating neutralization in contexts in which the stop release is missing. Flemming (1995) extends this line of reasoning in the framework of Optimality Theory.

Category discriminability also has important ramifications in sociolinguistics. An important set of studies (Labov et al., 1991; Faber and DiPaolo, 1995) deals with cases of near-merger, in which the phonetic distributions of two categories have become heavily overlapped (due to historical changes and/or dialect variation) but a statistically significant difference is still observable in productions. A case in point is the near-merger between *ferry* and *furry* in Philidelphia English, as discussed in Labov et al. (1991). A surprising behavioral finding is that subjects whose productions display an acoustic difference between the vowels of such word pairs are unable to distinguish these words at above chance levels. This is true even if they are listening to their own speech.

In the class of model I have been discussing, it is to be expected that discrimination in perception will be worse than the discrimination analysis which the scientist can carry out on objective acoustic measures. This is the case because of intrinsic noise in the perceptual system. The scientist can

carry out a statistically optimal analysis on a microphone signal of the highest quality, whereas the perceptual system is limited by the critical bands of the auditory encoding, by the background noise from blood rushing through the ears, and so forth. However, it is also the case in this model that it is impossible to acquire a phonetic distinction which one cannot perceive. The subjects in the near-merger study must have been able to perceive a difference between *ferry* and *furry* at the time their production patterns were being established. Otherwise, the labeling and phonetic distributions needed to produce such a difference would never have been acquired. The model predicts that whenever a contrast becomes truly imperceptible in a language, then the contrast will collapse and this collapse will be irreversible. This prediction is born out by the assertion in Labov (1994) that total neutralizations are never reversed. Apparent cases of reversal can be traced to the survival of the distinction in some context. This can be a social context, as in cases in which a contrast is reimported into the speech community from a dialect or an influx of borrowings. Or the contrast may survive in some other context within the cognitive system, as when a contrast is reimported from morphologically rich orthographic system. Similarly, detailed phonetic studies also reveal that morphological relatives can serve this role through paradigm uniformity effects at the allophonic level. A case in point is Port et al's (1981) study of incomplete neutralization of obstruent voicing in German.

The failure of Labov et al.'s subjects to perceive a distinction in their own speech thus requires more explanation, assuming that this failure is more severe than perceptual encoding noise alone would predict. The levels of representation I have outlined provide an avenue of explanation. The task in Labov et al. (1991) is a word judgment task, and hence involves access to the lexicon. Thus, one must consider not only the phonetic encoding in perception, but also the relationship of this encoding to the stored word-form. If subjects have learned, from exposure to the varied dialect community of Philadelphia, that vowel quality information does not reliably distinguish *ferry* and *furry* , then they can downweight this information as a perceptual cue in lexical access. They learn not to pay attention to a cue they cannot rely on. This interpretation of the situation is born out by a related study by Schulman (1983) on the perception of a *sit, set, sat, sot* continuum by bilingual Swedish-English speakers from Lyksele. The speakers were unable to hear the *set, sat* distinction when experimental instructions were delivered in Swedish, a language in which dialect diversity renders the distinction unreliable. However, they could hear the distinction when instructions were delivered in English. In short, this study indicates that the attentional weighting of different phonetic cues is not an entrenched feature of the cognitive system; instead, it can be varied according to the speech situation.

4.2 Robustness of phonological constraints

Statistical robustness is a well established and major theme of the literature on phonetic encoding. Less widely acknowledged is the fact that it also plays an important role at more abstract levels of representation. In fact, the small size of lexicon, in comparison with the huge body of experienced speech, places severe limits on statistical inference over word forms as well as correspondences between word forms.

This issue is taken up in Pierrehumbert (2001b), a set of pilot calculations addressed to the issue of why known phonological constraints are so coarse-grained. Though a DOP approach would permit us to construct a huge proliferation of phonological templates using the logical resources of

phonology, the constraints which appear to be psychologically real by any test are relatively simple and nondetailed. Note in particular that they are simple in comparison to the lexical representations of individual words, as we have already seen from the comparison of lexical neighborhood effects with phonotactic effects.

The method used in this paper was a Monte Carlo simulation of vocabulary acquisition; hypothetical vocabularies of different individuals at different levels of development were estimated by random sampling of the 11383 monomorphemic words in Celex. Since the goal was to discover the constraints on within-word sequences, morphologically complex words liable to contain internal word boundaries are excluded from this training set. The sampling was weighted by word frequency, on the assumption that a individual is more likely to learn a frequent word than an infrequent one. Inventories for 20 different "individuals" at each vocabulary level were computed. 400, 800, 1600, 3200, and 6400 word samples were computed. The vocabulary level needed to learn a given constraint reliably is determined by inspecting these individual vocabularies to see whether the constraint is actually manifested for most or all individuals. The paper compares the learnability of two real phonological constraints of English to that of two unrealistic constraints which are quite a bit more detailed.

The realistic phonological constraints are the English constraint on foot alignment in the word, and a constraint set governing word-medial nasal obstruent clusters. The first refers to the fact that the basically trochaic foot structure of English is extended to trisyllables by aligning the foot to the left edge rather than the right edge of the word. The result is a 100 stress pattern (as in *parity*) rather than the 010 stress pattern that one finds in some other languages. The treatment of the nasal- obstruent clusters is taken from the detailed experimental study of Hay, et al (in press). This study, also discussed above, showed that the nasal-obstruent (NO) constraints are part of the speaker's implicit knowledge and that perceived well-formedness is gradiently related to the frequency of the different clusters.

One of the two unrealistic constraints evaluated in this paper is a hypothetical constraint involving a combination of the 100 stress template with the nasal-obstruent regularities. That is, we pose the question of whether it is reasonable for a language to have different nasal-obstruent constraints for trisyllabic words with a specific stress pattern. The other is the statistical pattern which is the object of an experimental study by Moreton (1997). The experiment looked for a phonemic bias effect as a corollary of a difference in probability, comparing word final stressed /gri/ as in *degree* and word final stressed /kri/, as in *decree* /gri/ is much more common than /kri/ in running speech, chiefly because of the high frequency of the word *agree*. There is also a contrast in type frequency of the two patterns, but it is much smaller. This example was selected because Moreton obtained a null result in his experiment. It presents one of the few cases anywhere in the experimental literature for a negative finding on the psychological reality of a statistical phonotactic pattern.

The calculations showed that the 100 stress pattern is extremely robust, learnable by all 20 individuals from a mere 400 words. A rank order of five nasal-obstruent combinations could be reliably learned from 3200 words, a vocabulary level realistically achievable by an elementary school child (see above). The two unrealistic constraints were not reliably learnable even at a vocabulary level of 6400 monomorphemic words. The fact that Moreton's subjects had not (apparently) learned the /gri/-/kri/ constraint is therefore unsurprising.

Pierrehumbert (2001b) views vocabulary levels in terms of stage in language acquisition. The

same line of reasoning, however, leads to the prediction that adult speakers with different vocabulary levels should have somewhat different grammars. In particular, adults with large vocabularies should be in a better position to estimate low, but nonzero, probabilities. This prediction is born out by results in Frisch et al. (2001). They found that adults with large vocabularies have a more generous threshhold for viewing nonce forms as acceptable, and they present calculations showing that this effect, for their stimuli, cannot be attributed to lexical neighborhood effects.

A new series of calculations extends this line of reasoning by exploring phonemic n-phones more systematically. The nasal-obstruent sequences contain two phonemes, and as such are examples of diphones. Even disregarding the positional information, /gri/ and /kri/ are triphones. Bailey and Hahn (2001)'s study of lexical neighborhood densities and phonotactic statistics, discussed above, systematically evaluated the role of diphone statistics in comparison to triphone statistics. Diphone statistics were an important predictor of wordlikeness judgments, but triphone statistics were not found to add any predictive power beyond that provided by diphone statistics. The goal of the present calculations was to determine the extent to which diphone and triphone constraints on the internal form of words can or must be learned from the lexicon. If triphone constraints are not learnable as such, then the well-formedness of triphones must be estimated from its subparts. Under this assumption, the well-formedness of /str/ sequence in *street* would follow from the well-formedness of /st/ and /tr/; the ill-formedness of /stl/ in **stleet* would follow from its ill-formed subpart /tl/. If triphones are generally and necessarily parsed in terms of their subparts, then triphone statistics would add no predictive power in a model of perceived well-formedness, just as Bailey and Hahn report.

The calculations were again made on the Celex monomorphemes. The transcription set contains 37 phonemes, disregarding 11 examples of distinctively nasalized vowels in French borrowings. The transcribed distinction between syllabic and nonsyllabic sonorant consonant was also disregarded on the grounds that it is predictable from sonority sequencing. When these are sonority peaks in their sequence (as in *apple*) they are syllabic, otherwise not (as in *mill*). This means that the full cross-product of the phonemes, providing the baseline against which constraints must be evaluated, generates 1369 different diphones and 50,653 triphones.

Basic counts show the general feasibility of training word-internal diphone statistics from the inventory of extant monomorphemes. The 11383 Celex monomorphemes display 51,257 tokens of diphones. Thus, the data set is 37 times bigger than the constraint set that the learner is attempting to parameterize. Downsampling the data set to 3200 words (the count around which nasal-obstruent constraints began to be reliably learnable) still yields a ratio of approximately 9 data points per diphone, on the average. This is an order of magnitude more data than constraint set parameters, and equivalent to the count that permitted *CCC to be detected at $P < 0.05$ in the tutorial example above.

The learnability situation is quite different when we turn to the triphones. There are fewer triphones per word than diphones. For example, a four phoneme word contains two triphones, namely those starting at the first and second phonemic positions, but three diphones. Thus, there are only 39877 examples of triphones the training set. Meanwhile, the candidate constraint set has exploded to 50,653 different forms. This leaves an average of less than one data point per constraint parameter, even on the implausible assumption that the whole vocabulary was available to the young language learner. Even the number of triphones which actually occur at least once in the monomorphemic inventory – namely 5414 – is large relation to the level of lexical knowledge

of children. In short, it is mathematically impossible to assign probabilities to triphones generally, on the basis of knowledge of monomorphemic words. If the general assumptions presented here are valid, then it should be possible to predict the well-formedness of triphones from more general constraints, which are learnable from the available data.

This raises the issue of how well triphones may be predicted from their diphone subparts. We look at this issue in two ways. For table [2], the expected count of each triphone in an inventory of the actual size of the training set has been computed. To compute these numbers, for each triphone P1P2P3, P3 is appended to P1 P2 with the conditional probability of P3 following P2 in a diphone. If this expected count (rounded to the nearest integer) is at least one, then the triphone is expected to occur. If it rounds to zero, the triphone is expected to be absent. The table breaks out these counts according to whether the triphone is or is not exemplified in the data set. This prediction is quite successful. Out of 45239 triphones which fail to occur, 42,776, or 94%, are predicted to be absent from the rarity of their subparts. Incorrect predictions comprise only 6.6% of the data. Not revealed by the table is the fact that these cases are overwhelming ones in which the count was predicted to be zero and is actually one, or via versa; in short, cases which sit right on the threshhold applied in setting up the table.

PUT TABLE 2 ABOUT HERE

A way of viewing the data which overcomes the threshholding issue and also provides more insight into the relative well-formedness is to plot the predicted rate of occurrence of triphones against their expected rate of occurrence. Such a plot is presented in Figure 7. In Figure 7, the triphones have been ranked by expected count and 10 bins of expected count are established. Within each bin, the median actual rate of occurrence is plotted along with with upper and lower quartiles. Overall, the rate of occurrence is shown to be strongly related to the expected rate of occurrence. The fact that the relationship is not perfect arises both from sampling issues (the actual lexicon can be viewed as a sampling of the potential lexicon) and, more importantly, from the fact that triphones are subject to additional constraints beyond the biphone conditions. An example is the OCP constraint disfavoring C1C2C1 combinations, as discussed above.

PUT FIGURE 7 ABOUT HERE

A comparison of the situation with diphones to that of triphones reveals that it is not only possible, but also necessary, to learn constraints on diphones rather than estimating their likelihood from their (phone) subparts. Table 3, constructed along the same lines as the triphone table, shows that only a few of the absent diphones are predicted to be absent on the basis of phone frequencies alone. Of the 1322 diphones which are predicted to exist under the null hypothesis, 44 percent are unattested. This outcome is not surprising, since the null hypothesis provides no coverage of the pattern of sonority alternation which imposes a syllabic rhythm on the speech stream. One consequence is that a pair of identical high frequency phonemes is predicted to be frequent; due to the high frequency of /ɪ/, the sequence /ɪɪ/ is expected to be the most frequent combination of all, whereas in fact, it is impossible. Most of the numerous unexplained absences in this table arise through factors which a linguist would view as systematic.

PUT TABLE 3 ABOUT HERE

For attested diphones, the phoneme frequencies are also rather poor at predicting the rate of occurrence. This is indicated by in Figure 8, constructed along the same lines as Figure 7. Though biphones containing frequent phonemes are on the average more frequent, the spread around this trend is much greater than in Figure 7. In particular, the lower quartile remains almost at zero all

the way up to the highest expected count.

PUT FIGURE 8 ABOUT HERE

In the absence of additional evidence, it is not possible to conclude that all triphones are necessarily evaluated via their subparts. The most high frequency triphones have counts in the three hundreds, amply sufficient to support an evaluation of their observed frequency relative to the expected frequency from a diphone grammar. It is possible that some triphones do acquire probabilities in their own right. The initial triphone of *Advent* occurs in seven different words despite having a predicted rate of occurrence of zero. The final triphone of *raft* occurs in 13 words despite have a predicted rate of occurrence of one. Possibly, people may have some implicit knowledge of these facts; but it would be difficult to demonstrate that any such knowledge goes beyond lexical neighborhood effects. However, in inspecting the list of predicted and attested triphones, it was surprising difficult to find these examples, as cases of blatantly overrepresented triphones are few. The mere assumption that the lexicon is a random sampling of a bigger universe of possible words means that it would exhibit some random variation in the over and under representation of particular combinations. It is not known whether cases of this type are treated as within-the-noise by the cognitive system, or whether they are remembered as important.

The fact that triphones are an unpromising field for larger scale phonological constraints also points up the importance of investigating large scale constraint candidates which are closer to the fruits of linguistic scholarship. In the linguistic literature, constraints with a larger temporal scale (such as vowel harmony constraints and foot structure constraints) generally have the property of referring to classes of phonemes rather than to individual phonemes, if they refer to phonemes at all. As far as trainability goes, an increase in temporal scale is offset by a decrease in featural specificity. As the results on 100 versus 010 stress patterns show, such constraints can be extremely robust.

The nonlearnability of triphones at the level of the phonological grammar provides a sharp contrast with the amount of detail people learn about specific words. Even the most categorical description of a good-sized word, such as *ambrosia* or *budgerigar* is highly complex compared to any known phonological constraint, indicating that brute cognitive limits on representational complexity are not the issue. Moreover, cases discussed above in relation to word-specific allophonic detail and subphonemic morpholophonological paradigm effects show that the representations of words can be extremely detailed. This detail can be acquired because reasonably frequent words are encountered so many times in speech. Counts in Carterette and Jones' (1974) tabulation of conversational speech by adults and children show that the most frequent triphonemic function words (such as *was* and *what*) are up to 50 times as frequent in speech as triconsonantal syllable onsets such as /spr/. A word with a frequency of 6 per million (a threshhold above which word frequency effects are experimentally established, clearly demonstrating long-term storage) would show up approximately once per 15 hours of speech, assuming an average rate of three words per second. Athough people can learn the existence of new words from very few examples, once they have learned a word, they have plenty of evidence to fine-tune its representation.

5 Correlations across levels

The previous discussion emphasizes the need to distinguish levels of representation, each with its own statistical effects. This is reminiscent of the modularity claims of classical generative theory. However, there are important differences. In the generative approach, the task of setting up modules is viewed as a task of partitioning the system into maximally independent components. Each module is as stripped down as possible. In evaluating specific theoretical proposals, any observed redundancy between information encoded in one module and that in another is viewed as a weakness. However, it has proved impossible to eliminate redundancy across modules. The present approach, in contrast, uses rich representations (as also discussed in Baayen). Correlations across levels of representation are viewed as a necessary consequence of the way that more abstract levels are projected off of less abstract levels. Some of these correlations are extremely obvious and would show up in any current approach. For example, in essentially all current approaches, the phonological entities manipulated in the grammar and morphophonology are viewed as contentful; their behavior reflects their general phonetic character as it percolated up the system through successive levels of abstraction.

Some other examples of confluence across levels are far less obvious, and reveal subtle properties of human language.

One nonobvious correlation concerns the relationship of the phonetic encoding system to the phonological grammar. It is known that children can use statistical cues to decompose the speech stream into chunks at a very early age, before they know that words refer to objects and events (see Jusczyk et al, 1994; Mattys et al, 1999). Low-frequency transitions are taken to be boundaries, and these transition frequencies must evidently be estimated using surface statistics, since type statistics would depend on a lexicon which the child has not yet formed. Luckily, these surface (or token) statistics on diphones are highly correlated with type statistics in the lexicon. In general, phoneme combinations which are infrequent in running speech are also infrequent word-internally in the lexicon. This means that a low frequency phoneme transition is a viable cue to the possible existence of a word boundary. As pointed out in Hay (2000), it is possible to design a formal language in which this correlation does not obtain. Imagine, for example, that a language had an invariant start phoneme for words (say, /t/) and an invariant stop phoneme (say, /k/), so that every word began with /t/ and ended with /k/. In this case, the combination /kt/ would be extremely common in running speech, but would nonetheless be the cue to the presence of a word boundary. Human languages do not have this property. Instead, show maximal contrast sets in word-initial position, and the proliferation of alternatives at word onsets leads to low frequency combinations across the word boundary. The confluence between type frequency and token frequency thus supports bootstrapping of the system during language acquisition and smoothes the communication between levels in the mature system.

A second important case of confluence brings together the issues in phonetic categorization and n-phone statistics which were discussed above. A substantial body of work in phonetics shows that the acoustic landmarks which are provided by transitions between the physical regimes of the vocal tract play a key role in forming discriminable categories. For example, at an obstruent-vowel transition, an aerodynamic factor (the build-up of pressure during the stop closure) conspires with an acoustic factor (a jump in the number of resonances as the vocal tract moves from a closed to an open position) and a psychoacoustic factor (the way the ear adjusts its sensitivity to changes in

amplitude). These factors together mean that obstruent-vowel transitions are perceptually salient and reliably distinguished from each other, a point developed in more detail in acoustic landmark landmark theory (Stevens, 1998). Steriade (1993) provides a typological survey showing the ramifications of this phonetic situation with regard to stops. Transitions between segments are also important in production, since coproduction of adjacent segments can result in substantial modifications of their realizations, including occlusions and gestural reorganizations. Such phenomena are one of the main topics of Articulatory Phonology as developed in Browman and Goldstein (1986, 1992).

The connection made in Steriade's work between phonetic information and phonological sequencing is made in a more broad-based way by the importance of diphone statistics in phonotactics. Any phonological diphone implicitly defines a phonetic transition, which has the potential to be a dynamic cue in speech perception. Some of these transitions are very robust from an encoding standpoint, and others are fragile. Fragility can arise either because the transition lacks perceptual landmarks, or because it is in an unstable region of the phonetic control space and yields excessively variable outcomes. Thus, there is as much reason for a language to selectively utilize the universe of possible transitions as to selectively utilize the phonetic space in its phoneme inventory. The learnability of diphone statistics thus delineates a confluence between the phonetic space, the inventory size, the size of the lexicon, and the complexity of the grammar.

Much recent work in stochastic Optimality Theory takes a different stance on correlations across levels of representation. In Flemming (1995), Kirchner (1997), and Boersma (1998), the response to correlations has been to dissolve the boundaries between levels by treating all effects through a single OT grammar with a single constraint ranking. The Gradual Learning Algorithm of Boersma (1998) and Boersma and Hayes (2001) is the most mathematically elaborated model of this type, and so it is the one I will discussed here.

The conceptual core of stochastic Optimality Theory is morphophonological alternations which are biproducts of general phonological constraints. A landmark paper by Anttila (available in 1994 and published in 1997) undertook to explain variability in the form of the Finnish genitive plural as a response to constraints on metrical prominence, word stress, and metrical alternation, all of which are active in the language generally. Anttila hypothesizes that the variability in outcome for the genitive plural arises because some constraints are unranked. On each individual instance of word production, a specific constraint ranking is imposed through a random selection of ranking for unranked constraints. Depending on which of constraint ranks ahead of the others on that specific trial, one or another of the surface goals takes priority in selecting the form on that particular occasion.

A major extension of this approach is the GLA model of Boersma (1998) and Boersma and Hayes (2001). In this model, constraints are ranked on a real-valued scale. Each individual constraint has a Gaussian distribution of ranking values on this scale. The mean of the distribution is incrementally learned through language exposure, and the variance is taken to be constant. The exact nature of the training algorithm has important repercussions for the overall behavior of the model. After a careful evaluation of statistical robustness and stability, Boersma proposed a training algorithm in which each individual extant word causes downward adjustment of all constraints it violates. The result is that the model is much closer to a standard stochastic grammar than would at first appear; phonological templates which are abundantly instantiated in the training set end up being highly favored by the grammar, and those which are poorly instantiated end up disfavored. Thus,

the rankings of constraints closely track the frequency values that would be assigned to the same constraints in some stochastic grammars.

This model permits much finer tracking of probability distributions for different outcomes that Anttila's model was capable of. Indeed, the model has such general mathematical power that the perceptual classification phenomena introduced in section 2 can be expressed in the model by positing large (in fact, arbitrary large) constraint families which describe arbitrarily fine regions of the parametric space. Fine tuning of constraint rankings has the effect of performing the metric calculations of a standard psychoacoustic model. Conceptually, then, phonetic encoding is raised into the phonological grammar by treating the parametric phonetic space as if it were a very large set of categories.

Morphophonological correspondences are also folded down onto the same constraint ranking. The need to express paradigm uniformity effects is generally accepted in Optimality Theory, and is addressed by Correspondence and Sympathy constraints (see McCarthy and Prince, 1995; McCarthy, 1999). A more challenging case is presented by morphophonological correspondences which are unnatural in the sense of Anderson (1981). A case in point is the alternation of /k/ with /s/ in English word pairs such as *electric, electricity, electricism.* Though this alternation (Velar Softening) has a historical basis in a series of phonetic natural reductions, it is not natural as it stands. If /s/ were in any sense an unmarked pronunciation of /k/ before a schwa, then it should be less frequent than /k/ before schwa generally. According to Celex, however, /s/ is less than half as frequent as /k/ before schwa in the lexicon as a whole. Furthermore, any phonetic generalization which pressured /k/ in the direction of /s/ should also tend to fricate /t/. However, in the cases in which /t/ appears before one of the triggering affixes, it remains a stop: *magnet, magnetism, Jesuit, Jesuitism, mute, mutism.*

The /k/-/s/ alternation is nonetheless productive, as we would predict from statistics over the relevant universe. Celex shows 72 word pairs involving a base form in /k/ with a related form in *-ism* or *-ity.* Velar Softening is found in all of them. Such a perfect regularity is expected to be extended, and a recent pilot experiment indicated that this is the case. Subjects in the experiment were led to believe that the experiment concerned the (semantic) choice amongst various affixes which turn adjectives into abstract nouns. They completed 18 discourse fragments such as the following, involving affixation on a novel stem. Baseline sentences and fillers were also included

5) Janet is criotic about environmental issues. Her ????? manifests itself in avid involvement in environmental groups.

All six subjects softened the /k/ to /s/ in every single case in which they selected the suffix *-ity* or *-ism* over the semantic competitor *-ness.*

This outcome can be captured in the GLA model in the following way. (I am grateful to Paul Boersma (p.c) for suggesting the specifics of this analysis.) A constraint disfavoring /k/ before *-ity* and *-ism* becomes extremely highly ranked as the learner encounters words such as *electricity.* At the same time, universal grammar is presumed to supply a universal set of constraints disfavoring replacement of any phoneme with any other phoneme (e.g disfavoring phonemic changes for a full cross-product of the phonemes). The constraint disfavoring replacement of /k/ with /s/ comes to be ranked low as the learner encounters words in which the replacement has occurred. /t/ is unaffected, since the key constraint targets /k/ only and, not voiceless stops in general. Once the constraint rankings have been learned, the same replacement will occur in any novel form.

Thus, the price of folding unnatural morphophonological correspondences into the phonological grammar is splitting the correspondences into unrelated constraint pairs, which are ranked separately. Probabilities are not directly encoded on correspondences, but rather indirectly impact the state of the grammar through incremental training.

Thus, the GLA model is very powerful. It can encode statistical regularities at all levels, from phonetic encoding up through morphophonological correspondences. For regularities which are either more or less abstract than the level of its core strengths, some researchers might find the encoding to be indirect and inperspicuous. In particular, the treatment of phonetic encoding appears to eschew the well-established resources of mathematical psychology.

In the GLA model, all constraints are at the same level and all constraint rankings are trained on the same data set. Thus, the connection drawn above between the granularity of constraints and the size of the effective training set does not appear to be available. In the model presented above, the probability distributions for different parts of the system are established directly from experience, and thus non-Gaussian distributions will be automatically discovered. Phonetics, in common with many other physical processes, provides examples of skewed distributions relating to physical nonlinearities and saturations. (See e.g. Duarte et al., 2001, who show that the distribution of consonantal interval durations in speech is skewed and obeys a gamma distribution, for many different languages). In contrast, distributions arising from repeated independent decisions (as in coin-flipping, or a forced choice experiment) tend to be Gaussian. Since the assumption of Gaussian distributions is critical to the mathematical tractability of the model, the existence of non-Gaussian distributions appears to be problematic. The GLA model also does not distinguish effects relating to type frequency from effects relating to surface, or token, frequency. It also provides no way to downweight the grammatical impact of extremely frequent words, as Bybee (2001) and Bailey and Hahn (2001) show to be necessary.

In the presentation above, we have suggested that well-formedness judgments represent a decision with weighted inputs from two levels, the lexicon and a score established by the phonological grammar as the likelihood of the best parse. Well-formedness judgments come about differently in the GLA model. Boersma and Hayes (2001) propose that they arise from the likelihood that the given word would emerge as such under repeated runs of the grammar. The word is judged as poor if it would often be modified to some better form on repeated trials. This kind of virtual reality calculation is available in a closed form (e.g. without actually running the grammar many times) because of the simplifying assumptions of the model. The idea is applied with some success in an experiment on morphophonological alternation in Tagalog reported in Zuraw (2000).

This assumption is problematic for phonotactic judgments which have made up most of the literature. Three experiments cited above show high accuracy rates for some sequences which are relatively infrequent and are judged as poor. The transcription data in Hay et al. (in press) found that rates of correction of unusual clusters depended both on the cluster frequency and on the existence of an acoustically similar competitor. Rare clusters without a similar competitor were rated low but not corrected often, even though they were judged as poor. In an imitation experiment, Munson (2001) also found that error rates in adult productions were not significantly different for infrequent and frequent clusters, though frequency did affect wordlikeness judgments. This outcome also occurred in the adult baseline data for Zamuner's (2001) acquisition study. Results such as these follow from the assumption that small phonetic effects can pile up over time in shaping the lexicon, which in turn shapes the grammer. With the lexicon standing in between

the phonetics and the grammar, it is possible for low rates of phonetic instability to coexist with strong lexical statistics. The view of the lexicon presented here also allows lexical neighborhood effects to impact well-formedness judgments. The observed combination of factors is not captured in the GLA model, in which well-formedness judgments are based on the grammar alone.

I have discussed the GLA in this much detail because it is by far the most coherent and comprehensive proposal to fold effects at different levels, from phonetic encoding up through morphophonological correspondences, into a single grammar. Its successes provide further evidence for the importance of stochastic grammars and robust learning algorithms to our understanding of phonology. Its specific weaknesses reveal the general weaknesses of responding to confluences across levels by conflating them.

6 Conclusion

In conclusion, entities at all levels of representation in phonetics and phonology display statistical variation. A wide assortment of behaviors reveals that speakers have implicit knowledge of this variation. It is relevant to speech processing, where it affects perceptual classification as well as speed and accuracy in perception and production. It is also reflected in long term properties of the system, such as allophonic outcomes and the compositionality of complex patterns from subparts.

For any level in the system, we must consider not only its overall probability of occurrence, but also its probability distribution over the less abstract level which gave rise to it. Each category of phonetic encoding has both a total rate of occurrence and a distribution over the parametric phonetic space. The availability of both is an automatic feature of exemplar theory, in which empirical distributions are built up incrementally from experienced tokens of speech. The range and likelihood of various phonetic realizations is revealed by the local density of memory traces on the parametric space, and the frequency of the category as a whole is revealed by the total quantity of traces. Analogous effects are found at higher levels, with word-forms also having probability distributions over phonetic outcomes, and phonological constraints having probability distributions over the space of word-forms.

Comparison of probabilities plays a crucial role in scientific inference and language learning. Both scientists and language learners posit a more complicated grammar only if systematic deviations from the output patterns of a simpler grammar are observable. The level of language exposure thus places bounds on the complexity of inferences that can be made. Children should be able to make gross inferences about the phonological system before they make subtler ones, and even for adults, the subtlety of inferences which are cognitively viable is limited by the size of the data set to which the generalization pertains. In particular, thanks to the tremendous volume of speech which people encounter, fine details of allophony can be learned as well as a large number of word-specific properties. Because of the much smaller size of the lexicon, general knowledge of words is more coarse-grained. Thus, a probabilistic framework allows us to make inferences about the utilization of the phonetic space, and the possible constraint set in phonology, in a way which is not possible in a purely categorical approach.

The starting point of this discussion was the claim that cognitive entities have probabilities and probability distributions and that comparisons of probabilities are involved in comparing alternative models of the whole system. Comparisons of probabilities also play a role in processing, when there

are two alternative analyses of the same form. In phonetic encoding of speech events, each event is categorized by finding the most probable of the competing labels. In parsing speech into words, a relevant comparison is the likelihood that a given sequence is word internal versus the likelihood that it bridges a word boundary. For example, in Hay et al. (in press), the judged well-formedness of the nonsense forms was found to depend on the statistically most likely parse. For a form such as /strɪnpi/, containing a cluster which is impossible within words, the most likely parse includes a word boundary: /strɪn#pi/. The score for such a form reflected the likelihood of the winning parse. In general, this means that in speech processing, the relationship of probabilities will be relevant exactly when there is more than one competing analysis in the cognitive system at the time when the speech processing takes place. The expected values which played a role in the original inferences about the form of the system are not necessarily relevant in the adult system; they may be associated with grammars which were supplanted long ago by a more mature conceptualization.

The phonological system as I have described it exhibits confluences across levels which permit bootstrapping of the system from surface regularities to more abstract ones, and which are implicated in the astonishing speed and accuracy of the adult system. One important case of confluence is the correlation of surface (token) statistics with type statistics, a correlation related to the special status of word-initial position as a locus for maximal contrasts. A second case is the privileged status of diphones with regards to acoustic distinctiveness, coarticulatory control, and complexity of the phonological grammar in relation to the size of the lexicon.

7 Acknowledgments

I am very grateful to Jen Hay, Dan Jurafsky, and Norma Mendoza-Denton for useful discussions which contributed substantially to the structure of this article. Thanks, too, to Stef Jannedy for replotting Peterson and Barney (1952). Reactions from audiences at the LSA, Carry-le-Port, and LCSP are also much appreciated.

Table 1: Lexical neighbors for a nonsense word

ORTHOGRAPHY	IPA
canvle	**kænvəl**
canvas	**kænvəs**
anvil	**ænvəl**
candle	**kændəl**
Campbell	**kæmbəl**

7.1 Tables, con't

Table 2: Existence and absence of triphones in relation to expected counts.

	Absent	Exist
Predicted Absent	42,776	892
Predicted to Exist	2463	4522

Table 3: Existence and absence of diphones in relation to expected counts.

	Absent	Exist
Predicted Absent	48	7
Predicted to Exist	582	732

8 Figure captions and Figures

Note to the typesetter: Figures are provided in sequential order, one panel per page, Note that Figure 4 (4A, 4B, 4C) has three panels. Figure 5 (5A, 5B, 5C) also has three panels.

Figure 1: The F1 F2 vowel space. Data, taken from Peterson and Barney (1952), combine data over children, women, and men. Regions corresponding to each of the various vowels are indicated. Only tokens which were unambiguously identified by listeners are included. In cases of overlap between regions, F3 or some other feature disambiguated the percept.

Figure 2: Classification of an unknown stimulus token, in the framework of exemplar theory. The asterisk indicates the F2 location of the unknown stimulus. The arrows show the window over which the evaluation is made.

Figure 3: Probability distributions for $P(CCC) = 0$ and $P(CCC) = 0.296$, for a sample size of $n = 14$. Distributions computed directly by a binomial calculation.

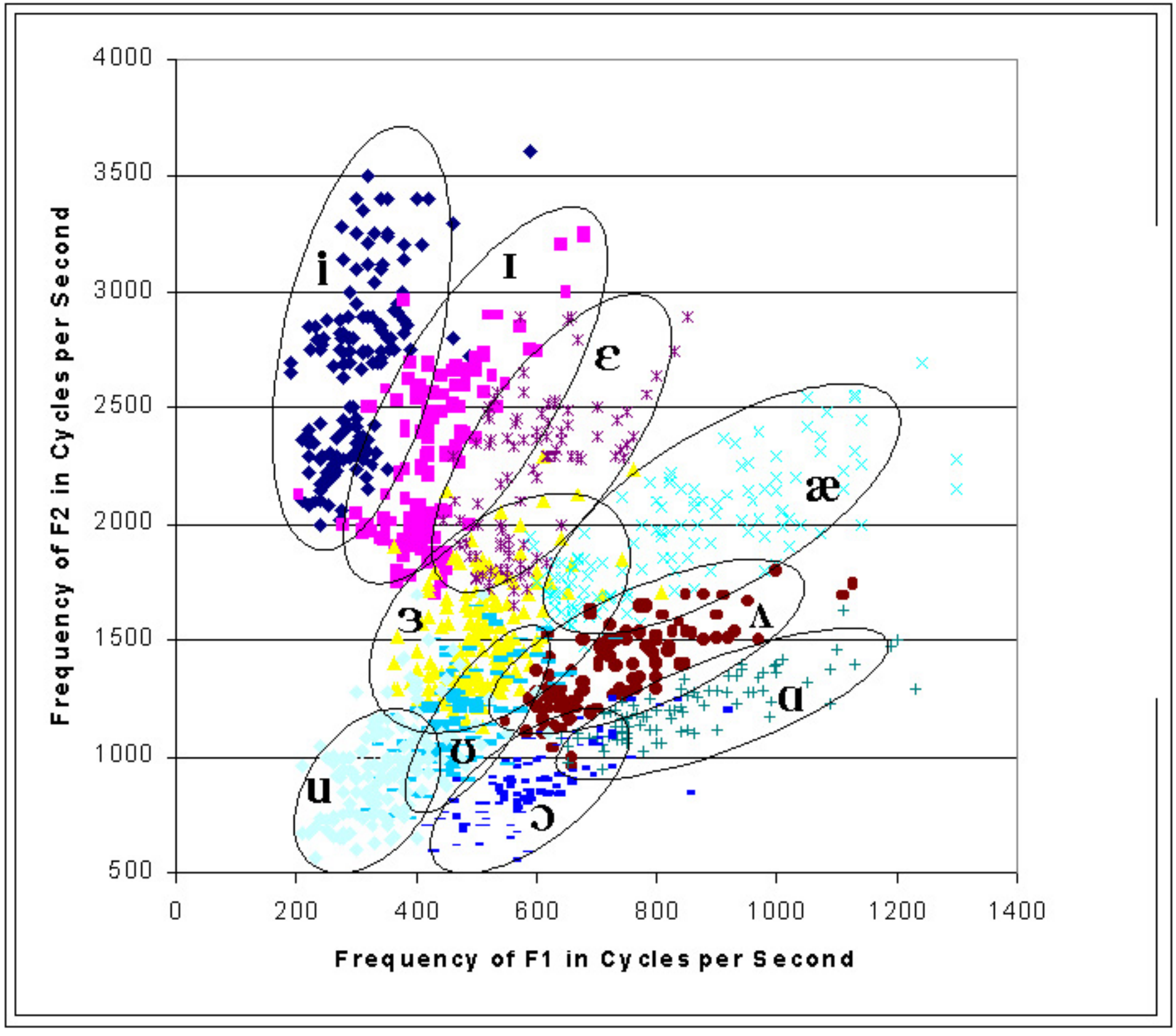
Figure 4: Probability distributions for $P(\#s) = 0.1153$ and $P(s\#) = 0.0763$, for a sample sizes of $n = 14$ (panel A) $n = 140$ (panel B), and $n = 1400$ (panelC). Distributions computed directly by a binomial calculation.

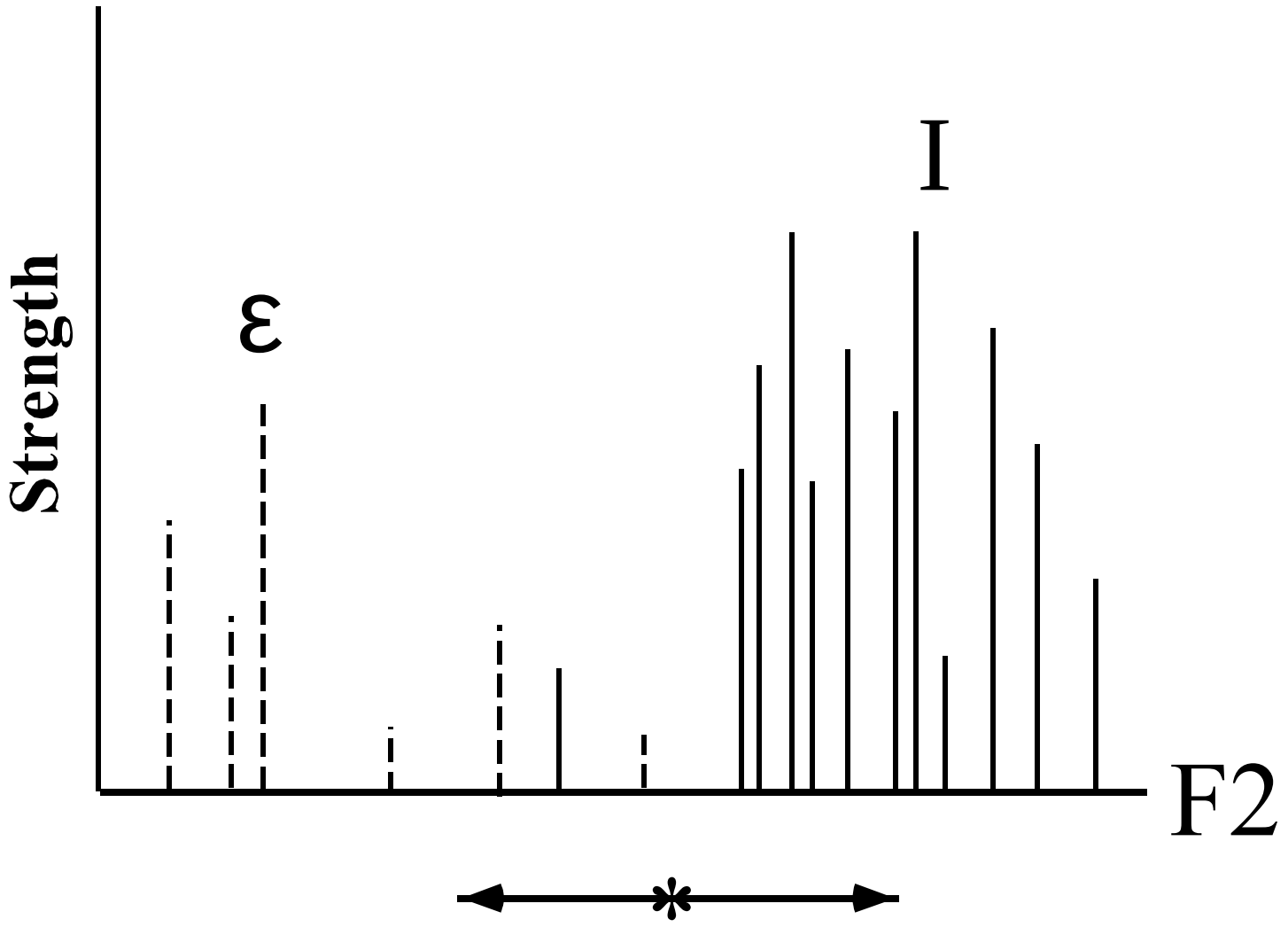
Figure 5: Relationship of two competing categories having idealized Gaussian distributions over the phonetic space. Panel A: 200 Hz separation of means. Low variance leads to little overlap of the distributions. Panel B: 100 Hz separation leads to more overlap. Panel C: Greater variance also leads to more overlap.

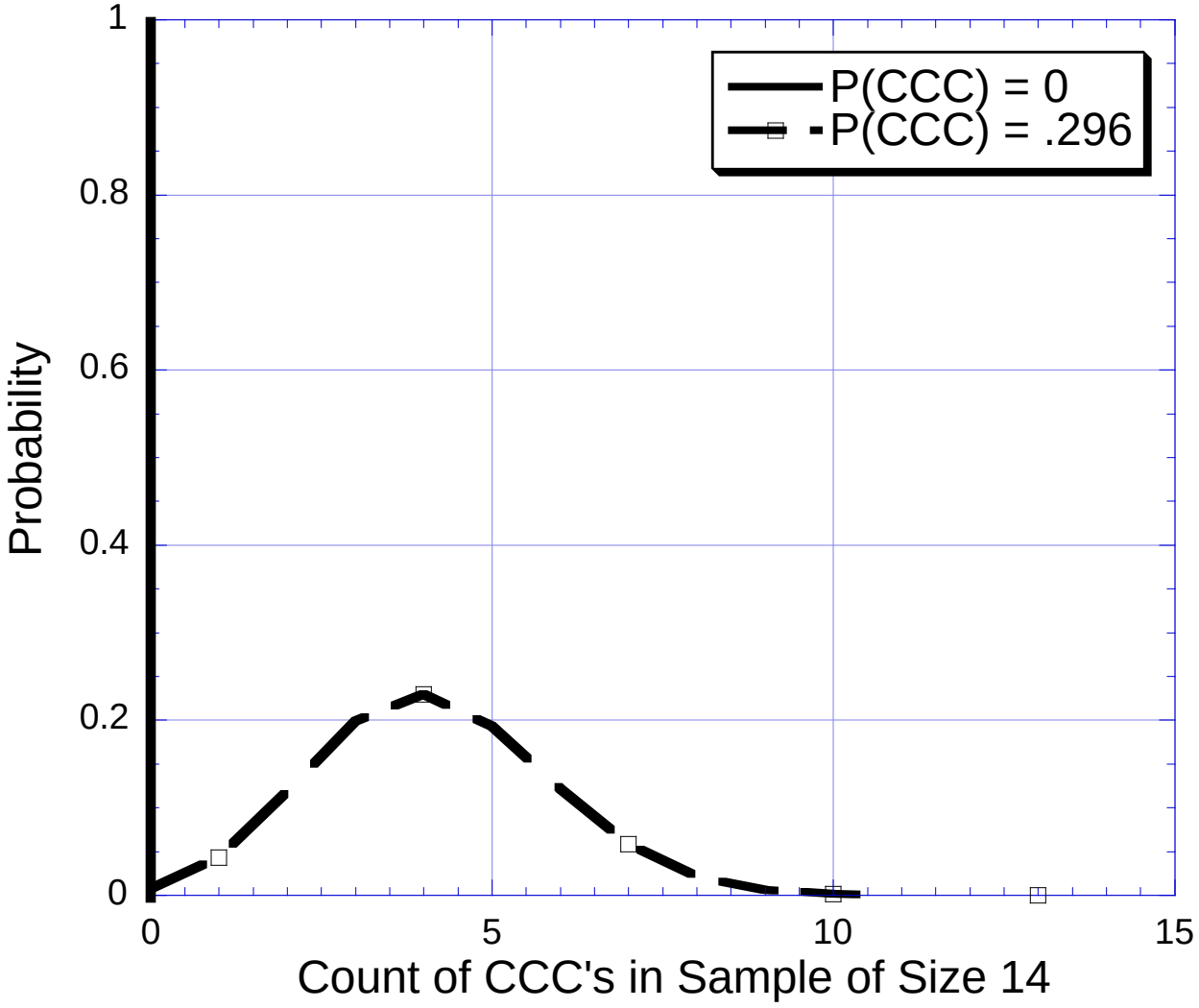
Figure 6: Figure 5C, with the discrimination threshhold indicated.

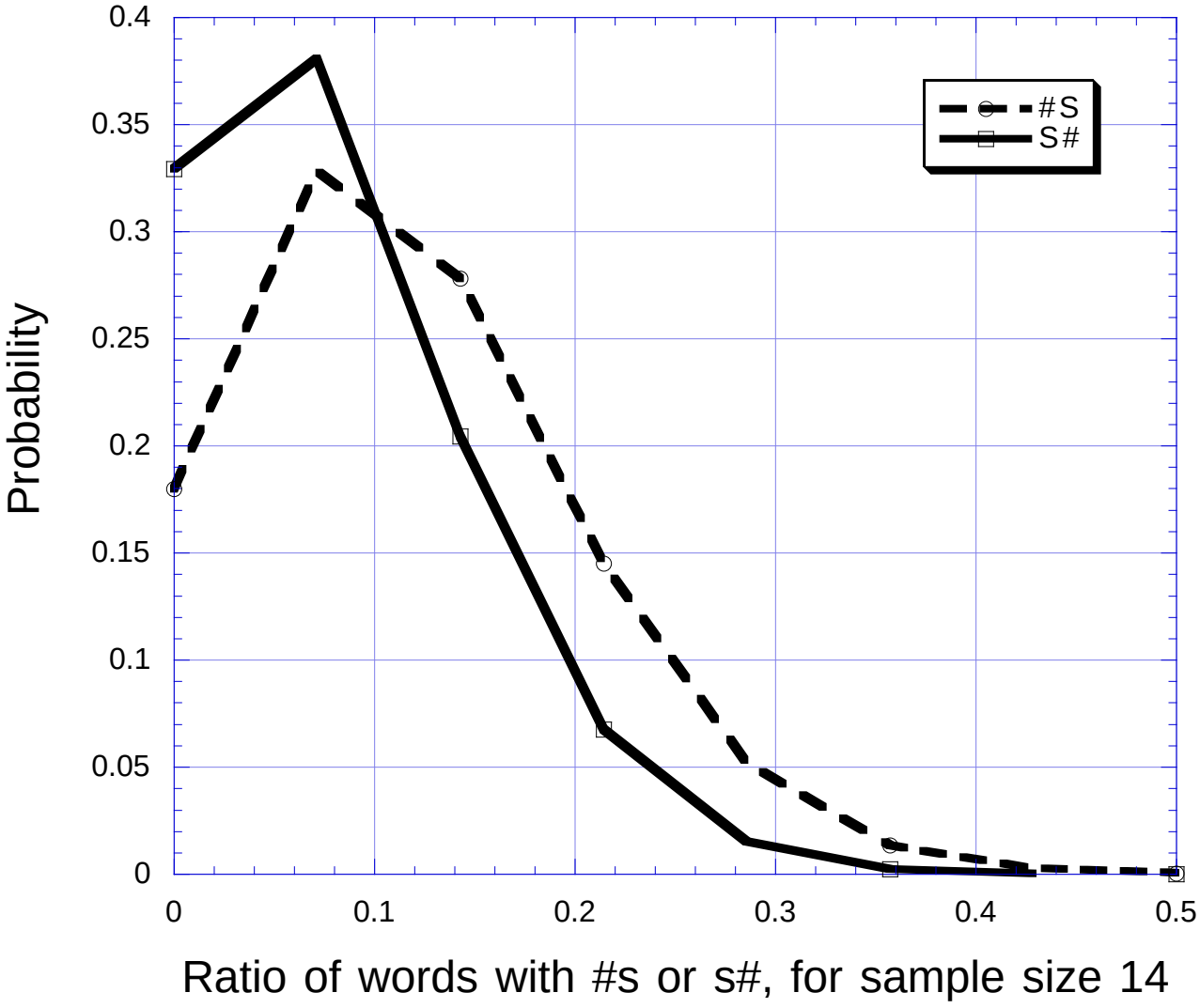
Figure 7: Quartiles of observed counts for triphones, plotted against median expected count.

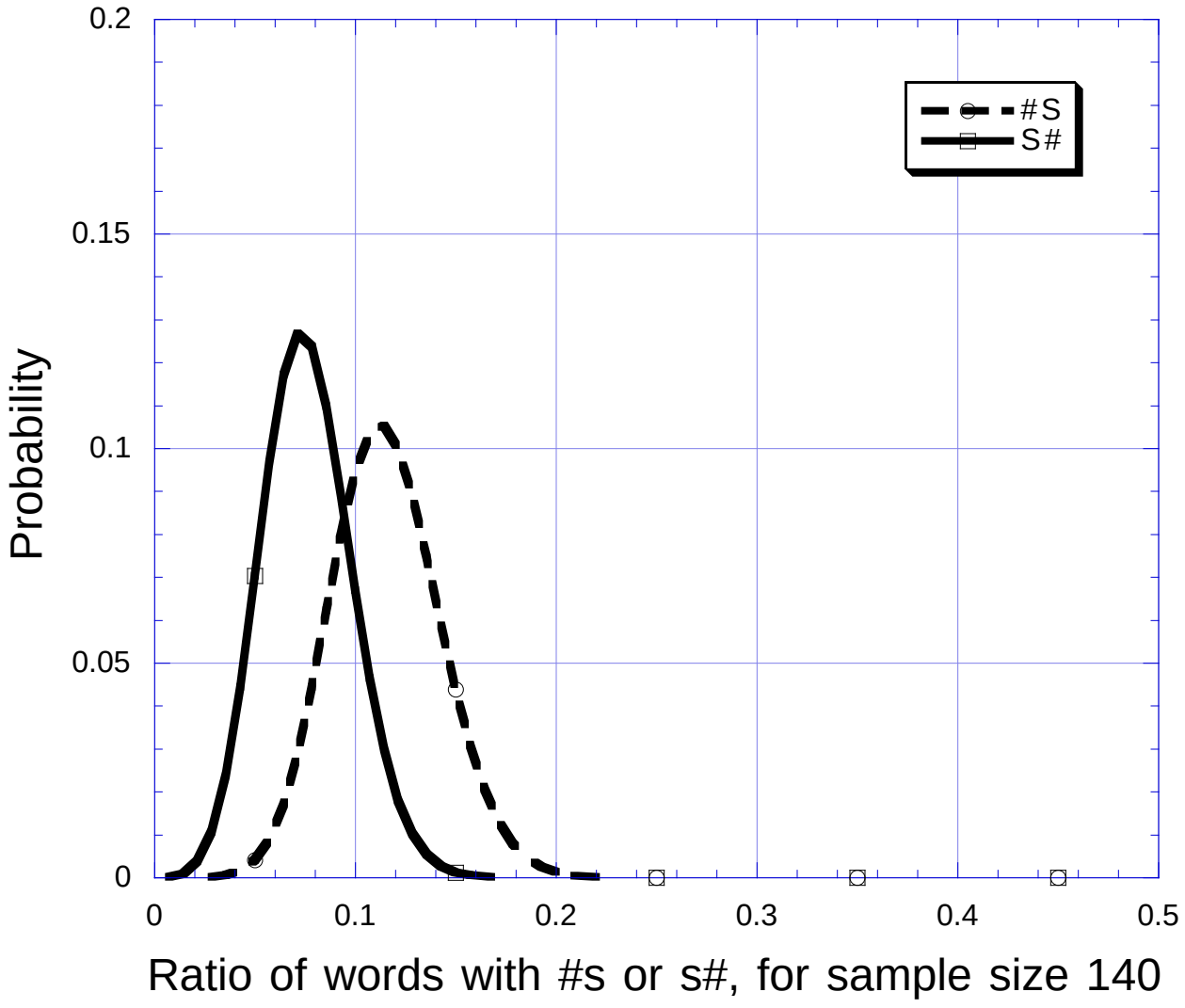
Figure 8: Quartiles of observed counts for diphones, plotted against median expected count.

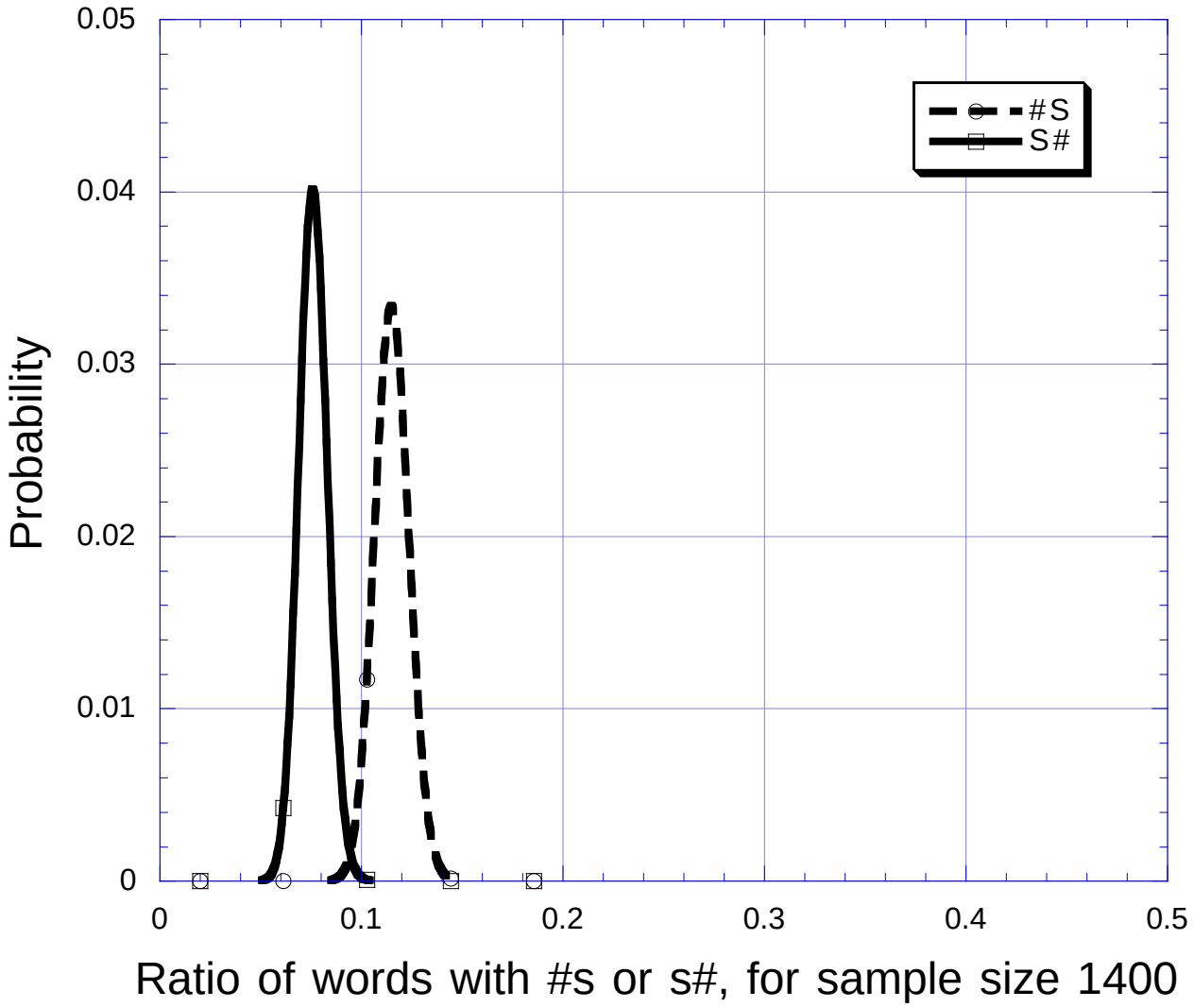

4000
3500
3000
2500
2000
1500
1000
500
0
200
400
600
800
1000
1200
1400
Frequency of F2 in Cycles per Second
Frequency of F1 in Cycles per Second
i
ɪ
ɛ
æ
ɜ
ʌ
ɑ
ʊ
u
ɔ

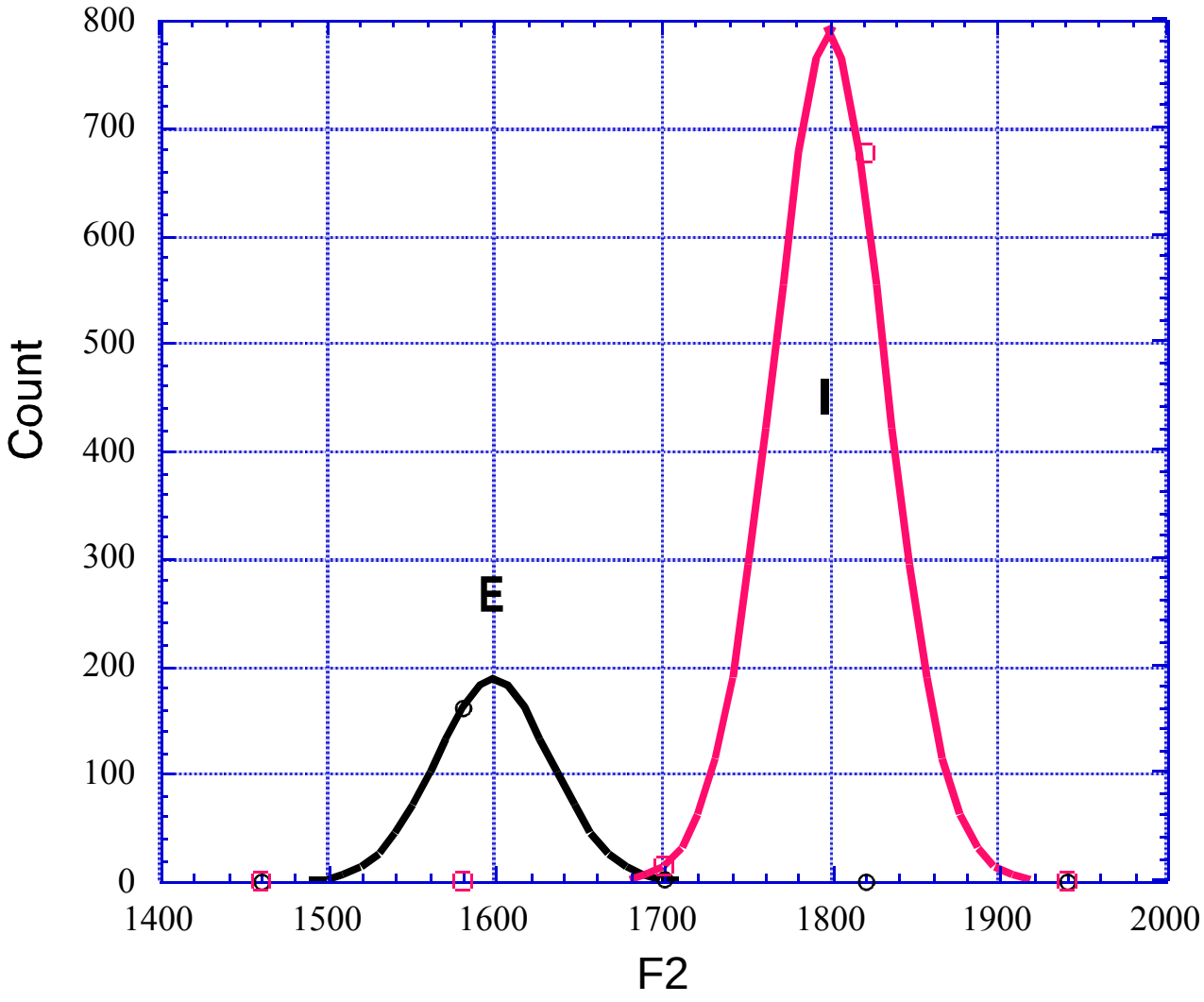

Strength
ε
I
F2
*

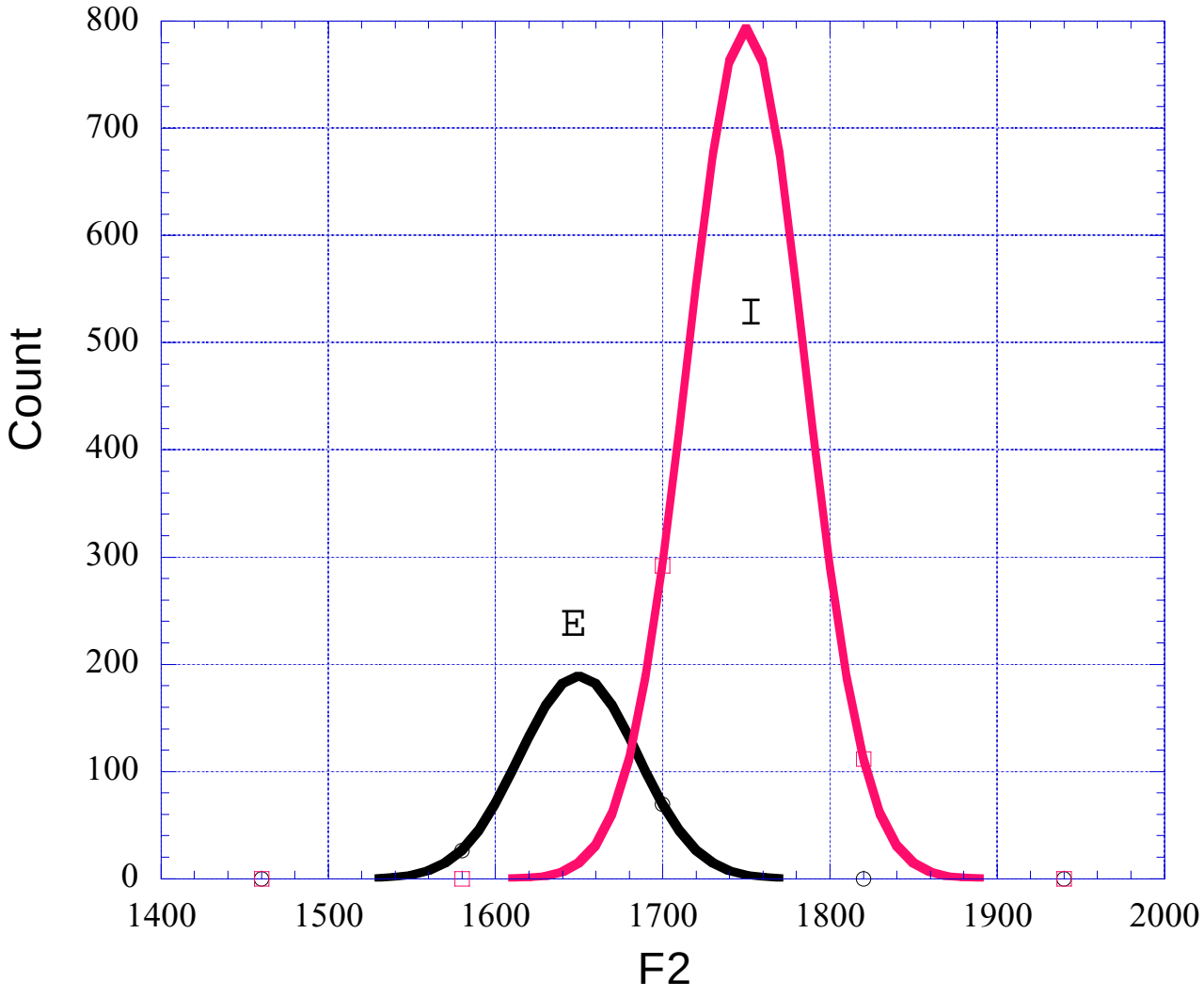

P(CCC) = 0
P(CCC) = .296
Probability
1
0.8
0.6
0.4
0.2
0
0
5
10
15
Count of CCC's in Sample of Size 14

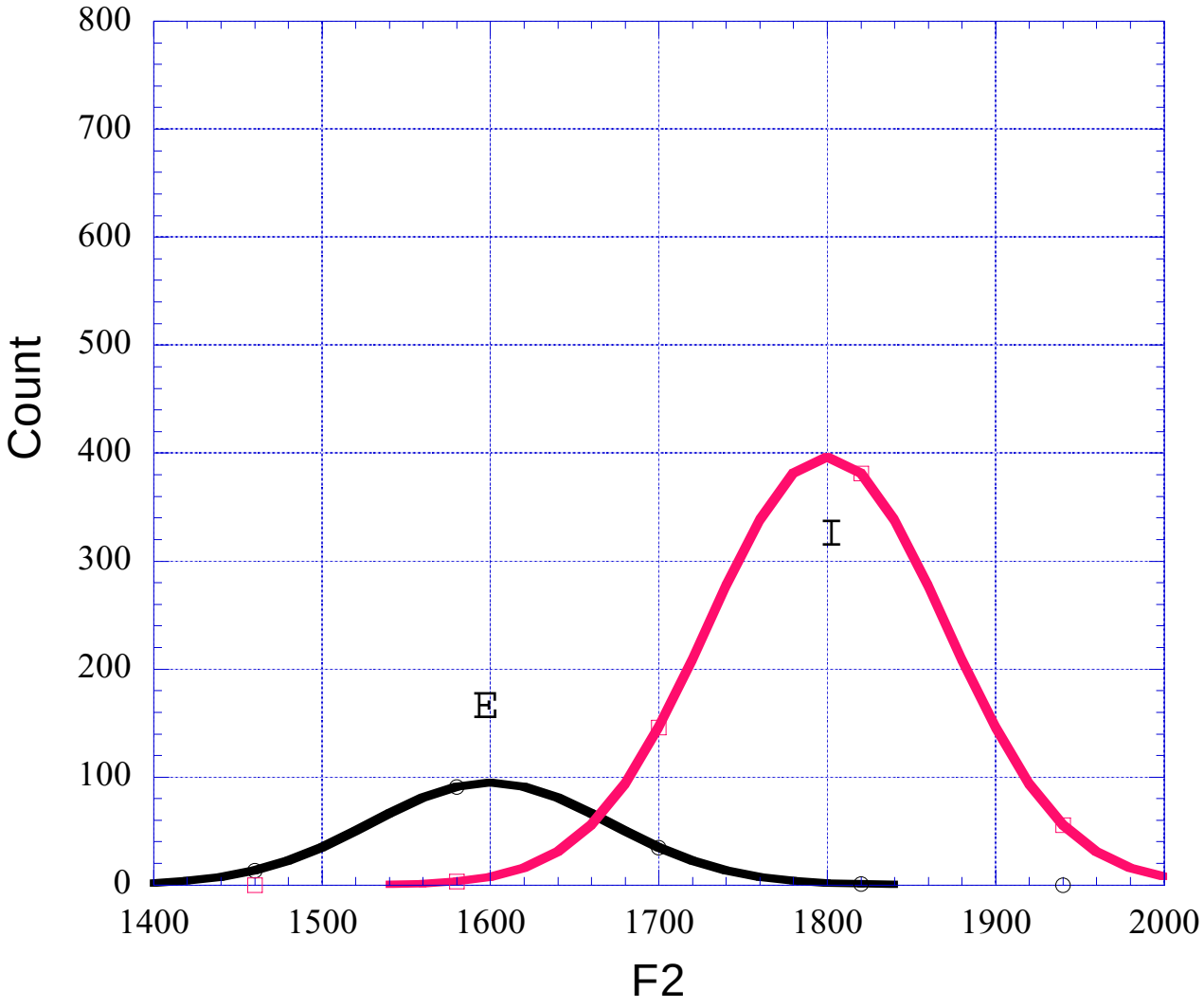

0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Probability
0
0.1
0.2
0.3
0.4
0.5
#S
S#
Ratio of words with #s or s#, for sample size 14

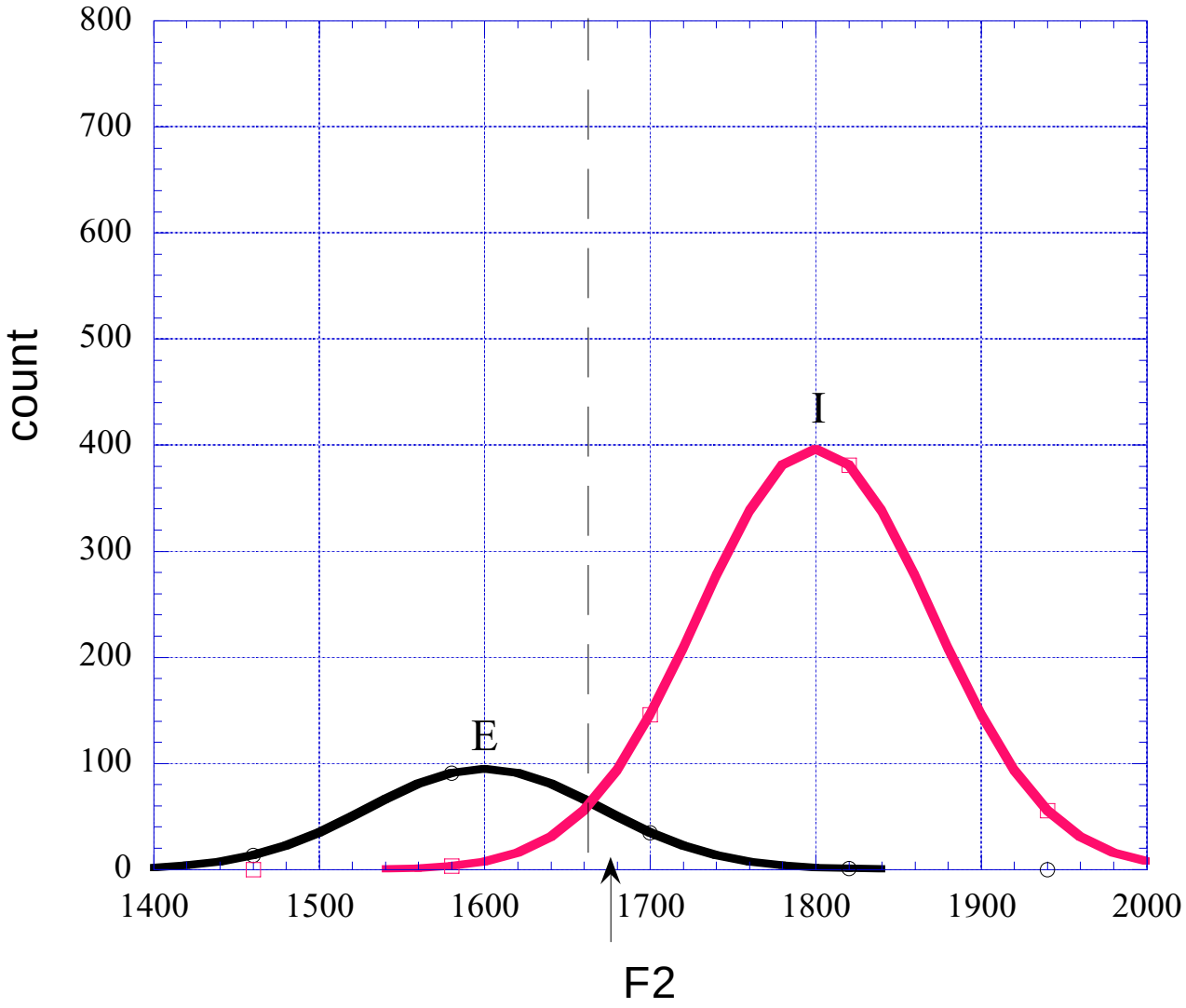

0.2
0.15
0.1
0.05
0
Probability
#S
S#
0
0.1
0.2
0.3
0.4
0.5
Ratio of words with #s or s#, for sample size 140


0.05
0.04
0.03
0.02
0.01
0
Probability
0
0.1
0.2
0.3
0.4
0.5
#S
S#
Ratio of words with #s or s#, for sample size 1400

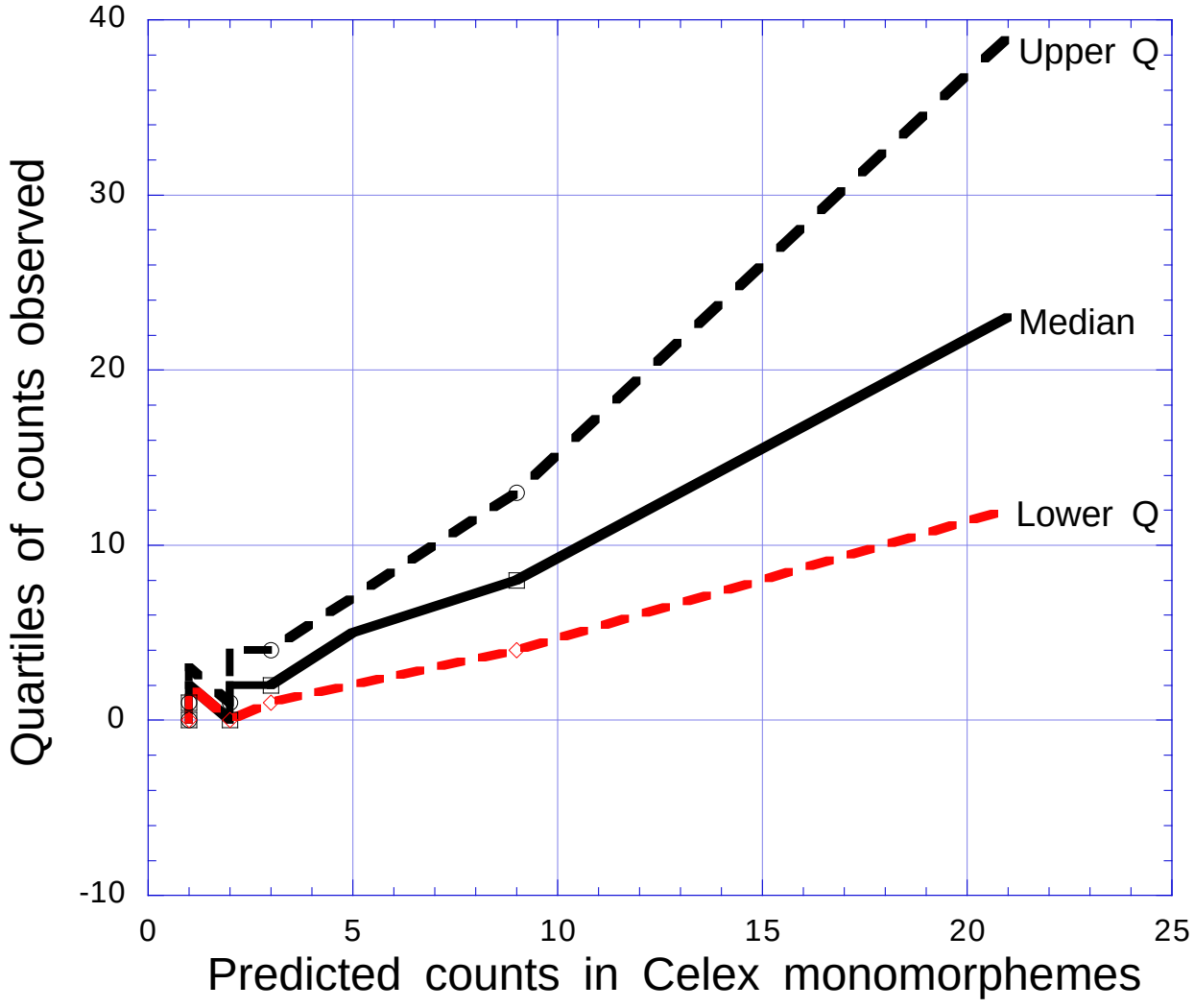

800
700
600
500
400
300
200
100
0
Count
E
I
1400
1500
1600
1700
1800
1900
2000
F2

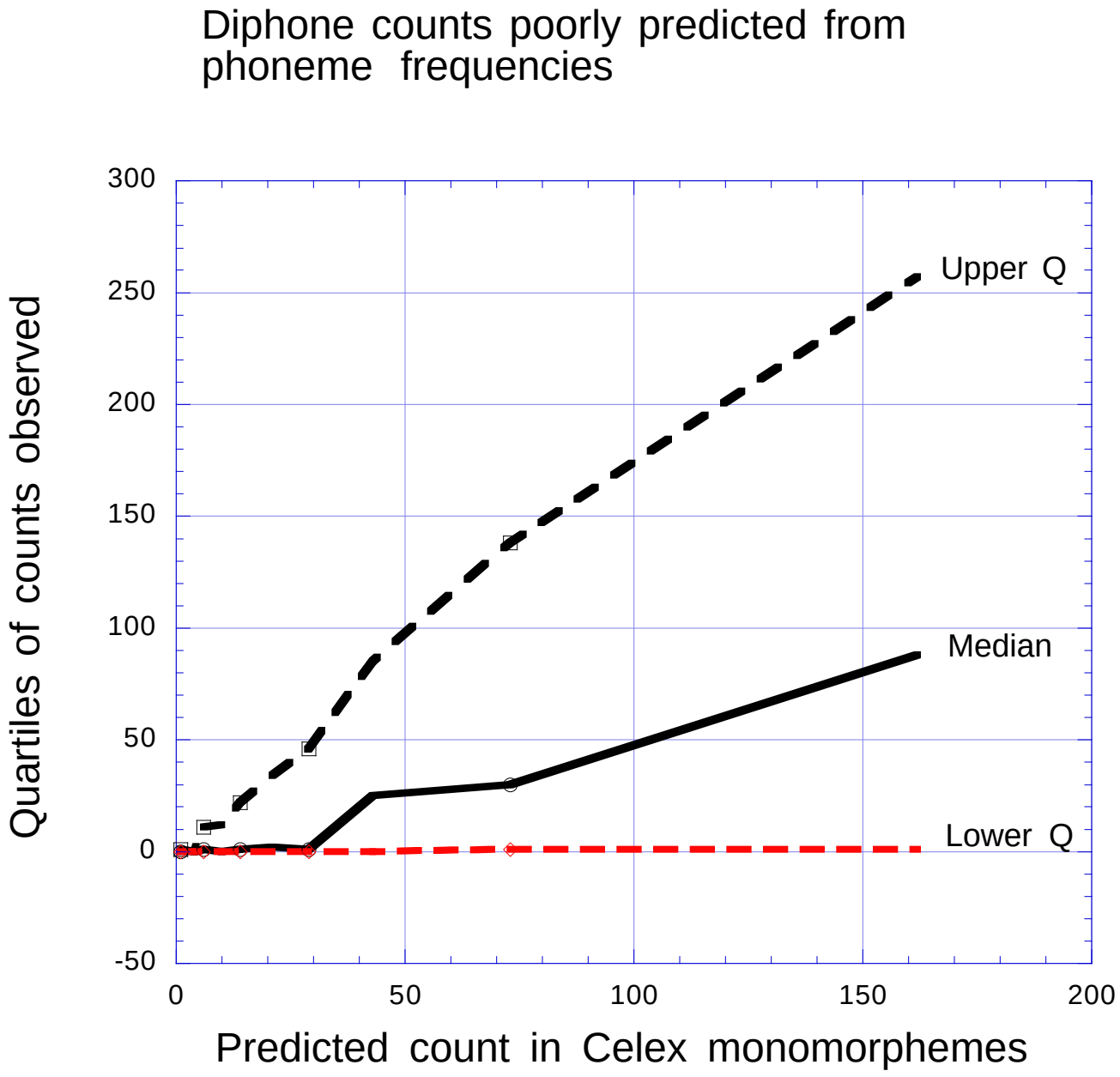

800
700
600
500
400
300
200
100
0
Count
I
E
1400
1500
1600
1700
1800
1900
2000
F2


800
700
600
500
400
300
200
100
0
Count
E
I
1400
1500
1600
1700
1800
1900
2000
F2


800
700
600
500
400
300
200
100
0
count
I
E
1400
1500
1600
1700
1800
1900
2000
F2

Triphone counts as predicted from diphone frequencies




Diphone counts poorly predicted from phoneme frequencies
Quartiles of counts observed
Predicted count in Celex monomorphemes
Upper Q
Median
Lower Q
300
250
200
150
100
50
0
-50
0
50
100
150
200