Stochastic phonology

By Janet B. Pierrehumbert

1. Introduction

In classic generative phonology, linguistic competence in the area of sound structure is modeled by a phonological grammar. The theory takes a grammatical form because it posits an inventory of categories (such as features, phonemes, syllables or feet) and a set of principles which specify the well-formed combinations of these categories. In any particular language, a particular set of principles delineates phonological well-formedness. By comparing phonologies of diverse languages, we can identify commonalities – both in the categories and in the principles for combining them – which suggest the existence of a universal grammar for sound structure.

The classical generative models are non-probabilistic. Any given sequence is either well-formed under a grammar, or it is completely impossible. Under this approach, statistical variation in observed data is viewed as related to variation in performance, rather than illuminating core competence. In contrast, work on sound structure in intellectual circles outside of generative linguistics proper has used probabilistic models for many decades. This line of research has established that the cognitive representation of sound structure is probabilistic, with frequencies playing a crucial role in the acquisition of phonological and phonetic competence, in speech production and perception, and in long-term mental representations.

In this paper, I will first summarize these findings, since these are findings that phonological theory needs to explain. Then I will present some formal ingredients for a stochastic theory of phonology, with the ingredients originating from several different intellectual circles. Lastly, I will summarize some proposals for putting these ingredients together and identify some of the main outstanding issues.

2. External versus cognitive probabilities

In introducing their topic of bioinformatics (applications of computational grammars in DNA sequencing and genetic analysis), Baldi & Brunak (1998) observe that 'A scientific discourse on sequence models – how well they fit the data and how they can be compared with each other – is impossible if the likelihood issue is not addressed honestly.' This observation pertains equally to the empirical study of sound structure. In any given empirical study, we wish to identify the basic units of description (the phonological units analogous to the amino acids of DNA), as well as the grouping and functionality of these units. Data, alas, are fraught with variation due to coding errors, variation amongst speakers, reflexes of undiscovered factors, and so forth. The standard scientific tool for assessing theoretical progress in the face of such variability is probability theory. Lines of research in which a database of any type is first established and then analyzed need probability theory. This is the case for bioinformatics, in which large genome databases are created, often by pooling data from many laboratories, and then analyzed grammatically. It is equally the case for lines of research on sound structure in which corpora are first established, and then analyzed. Thus, the initial thrust of stochastic theory in phonology came from sociolinguistics (in which field recordings are analyzed post-hoc), and psycholinguistics (in which the analysis is responsible for all data collected during an experimental run). Historically, work in generative phonology has emphasized exegesis rather than comprehensive coverage of a corpus. That is, the theory has gradually developed through argumentation about particular phenomena that appear to provide theoretical leverage even if they are rare. Though probabilities may be helpful in this mode of analysis, as we will see below, they are not so clearly inescapable.

The probabilistic reasoning involved in assessing the match between a model and a corpus will not, however, be my main topic here. Such probabilities feature in any mature scientific field and they do not tell us anything about language per se. The recent rise of stochastic phonology stems from a shift in the status of probabilities in the scientific effort on language. Advances in computer power and research methodology over several decades have led to results – initially in sociolinguistics and psycholinguistics – which suggest that cognitive representations are themselves probabilistic. The human language learner, faced with the same variable data that greets the scientist, does not (it would appear) abstract away a purely categorical model. Instead, s/he develops a cognitive system in which frequency information plays a central role. The cognitive system is still grammatical: It establishes the well-

Professor J.B. Pierrehumbert, Linguistics Department, Northwestern University, 2016 Sheridan Road, Northwestern University, Evanston, IL 60208, USA, jbp@northwestern.edu

formedness of complex forms from their subparts, and it has the power to create and to process completely novel forms. However, it is a probabilistic grammar, in the sense that it maintains frequency distributions, and the frequency of any given phonological unit is an important factor in how it behaves in the system.

3. Probabilities over what?

Probabilistic effects have been established at different levels of abstraction in phonology/phonetics. (a) Experimental results using a wide variety of paradigms indicate that people have probabilistic knowledge of the phonetic space as it relates to the phonological categories of their language. (b) They also have probabilistic knowledge of the frequencies with which these categories combine with each other to make up words in the lexicon. Lastly, (c) the relationships of words to each other provide the domain for implicit knowledge of morphophonological alternations. Distinguishing different levels of representation is crucial to our understanding of these results. In consequence, I will not present or discuss any reductionist models – e.g. models which claim that language should be viewed probabilistically **instead of** abstractly. Although many linguists presume that probabilistic models are inherently reductionist, there is widespread agreement amongst experimentalists and computationalists that reductionist models are not viable. In particular, the connectionist program exemplified by McClelland & Elman (1986) has evolved in the direction of models with more articulated levels of abstraction, such as Plaut *et al.* (1996). Dell's (2000) commentary emphasizes that progress in this paradigm depends on the progress of representational and architectural assumptions.

Thus, the cutting edge of research concerns non-reductionist models. In non-reductionist models, a representational framework is developed for each level of abstraction. Frequency distributions are associated with entities or relationships at some or all of these levels. Interactions within and across levels (as specified by the architecture of the model) generate predictions about the space of possible outcomes and the specifics of individual items in that space.

3.1. Probabilities over the phonetic space

The most superficial level of description of speech is that provided by the speech signal itself. The speech signal unfolds in the external physical world and is described using the equations of physics. Our mental representation of sound structure is 'about' speech. If we are talking about dogs, and say 'Dogs are mammals', the sentence is true if the creatures that we designate by the term *dogs* actually are mammals. Similarly, if we say '''pat' begins with /p/'', this is true if the speech events that count as examples of the word *pat* actually do begin with a segment of the abstract type /p/. An immense body of experimental literature (reviewed in Pierrehumbert, 2000 and Pierrehumbert, Beckman & Ladd, 2000) demonstrates that quantitative phonetic details for realizations of phonological units differ from one language to another. A speech signal that constitutes a /p/ in one language may provide an example of /b/ in another. Prototypical articulations and formant values for even the most closely analogous vowels differ from one language to another, as does the allowable range of variation for examples of the same vowel category. Even more tellingly, phonetic interactions differ quantitatively across languages. An example is provided by Flege & Hillenbrand's (1986) study of the production and perception of word-final voiced fricatives in French and English. The interaction of vowel duration and fricative duration as a function of the voicing of a coda fricative differs between the two languages, and listeners show attunement to the patterns of their own language during speech perception.

Establishing mental representations of phonetic distributions, including the contextual factors which play important roles, requires an immense amount of experience and considerable sophistication in encoding this experience. Although initial progress on acquiring these distributions is one of the earliest accomplishments of language acquisition – with considerable progress by 8 months (see review in Vihman, 1996) – adult mastery of allophony, stress/accent, and phonetic precision continues to develop for 6 to 12 year-olds (see Eguchi & Hirsh, 1969; Atkinson-King, 1973; Barton, 1980; Kent & Forner, 1980; Lee, Potamianos & Narayan, 1999; Raimy & Vogel, 2000; Chevrot, Beaud & Varga, 2000). There is also evidence that updating of these probability distributions continues throughout adult life. A striking example of such updating is provided by Harrington, Palethorpe & Watson's (2000) study of 40 years of BBC broadcasts by Her Majesty the Queen. The Queen's pronunciation in these broadcasts has drifted in the direction of Southern Standard British, reflecting social attunement to the speech norms of her younger British subjects.

Thus, a minimal conclusion is that the interface between phonological representations and phonetic outcomes must be modeled using probability distributions over a mental representation of the phonetic space. However, this probabilistic interface does not exhaust the theoretical importance of phonetic distributions. A number of recent studies have brought to light connections between phonetic patterns and various higher levels of representation. An experiment on flapping described in Steriade (2000) found that morphologically related word pairs share optional allophonic variants at far more than chance levels. These data are analyzed using Optimality Theory (OT) and provide evidence for Output–Output correspondence rules stated at the level of the allophone, rather than the phoneme. (Output–

Output Correspondence constraints, introduced in McCarthy & Prince (1995), are the device presently used in OT for enforcing uniformity amongst morphological relatives.) Gussenhoven (2000), also working in OT, reports a direct interaction between qualitative and quantitative constraints on the timing of boundary tones in a dialect of Dutch. Studies summarized in Bybee (2000, 2001) demonstrate a connection between word frequency and lenition, with more frequent words showing systematically higher likelihoods of more reduced pronunciations. For example, in the double-marked past tense verbs (such as 'left', 'felt'), the /t/ is more likely to be omitted in more frequent forms than in less frequent forms. Hay (2000) reports a production experiment on morphologically complex words in which the stem ends in /t/ (such as 'swiftly' and 'listless'). The results demonstrate a gradient effect of the degree of morphological decomposibility on the degree to which the /t/ is pronounced. Studies by Jurafsky, Bell and Girard (in press) also demonstrate effects of contextual predictability on segmental durations. The first two of these studies delineate a connection between phonetic detail and central theoretical issues; for the last three, the patterns have been documented in sufficient detail to plainly suggest probabilistic knowledge over the phonetic space.

3.2. Probabilities over lexical items

Phonological elements are about speech events. Words are made of phonological elements. Phonotactic constraints are about words. If we say that a phonotactic constraint is true of a language, we mean that it characterizes the words of the language. For example, if we say that Hawaiian has only CV syllables, we mean that all words of Hawaiian may be syllabified without recourse to any more complex syllable templates. A behavioral reflex of a phonotactic constraint is judgments about what is a possible word. For example, /mgla/ is judged by English speakers to be impossible, because extant English words contain no initial /mgl/ clusters. However it is a possible (and indeed an existing) word of Russian.

A fairly sizable, and rapidly accumulating, body of experimental literature establishes two major factors in well-formedness judgments of nonwords (or 'wordlikeness judgments', in the psycholinguistics literature.) One, known since Greenberg & Jenkins (1964), is the existence of close lexical neighbors, e.g. actual words which differ in just a few features or phonemes. The other important factor is general knowledge of the lexical statistics of the language. These factors are correlated, because a general pattern is more likely if many words exhibit it. However, they are not perfectly correlated, because a word that is made up of numerous probable subparts may have few close neighbors if the subparts were exhibited in disjoint sets of words.

Results demonstrating the importance of lexical statistics include the following: Treiman *et al.* (2000) show that the frequency of the rhyme in CVC stimuli is reflected in both well-formedness judgments and in decisions on a blending task. Frisch & Zaweydeh (2001) show that speakers of Jordanian Arabic apply general knowledge of lexical statistics in judging novel verbs with varying degrees of OCP violations. (The OCP, or Obligatory Contour Principle, disfavors forms with excessively similar consonants in close proximity.) Bailey & Hahn (1998) find a small but significant effect of general probabilistic knowledge of word form on wordlikeness judgments, when lexical neighborhoods are factored out. These same factors are also important in speech production and perception. Vitevich *et al.* (1997), Vitevich & Luce (1998), Vitevich *et al.* (1999) explore how lexical neighborhoods and phonotactic probability interact. Munson (2000, forthcoming) compares production data in adults and children.

Of particular importance to the theoretical architecture is the existence of cumulative probabilistic effects (phenomena in which the probabilities associated with two different constraints combine to yield the likelihood of the outcome). Hay, Pierrehumbert & Beckman (forthcoming) discuss an experimental study in which transcriptions and ratings of nonsense words containing nasal-obstruent clusters were obtained. They find that well-formedness judgments are gradiently related to the frequency of the cluster and interact cumulatively with an OCP effect on strident coronals. That is, evaluation of a form such as /strɪmsi/ reflects both the frequency of the /ms/ medial cluster, and the dispreference for a word with two strident coronals (here, the two /s/s). Frisch *et al.* (2000) map out the well-formedness of words containing two to four syllables, in which the syllables have either high or low lexical frequencies. The overall well-formedness of the outcome is a cumulative function of the frequencies of the subparts. Disyllabic words with low-frequency subparts are about as well-formed as four-syllable words with high-frequency subparts.

The idea that phonological descriptors – such as onsets, rhymes, syllable contacts, and metrical feet – have associated frequencies provides a number of additional benefits beyond the success in predicting gradient judgments of well-formedness. First, it provides an objective and valid way of assessing whether a gap in the lexical inventory is systematic or accidental. English lacks any words which contain the sequence /lfl/. Is this an accident, or is there a constraint targeting this cluster? Using probabilistic descriptors, it is possible to compute the count of such words we would expect in the lexicon under the scenario in which there is no constraint. Comparing this value (the 'expected value') to the number of examples found, clarifies whether the gap is accidental or systematic (cf. the analysis of triconsonantal clusters developed in Pierrehumbert, 1994). We only need posit a constraint when the absence of a set of forms defies a high expected rate of occurrence. This brings us to a second benefit of probabilistic descrip-

tors – the free ride. As discussed in Pierrehumbert (1994), the phonological grammar can be considerably simplified by assuming that complex patterns with low expected values are not, in fact, expected to occur. The absence of a complex pattern requires no explanation if the expected count is under one. Lastly, comparison of observed counts to counts expected under a null hypothesis permits a formal treatment of soft constraints. Frisch (1996) uses logistic equations to describe the relationship observed in Arabic between phonemic similarity and the statistical strength of the OCP. In this treatment, a hard (or fully grammatical) constraint emerges as the mathematical limit of a soft tendency. The relation between hard and soft constraints is delineated in a way which nonstochastic models cannot capture.

An important controversy in this literature is the issue of type frequency versus token frequency. A phonological pattern has high type frequency if it is instantiated in many different words. It has high token frequencies if it is found frequently in running speech. For example, word-final stressed /gri/ is found in four simplex words of English (*agree, degree, pedigree*, and *filigree*), and hence it has twice the type frequency of word-final stressed /kri/ (found only in *scree* and *decree*). However, the word *agree* is extremely common in running speech, and as a result the token frequency of /gri/# is about 45 times higher than that of /kri/#. If type frequency matters, then constraints are about words and words are about speech events. If token frequency matters, constraints and words are both about speech events – constraints are just more general descriptions of speech events. An experiment discussed in Moreton (1997) on /gri/# and /kri/# is based on the assumption that the token frequency is the relevant one, whereas Pierrehumbert (in press) argues that it is crucial to consider the type frequency.

Untangling this issue is difficult, because type and token frequencies are highly correlated with each other in natural language. This correlation is not mathematically necessary, and the fact that it exists is an important characteristic of language. Study of the outliers of this relationship (namely, high-frequency words with unusual phonological traits) leads to the conclusion that type frequency, at least, is important. Patterns exhibited in just a few words fail to generalize, no matter how high-frequency these words may be (see Bybee, 2001 for a review of findings to this effect). One way of interpreting such findings is that phonological constraints are abstractions, and abstractions are cognitively expensive. Abstraction is motivated when it is needed to handle variability, in the form of diverse and novel incoming forms. However, if enough words exist to motivate projection of an abstraction, the frequencies of these words may contribute to the strength and productivity of this abstraction. Even if type frequency clearly matters, token frequency may also matter.

3.3. Probabilities of relations between words

The generative approach to phonology was launched above all on the strength of morphophonological alternations, such as the vowel shift in *serene, serenity* or the stress shift in *Plato, platonic*. These are relations between words, with highly regular and productive patterns, such as *cat, cats* exhibited in many word pairs and marginal patterns, such as *ring, rang*, exhibited in few pairs. The earliest morphophonological alternations are acquired at approximately age two, i.e. substantially later than the first knowledge of phonetic form (demonstrable from 4 days old) or the first use of word shape (demonstrable in early toddlerhood). The acquisition of morphophonological alternations continues until age 18 at least (see Menn & Stoel-Gammon, 1993; Carroll, 1999). The more irregular and abstract alternations such as the English Vowel Shift are not productive for all speakers (McCawley, 1986). The late acquisition of morphophonological alternations reflects the fact that such alternations must be deduced from word pairs, and the learning of word pairs depends on the learning of words, which in turn depends on phonetic encoding. This perspective is clearly laid out in Bybee (2001), which integrates much previous work in the framework she originated, usage-based phonology. It also plays an important role in OT in the form of Output–Output Correspondence constraints, as discussed above.

Frequency is known to play a role in the cognitive representation of morphophonological relationships. The acquisition of any given rule depends on having a sufficient number of examples (although it is important to note that other factors such as phonological and semantic transparency also play a role, with the result that frequency is not sufficient to predict order of acquisition). Bybee & Pardo (1981) as well as other results reviewed in Bybee (2001) show that adult subjects only generalize patterns to novel forms if their lexicons include a sufficient number of examples. Patterns exhibited only by a few word pairs fail to generalize even if the words in the pairs are extremely frequent. For example, the highly irregular conjugation of the verb *avoir* ('to have', in French) will not generalize to a novel verb. A direct confirmation of this claim is provided by Ohala & Ohala's (1987) study (summarized in Ohala, 1987). In this study, perceived morphological relatedness was operationalized by asking how likely paired words were to have a common historical ancestor. In their comparison of common alternations with isolated patterns (such as *slay/slaughter* and *thumb/thimble*), they found that common alternations were judged as more derivationally related for any given degree of semantic relatedness.

4. Theoretical ingredients

There is no theory at present that provides an integrated treatment of all probabilistic effects in

phonology and phonetics. However, models have been proposed in different subdomains. In some subdomains (such as perceptual categorization), an immense research literature is available. Here I summarize the leading ideas of current models. Then I will move on to some recent ideas to integrate these theoretical ingredients so as to achieve a more comprehensive model which displays the predictiveness of the traditional generative ideal.

4.1. Probabilistic knowledge of phonetics

Implicit knowledge of the quantitative details of pronunciation forms part of linguistic competence by any reasonable definition, since it is fully productive (applying to new word combinations and new words as well as remembered ones) and it develops early and reliably through an apparently innate predisposition to attend to the speech signal. To model such knowledge, the two critical ingredients are a cognitive map and a set of labels. A cognitive map is an analogue representation of physical reality. For example, the lowest level of visual processing encodes the light pattern on the retina onto a sort of mental movie screen. For phonetics, the dimension of the cognitive map are the dimensions of articulatory and acoustic contrast. Part of this map is reflected in the familiar formant space for vowels in which F1 (the first resonance of the vocal tract) is plotted against F2 (the second resonance). The resonances are acoustic correlates of vowel height and frontness. An extremely critical feature, which is exemplified in the formant space, is that the space has an associated metric: it is possible to define degrees of proximity on any particular dimension, or across all dimension. Regions of the cognitive map are associated with labels (more categorical entities on a more abstract level of representation). For example, one region of the F1 – F2 space would be associated with the vowel /i/, and another (possibly overlapping) region would be associated with the vowel /ɪ/. Of course, the labels need not be phonemes, but could be any sort of phonological unit and indeed other units as well.

A gradual shift in phonetic detail – during initial acquisition, a dialect shift, or a historical change–can be readily modelled in a theory which has incremental updating of the probability distribution over the cognitive map which is associated with any given label. For example, children's gradual acquisition of adult levels of phonetic precision can be modelled by assuming that they gradually build up accurate probability distributions for the different phonemes of their language as they occur in context. It cannot be modelled in a 'pegboard' model of phonetic knowledge, in which a universal inventory of phones (such as the elements of the IPA) is available to the phonology. In the pegboard model, each hole either does or does not have a peg in any given language system, and any change must be described as a shift from one hole all the way to another one. If the pegboard model is extended so that it has thousands or millions of pegs, then the models will converge provided that a metric is defined on all dimensions of the pegboard. This line of extension would obviously amount to an admission that a cognitive map is the most superficial level of encoding for sound structure.

Recent papers on exemplar theory (Johnson, 1996; Pierrehumbert, 2001; Kirchner forthcoming) provide formal proposals about how probability distributions over cognitive maps are represented, updated, and used in speech perception and production. Exemplar theory originated in the field of psychology as a schematic account of perceptual classification. (In psychology, Goldinger (1996, 2000) presents a closely related proposal dealing with the memory of particular voices in connection with particular words.) The theory presupposes that extremely detailed memories of experiences are stored, an assumption which has a surprising degree of experimental support. These remembered percepts gradually fill in the region of the cognitive map corresponding to any given categorical label. A label which is encountered frequently will be represented by numerous memories which densely populate the region corresponding to the label. Infrequent categories have a more impoverished representation. The perceptual classification of a new token is accomplished by a statistical choice rule which computes the most probable label, given the location and count of competing distributions in the region of the new token (see Johnson, 1996 and Pierrehumbert, 2001 for equations). This approach is highly successful in capturing a variety of otherwise perplexing findings on speech perception. I therefore assume that it captures schematically some of the main features of the neural mechanisms that are actually used in perception.

In order to bring this approach to bear on linguistic issues, it must be extended to cover speech production. Proposals are provided by Pierrehumbert (2001) and Kirchner (forthcoming). Both proposals depend on the assumption that production is accomplished by activating a subregion of the exemplar space for a category, a claim also advanced in Goldinger (1996, 2000). The aggregate properties of this subregion serve as production goals for articulatory planning. Pierrehumbert (2001) presents calculations showing how a persisent leniting bias in such a model would give rise to Bybee's observations about the relationship of word frequency to the progress of a leniting historical change. She also shows how an unstable category collides and merges with a stable one in a situation where there is a neutralizing pressure on the system. Kirchner discusses how phonologization arises a model of this class.

4.2. Lexical networks

All current models of phonology assume the existence of a mental lexicon, in which the representation of each individual word provides in some fashion a distillation of its various manifestations in various contexts. This assumption is needed to explain why

we can recognize words produced by new speakers, as well as the ability to recognize words whose allophony is influenced by phrasal prosody and sociostylistic register. The nature and abstractness of these word representations differs in different theories. All theories provide the ability to abstract across allophonic variation, but not all theories provide explicit abstract treatment of principles of lexical phonology (e.g. morphophonological rules which apply only to particular word classes or which have idiosyncratic lexical exceptions). Psycholinguistic experiments are in clear agreement that the most irregular morphologically derived forms must be stored as whole words in the mental lexicon. Similarly, some form of abstraction over lexical items – whether explicit or on-line – makes it possible to generate novel forms in the most regular and productive areas of morphology. Controversy focuses on the relationship between the stored lexicon and the grammar.

In connectionist models of speech perception and speech production, the entries in the lexicon are organized in a network. Words with similar properties are linked to each other either directly or indirectly. Types of links include phonological links (e.g. two words share a phonological element, and therefore both have links to a node representing that element), morphological links (e.g. morphologically complex forms are linked to their base form), and syntactic and semantic links (e.g. a word is linked to its hypernym). Spreading activation and mutual inhibition amongst lexical forms in the network explains the time course and outcomes in both speech production and speech perception. In particular, speech perception proceeds incrementally as the speech stream comes in; activation spreads from phonological elements which are discerned in the signal up to all words which exhibit those elements in that order; words compete to be recognized, and a successful candidate inhibits its phonologically similar competitors. Frequency plays a key role in such networks, because nodes or links which are used frequently acquire high resting activation levels. Differential activation levels explain a battery of experimental results on speed, accuracy, priming, and biases in speech processing. This general picture of lexical access is now standard in psycholinguistics, and is found in one form or another in all current models of speech processing (see McClelland & Elman, 1986; Vitevich & Luce, 1998; Norris, 1994; Dell, 2000; Norris, McQueen & Cutler, 2000) There is no competing approach which explains the large experimental literature in this area.

The traditional distinction between competence and performance means that linguists have not always been interested in the experimental results which have motivated the concept of a lexical network. However, a growing body of work demonstrates the implications of the lexical network for traditional concerns of phonology. Bybee (2001) surveys findings on productivity, regularization, and historical change. Dell (2000) and Frisch (1996, 2000) discuss the role of similarity and frequency in phonology. Hay (2000) shows how lexical networks give rise to degrees of morphological relatedness and decomposibility. She also shows how models of morphological processing such as Baayen & Schreuder (1999) give rise to both the trends and the pattern of exceptions in level-ordering of affixes (the tendency to place unproductive and relatively opaque affixes closer to the stem than productive and transparent ones). McClelland & Seidenberg (2000) reiterates the general capability of connectionist networks for capturing gradient productivity and exceptionality, noting that this mechanism is now also adopted by Pinker (1999).

4.3. Stochastic grammars

In the speech engineering and Natural Language Processing literature, the primary tool is the stochastic grammar. The two types of grammars most frequently used in this approach are finite-state grammars and context-free grammars. These are the stochastic versions of the two lowest or simplest types of grammars on the Chomsky hierarchy, and as such they offer very attractive computational properties compared to context-sensitive and transformational grammars. In particular, they are subject to well-defined training algorithms that make it possible to estimate grammar parameters from labeled corpora. In addition, they can be run in either a forward direction (to enumerate the language described by the grammar) or as analyzers (to parse and accept or reject incoming forms). Thus they provide a conceptual baseline for any model relating production to perception, or generation to analysis.

In a stochastic finite-grammar, a set of terminal elements – for example, phonemes – is defined. Probabilities pertain to the transition from one terminal element to the next. This type of grammar is readily conceptualized by imagining a walk through a network of paths, for example in a garden. At each junction of paths, the stroller picks a direction, and the different alternatives may have different degrees of attractiveness and therefore attract different numbers of strollers on the average. An output of such a grammar is a sequence of path segments from the entrance to the exit. Because phonology does not have the level of recursion found in syntax (in particular, there appears to be no evidence for phonological structures with unbounded center-embedding), finite-state models are much more successful in the domain of sound structure than most linguists expect. Their unexpected power arises from two factors. First, the terminal nodes need not be phonemes, but can be formal objects of any type. Hierarchical effects on phoneme licensing and allophony can be handled by using phoneme nodes which are labeled with their prosodic position, such as stressed/unstressed, final/non-final, and so forth. Similarly, the terminal nodes can be set up as correspondences between elements on various autosegmental tiers. Secondly, finite-state grammars can be built up in layers. One

layer can handle large-scale dependencies, with each of its nodes expanded into a grammar on another layer. The power and flexibility of finite-state methods is illustrated in Koskeniemi (1983), Karttunen (1998), as well as the proceedings of the recent SIGPHON conference on Finite State Phonology (Eisner, Karttunen & Thériault, 2000).

In a stochastic context-free grammar, both nonterminal and terminal nodes are defined. The probabilities define the likelihood of alterative expansions of the nonterminal nodes. Coleman (2000) uses this class of grammar to model the stress rules of English.

Work in the framework of data oriented parsing (DOP) provides a perspective on both of these approaches. DOP, a research program in Natural Language Processing (see Bod, 1998) undertakes to train parsers for syntactic and semantic analysis by collating large inventories of syntactic descriptions, together with their frequencies of occurrence, in relevant corpora. Of course any complete parse of a complex utterance in a corpus is likely to be found only once; the workhorse of the theory is the partial or fragmentary tree structures that can be assembled to make complex utterances. The thrust of research is to identify the specific sorts of fragments (including bounds on width and depth) whose frequencies most usefully predict the parses of novel forms. A finite-state grammar can be viewed as a DOP in which sequences of terminal elements are the only descriptions for which frequencies are collated. Similarly, a stochastic context-free grammar corresponds to the decision to collate all tree fragments of depth two. In either case, the elements of the grammar are projected directly from the structures observed in the corpus.

When applied to phonology, this approach provides a very direct interpretation of the fact that phonological grammars track the lexicon. The elements of the grammar are partial descriptions of observed words (either observed in the lexicon, for type frequency, or observed in continuous speech, for token frequency). By definition, these elements correspond formally to the mental representations of words, and their frequencies correspond to how often the patterns are observed in words.

4.4. Variable rules and stochastic grammars

In Chomsky & Halle (1968), regular relationships amongst lexical items are treated through the interaction of underlying representations with transformational rules. The underlying representation of a morpheme distills – sometimes in an abstract and indirect way – the commonalities in its manifestations in different words. The differences amongst these manifestations come about because of transformational rules, which are triggered by some but not all contexts in which the morpheme occurs. For example, the contrast in vowel quality between *serene* and *serenity* comes about because the suffix /ɪti/ provides the context for the rule of Trisyllabic Laxing, a rule which is inapplicable to the base form. In this model, a rule either applies absolutely, or entirely fails to apply, to any given form. Similarly, a given language either does, or does not, have a given rule.

Sociolinguistics developed an extension of this approach in which rules have probabilities rather than applying absolutely. This extension responds to findings that speakers do not always use the same pronunciation of a sound sequence. For example, a speaker of African–American Vernacular English may monophthongize the diphthong /aɪ/ on many, but not all, occasions. Assigning a probability to the monophthongization rule readily describes this fact. Just as in the non-stochastic model, the structural description for the rule is met absolutely, or not at all; however, whenever it is met, there is only a probability that the rule will apply. In some cases in which the structural description is met, the input form is passed on unmodified. It is important to note that such probabilities are established for individual speakers (e.g. they are not artifacts of averaging over a dialectally diverse group). Thus, they represent long-term cognitive properties, and as such are part of the mental representation of language. A standard statistical package, Varbrul, exists for fitting models of this class to data sets, and a large literature in sociolinguistics uses this package. A useful review of the underlying assumptions is provided by Sankoff (1987), and the primary journal is this area is *Language Variation and Change*.

There proves to be fascinating systematicity in the probabilities of various processes. This systematicity shows up with regard to both social and cognitive factors. When an allophonic rule enters a language as a historical change in progress, its rate of application is much higher in some social groups than in others. By comparing the rule probabilities for different groups, we learn something about how social roles and social interactions affect people's mental representations. An example of a morphosyntactic effect of probabilities is provided by Guy's studies of /t/ deletion. Guy (1991a,b) found that rates of /t/ deletion are systematically different in monomorphemic words (such as *past*), double-marked past tenses (such as *left*, past of *leave*), and regular past tenses (such as *passed*). He develops an exegesis of these results using a probabilistic extension of Lexical Phonology (see Kiparsky, 1985). This work represents the epitome of probabilistic derivational models of phonology.

A close relative of probabilistic rules in variationist theory is provided by probabilistic constraint ranking in OT. OT, like the model of Chomsky & Halle (1968), draws a separation between the grammar and the lexicon. The grammar consists of ranked constraints rather than rewrite rules. As in Chomsky & Halle (1968), the lexical representation for any given morpheme distills its manifestations in different words. Qualitatively different outcomes for the same morpheme can occur if a high-ranked constraint invoked by its context in one word results in a variant

of the underlying representation being selected, which is not selected for the morpheme in some other context. If the same surface representation were selected for all contexts in which the morpheme occurred, then an abstract lexical representation which differed from the surface outcome would not survive the acquisition process. Instead, a more transparent form would be selected which emerged unmodified from evaluation by the grammar. This principle provides a broad analogue to the Strict Cycle Condition of Lexical Phonology, the most elaborated derivationalist model.

Anttila (1997) already noted the potential of OT for explaining variable outcomes for the same form. His analysis of the variation in the Finnish genitive plural established the probabilities of different suffix variants for words of various lengths and prosodic structure. The assumption that certain constraints are tied permitted him to model these statistics. He assumed that during the production of any individual word token, two tied constraints A and B are randomly ranked. In some productions, A outranks B whereas in others, B outranks A. If A and B sufficiently highly ranked to be decisive in the outcome, then variation will be observed. Note that, just as in variationist theory, the underlying cause of variation is imputed to the minds of individuals and is an intrinsic part of linguistic competence.

Work by Hayes & MacEachern (1998), Boersma (1998) and Boersma & Hayes (2001) refines and extends this approach by providing each constraint with a probability distribution on a ranking scale. In Hayes and MacEachern, each constraint has a ranking interval, that is, the probability distributions are taken to be rectangular. If the interval has no overlap with the interval for any other constraint, then there is no variability in the way that that constraint interacts with others. The case of complete overlap of the two intervals reduces to the situation Anttila explored. When the overlap is partial between the intervals for constraints A and B, then the probability that A outranks B is not equal to the probability that B outranks A. In Boersma and Hayes, the distributions are Gaussian rather than rectangular. This means that there is always a finite probability that the generally lower ranked constraint will outrank the generally higher ranked constraint on a given trial; however, the Gaussian distribution tails off so fast that this probability can become vanishingly small with respect to any realistically sized corpus. As Boersma (1998) demonstrates, this approach permits fine-grained modeling of variability in outcomes. In addition, he presents a training algorithm under which incremental exposure to linguistic outcomes leads to incremental updating of ranking distributions. This algorithm offers considerable advantages over the Tesar learning algorithm for OT (Tesar & Smolensky, 1998) because it is more robust under variability in linguistic exposure and it behaves gracefully under sporadic exposure to exceptional forms. This kind of robustness is, in fact, characteristic of human learning of language and provides strong evidence for a probabilistic component of the learning model.

Probabilistic OT models have a strong potential for explaining why the lexicon tracks the grammar. In a non-stochastic version of OT, the preference for maintaining the most direct possible correspondence between underlying and surface representations has the consequence that lexical items are encoded as they appear on the surface unless there is reason to do otherwise. In a stochastic version, the same word surfaces in different variants with different probabilities. On the assumption that the dominant variant is internalized by language learners or used to update the lexicons of adult speakers, the end result will be a lexicon which reflects the preferred constraint rankings. To evaluate this suggestion, it will be crucial to carry out full-scale computational modeling of how lexical development proceeds via an OT grammar. At present, OT offers less insight into why the grammar tracks the lexicon. This is because the constraint set is treated in most papers as if it were a priori. Though the original assumption that the constraint set is universal has been conspicuously relaxed in more recent work, in favor of grammars which include idiosyncratic and language-particular generalizations, there is no generally accepted formal mechanism for generating the full set of relevant descriptors, as there is for the stochastic grammars of the previous section.

4.5. Unified theory

The formal ingredients I have just described originate from several different circles. No present theory uses them all in an integrated fashion. However, there is some noteworthy progress in this direction. Here, I provide my own perspective on the basis and direction of this integration.

The most thoroughly supported theoretical ingredients are the lexical network and the cognitive map. Each explains a large and diverse battery of findings about implicit knowledge of speech, and no viable alternative has been proposed for either concept. Thus, the most important area of contention is the architecture of the system in between the cognitive map (representing low-level phonetic encoding) and the lexical network (representing our mental store of words as they relate to each other). Is this system more like a network or more like a grammar? How does it come about that it is attuned both to the nature of the phonetic space and to the nature of the lexical inventory?

An important issue in defining this architecture is the extent and abstractness of pattern generalization. A very significant degree of generalization can be achieved in models such as McClelland *et al.* (1986) by assuming that a novel incoming signal activates the entire group of words which are highly similar to it, and that the results (of whatever kind) represent the aggregate nature of these activated words (see also Seidenberg, 1997, for discussion of this point). However, aggregating over word groups does not

reproduce in full the effect of a constraint containing a variable whose domain is an abstract type. For example, as discussed in Marcus (1998), McClelland *et al.* (1986) does not generalize a trochaic foot pattern to words which are longer than those in the training set. To achieve this generalization, it is necessary to quantify over feet in a template specifying 'any number of feet'. Similarly, an associative network can implicitly extend the obligatory contour principle (OCP) effect, which favors combinations of dissimilar and nonhomorganic consonants, to many new words exhibiting attested combinations of consonants. However, as explained in Berent, Everett & Shimron (2000) and Zuraw (2000), it will fail to abstract across place and similarity generally so as to properly admit all novel solutions to the satisfaction of this constraint. Results such as these indicate that the model needs schemas containing abstract variables. Schemas with abstract variables are a shared feature of usage-based phonology (as discussed in Bybee, 2000, 2001) as well as approaches closer to the generative tradition (e.g. Marcus, 1998a,b; Pinker, 1999).

A tension can be identified in the literature between people proposing stochastic grammars (generally from a background of Natural Language Processing) and people proposing stochastic versions of OT (generally from a generative linguistics background). These approaches are not as divergent as would at first appear. As shown in Karttunen (1998) and papers in Eisner *et al.* (2000), stochastic OT models are actually equivalent to finite-state models given some reasonable restrictions on the constraint sets. This equivalence does not appear to obtain for some of the more radical innovations in OT. In particular, it is problematic if continuously-valued phonetic goals are interspersed with qualitative ones (as in Gussenhoven, 2000), or if meta-level constraints are interspersed with other constraints. Meta-level constraints are ones which refer to entities which are the not primitives in the descriptive language, but rather outcomes of other constraints. For example, an OCP constraint is a meta-level constraint in a model which generates phonemes epiphenomenally from the interaction of phonetic functions. Thus, future progress will depend on comprehensive and exact assessment of what meta-level constraints are needed in phonology and how they interact with other constraints.

I have also said that stochastic grammars are quite good at capturing the way that grammars track the state of the lexicon. Stochastic OT shows considerable promise in explaining why the lexicon reflects the grammar. The actual state of affairs is that the grammar and the lexicon are attuned to each other. It is important not to get stuck on a chicken-and-egg question ('Which came first? The lexicon or the grammar?'). Chickens come from eggs, and eggs come from chickens. Analogously, the grammar is acquired through experience with the lexicon and items in the lexicon are acquired via the grammar, as it acts in speech perception and production. Thus, the ultimate answer to the issue of how the grammar and the lexicon are related will come from modeling the equilibrium state of the production-perception loop. Looking at the joint state of the grammar and the lexicon as they stabilize over many instances of production and perception promises to reveal how they are attuned to each other.

Only a few papers now explore the end result for the linguistic system of a production/perception loop. Pierrehumbert (2001) and Kirchner (forthcoming) show how production/perception loops play out in exemplar theory in relation to allophony and phoneme licensing. The treatment of phonological grammar in both of these papers is extremely sketchy. Hay (2000) presents some consequences of the production/perception loop as it plays out in a morphological processing model with whole word and decompositional access to complex forms. The main prediction is that semi-decomposed forms can exist, but that they tend to evolve towards the extremes of the system (e.g. to become either non-decomposed, or fully decomposed). This provides a more abstract counterpart to the phonologization discussed by Kirchner, and shows how iteration can produce sharpening of what would otherwise be a soft tendency. Lastly Zuraw (2000) makes a significant extension of Boersma & Hayes (2001) by adding a perception-lexicalization component which uses Bayesian inference to probabilistically estimate underlying representations of novel words. She applies this model to the issue of vocabulary evolution in Tagalog, showing how a relatively weak phonological constraint ends up impacting lexical statistics.

Zuraw's data concern a semi-productive morphophonological rule of Tagalog by which a nasal consonant in a prefix coalesces with the first consonant of the stem. Over the lexicon as a whole, this rule is probabilistically dependent on voicing and place; however any given extant word either does or does not display the rule. By introducing new words in context, Zuraw obtained rates of rule application for novel stems, and she also obtained well-formedness judgments for target words presented as relatives of the base form. The same approach is also applied to an alternation involving vowel height.

Recall that in models of phonotactic well-formedness built on stochastic grammars, the likelihood of a novel form as determined from its subparts directly predicts its well-formedness score. In Zuraw's model, in contrast, well-formedness judgments come about in a different way, for two reasons. First, the data to be explained are word relationships (morphologically complex forms presented together with the putative base), rather than simplex words in isolation. Second, she is working with a model in which constraint rankings have probabilities, and not constraints themselves. Following Boersma & Hayes (2001), Zuraw proposes that the well-formedness of a novel form is judged by carrying out a kind of virtual calculation about how it would come out over numerous productions using a stochastic OT grammar. If the

form always comes out exactly as is, then it is rated highly, whereas if some productions revise it to a less marked form, then it is rated somewhat lower. It is rated very low if it is almost always revised to something else.

The primary challenge for an OT treatment of gradient well-formedness is raised in Berkley (1994): Novel forms which violate a probabilistic constraint are judged to be marginal, whereas extant forms violating the same constraint are rather stable in the vocabularies of adult speakers. For example, English speakers feel that morphological neologisms such as *distinctity* (containing an OCP violation) are rather poor, but the two exceptional forms which contain this same configuration, *chastity* and *sanctity*, show little tendency to alternate with less marked forms. Zuraw (2000) addresses this problem with a proposal about the ranking of Input–Output Correspondence constraints (constraints which favor identity between the underlying form of an morpheme and its surface manifestation; see McCarthy & Prince, 1995) and a new constraint USELISTED (an OT implementation of classic observations about morphological blocking, by which a lexicalized complex form takes priority over one which is productively formed on the fly). She proposes that such constraints become more and more highly ranked during the course of language acquisition, so that for adults producing known words, there is practically no alternation. Instability in the production of novel forms – and the impact on well-formedness judgments which is thereby entailed in her framework – arises from probabilistic variation while first constructing the mental representations of the forms.

Clearly, this proposal makes strong predictions about the course of language acquisition, and these predictions need to be verified through empirical studies. It also makes predictions about results for simplex novel forms, such as the stimulus sets for Hay *et al*. (forthcoming) and other experiments reviewed in Section 3.2. For simplex forms, there is no effective competition between Input–Output Correspondence and USELISTED since there is no question of a phonologically opaque morphological decomposition. To generate the probabilistic alternations which provide an underpinning for well-formedness judgments in this framework, it would be necessary to identify some other constraint class which stands in an unstable relationship to USELISTED. The most likely one is markedness constraints. That is, the approach leads us to expect that novel simplex forms would be judged as marginal insofar as they tend to probabilistically alternate with less marked forms when they are first being added to the lexicon.

It is difficult to find studies which address this point. The transcription data in Hay *et al*. (forthcoming) show an extremely imperfect correlation between the rating of a cluster and the tendency to misperceive it as some other cluster. A rather poor cluster may still be reliably transcribed if the system includes no acoustically similar competitors. One can imagine that it would be accurately reproduced in speech as well as in writing. Closer to the mark is the study discussed in Munson (2000), in which adults and children imitated novel words rather than transcribing them. Munson reports a statistically significant effect of phonotactic likelihood on well-formedness. However, for adult speakers the error rate in production was not significantly different for the high frequency and low frequency clusters in the study. Although there were somewhat more errors for the low frequency clusters, the error rate does not appear sufficient to explain the rating data. If such findings are replicated in more extensive experiments, then probabilistic alternations cannot be accepted as the sole source of gradient well-formedness.

5. Conclusion

In summary, probabilistic effects have been identified at all levels of representation. The main tools for capturing such effects are cognitive maps, lexical networks, and stochastic grammars or stochastic constraint ranking systems. It is clear that the phonological theory of the future will have lexical networks (in some form), cognitive maps (in some form), and an architecture for connecting them which includes the power to state patterns involving abstract variables. In order to sort out the open questions about this architecture, a key issue is modeling the perception/production loop. Such modeling critically involves probabilities, since it involves incremental learning over extensive and variable experience.

A Stochastic Phonology Bibliography

ANTTILA, A.T. (1997) 'Deriving variation from grammar', in F. HINSKENS, R. VAN HOUT & L. WETZELS (eds) *Variation, Change, and Phonological Theory*, 35–68. Amsterdam: John Benjamins Publishing Co., (Downloadable from the Rutgers Optimality Archive, ROA 63-0000).

ATKINSON-KING, K. (1973) Children's acquisition of phonological stress contrasts. *UCLA Working Papers in Phonetics*, no. 25.

BAAYEN, H. & SCHREUDER, R. (1999) War and peace: morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain and Language* 68, 27–32.

BAILEY, T.M. & HAHN, U. (1998) 'Determinants of wordlikeness', in M.A. GERNSBACHER, & S.J. DERRY (eds) *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, 90–95. Mahwah, NJ: Lawrence Erlbaum.

BALDI, P. & BRUNAK, S. (1998) *Bioinformatics: the Machine Learning Approach*. Cambridge, MA: MIT Press.

BARTON, D. (1980) 'Phonemic perception in children', in G.H. YENI-KOMSHIAN, J.F. KAVANAGH & C.A. FERGUSON (eds) *Child Phonology, Vol. 2, Perception*, 97–116. New York: Academic Press.

BERENT, I., EVERETT, D.L. & SHIMRON, J. (2000) Do phonological representations specify variables? Evidence from the obligatory contour principle. *Cognitive Psychology* 42, 1–60.

BERKLEY, D.M. (1994) The OCP and gradient data. *Studies in the Linguistic Sciences* 24/2, 59–72.

BOD, R. (1998) *Beyond Grammar: an Experience-Based Theory of Language*. Cambridge UK: CSLI Publications/Cambridge University Press.

BOERSMA, P. (1998) *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD Dissertation, Amsterdam: University of Amsterdam. (Downloadable from http://fonsg3.let.uva.nl/paul/).

BOERSMA, P. & HAYES, B. (2001) Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32, 45–86. (Downloadable from http://fonsg3.let.uva.nl/paul/).

BROE, M. & PIERREHUMBERT, J. (eds) (2000) *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge, UK: Cambridge University Press.

BYBEE, J. (2000) 'Lexicalization of sound change and alternating environments', in: BROE & PIERREHUMBERT (2000) pp 250–269.

BYBEE, J. (2000) 'The phonology of the lexicon: evidence from lexical diffusion', in M. BARLOW & S. KEMMER (eds) *Usage-Based Models of Language*. Palo Alto: CSLI Publications.

BYBEE, J. (2001) *Phonology and Language Use* (Cambridge Studies in Linguistics, 94). Cambridge UK: Cambridge University press.

BYBEE, J. & PARDO, E. (1981) Morphological and lexical conditioning of rules: Experimental evidence from Spanish. *Linguistics* 19, 937–968.

CARROLL, D.W. (1999) *Psychology of Language*. Pacific Grove, CA: Brooks/Cole Publishing Co.

CHEVROT, J.-P., BEAUD, L. & VARGA, R. (2000) Developmental data on a French sociolinguistic variable: the word-final post-consonantal /R/. *Language Variation and Change* 12/3, 295–319.

CHOMSKY, N. & HALLE, M. (1968) *The Sound Pattern of English*. New York: Harper & Row.

COLEMAN, J.S. (2000) Candidate selection. *Linguistic Review* 17, 167–179.

DELL, G. (2000) 'Commentary: Counting, connectionism, and lexical representation', in BROE & PIERREHUMBERT (2000) pp 335–348.

EGUCHI, S. & HIRSH, I. (1969) Development of speech sounds in children. *Acta Octo-Laryngologica Suppl* 257.

EISNER, J., KARTTUNEN, L. & THÉRIAULT, A., (eds) (2000) SIGPHON 2000 Finite-State Phonology: *Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology. 6 August, 2000 at the Centre Universitaire, Luxembourg*. (Downloadable from http://www.cogsci.edsac.uk/sigphon/CPpapersSP.html).

FLEGE, J.E. & HILLENBRAND, J. (1986) Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *Journal of the Acoustical Society of America* 79/2, 508–517.

FRISCH, S.A. (1996) *Similarity and frequency in phonology*. PhD Dissertation. Evanston, IL: Northwestern University.

FRISCH, S.A. (2000) 'Temporally organized lexical representations as phonological units', in BROE & PIERREHUMBERT (2000) pp 283–298.

FRISCH, S.A., LARGE, N.R. & PISONI, D.B. (2000) Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42, 481–496.

FRISCH, S.A. & ZAWAYDEH, B.A. (2001) The psychological reality of OCP-Place in Arabic. *Language*, March, 91–106.

GOLDINGER, S.D. (1996) Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 1166–1183.

GOLDINGER, S.D. (2000) 'The role of perceptual episodes in lexical processing', in A. Cutler, J.M. McQueen & R. Zondervan (eds) *Proceedings of SWAP (Spoken Word Access Processes)*, 155–159. Nijmegen: Max Planck Institute for Psycholinguistics.

GREENBERG, J.H. & JENKINS, J.J. (1964) Studies in the psychological correlates of the sound system of American English. *Word* 20, 157–177.

GUSSENHOVEN, C. (2000) 'The boundary tones are coming: on the nonperipheral realization of boundary tones', in BROE & PIERREHUMBERT (2000) pp 132–151.

GUY, G.R. (1991a) Contextual conditioning in variable lexical phonology. *Language Variation and Change* 3, 223–239.

GUY, G.R. (1991b) Explanation in variable phonology. *Language Variation and Change* 3, 1–22.

HARRINGTON, J., PALETHORPE, S. & WATSON, C.I. (2000) Does the Queen speak the Queen's English? *Nature* 408, 927–928.

HAY, J.B. (2000) *Causes and consequences of word structure*. PhD Dissertation. Evanston, IL: Northwestern University.

HAY, J.B., PIERREHUMBERT, J. & BECKMAN M.E. (forthcoming) 'Speech Perception, Well-Formedness, and the Statistics of the Lexicon', in R. OGDEN, J. LOCAL & R. TEMPLE (eds) *Papers in Laboratory Phonology VI*. Cambridge, UK: Cambridge University Press (Downloadable from. http://www.ling.nwu.edu/ jbp/publications.html).

HAYES, B. & MACEACHERN, P. (1998) Quatrain form in English fold verse. *Language* 74, 473–507.

JOHNSON, K. (1996) 'Speech perception without speaker normalization', in K. JOHNSON & J. MULLENNIX (eds) *Talker Variability in Speech Processing*. San Diego: Academic Press.

JURAFSKY, D., BELL, A. & GIRARD, C. (in press) 'The role of the lemma in form variation', in N. WARNER, & C. GUSSENHOVEN (eds) *Laboratory Phonology VII*. (Downloadable from http://www.colorado.edu/linguistics/jurafsky).

KARTTUNEN, L. (1998) The proper treatment of optimality in computational phonology. *Proceedings of FSMNLP'98. The International Workshop on Finite-State Methods in Natura Language Processing*. Ankara, Turkey: Bilkent University. June 29–July 1, 1998. (ROA-258-0498 and http://www.cis.upenn.edu/~karttunen/).

KENT, R. & FORNER, L. (1980) Speech segment durations in sentence recitations by children and adults. *Journal of Phonetics* 8, 157–168.

KIPARSKY, P. (1985) Some consequences of Lexical Phonology. *Phonology Yearbook* 2, 85–138.

KIRCHNER, R. (forthcoming). Preliminary thoughts on, 'phonologisation' within an exemplar-based speech processing model. *UCLA Working papers*, Vol. 6 (Downloadable from http://www.ualberta.ca/~kirchner).

KOSKENNIEMI, K. (1983) *Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production*. Publication 11. Helsinki: Department of General Linguistics, University of Helsinki.

LEE, S., POTAMIANOS, A. & NARAYAN, S. (1999) Acoustics of children's speech: developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America* 105, 1455–1468.

MARCUS, G.F. (1998a) Rethinking eliminative connectionism. *Cognitive Psychology* 37/3, 243–282.

MARCUS, G.F. (1998b) Can connectionism save constructivism? *Cognition* 66, 153–182.

MCCARTHY, J. & PRINCE, A. (1995) Faithfulness and reduplicative identity. University of Massachusetts Occasional Papers in Linguistics 18, 249–384. (Downloadable from Rutgers Optimality Archive, ROA 60-2000).

MCCAWLEY, J. (1986) Today phonology, tomorrow the world. *Phonology Yearbook* 3, 27–45.

MCCLELLAND, J.L. & ELMAN, J.L. (1986) The TRACE model of speech perception. *Cognitive Psychology* 18, 1–86.

MCCLELLAND, J.L. & SEIDENBERG, M.S. (2000) Why do kids say goed and brang? *Science* 287, 47–48.

MENN, L. & STOEL-GAMMON, C. (1993) 'Phonological development: learning sounds and sound patterns', in J. GLEASON (ed.) *The Development of Language*, 65–114. New York: Macmillan Publishing Group.

MORETON, E. (1997) *Phonotactic rules in speech perception, Abstract 2aSC4*. San Diego, CA: 134th Meeting of the Acoustical Society of America, December 1–5.

MUNSON, B. (2000) *Phonological pattern frequency and speech production in children and adults*. PhD Dissertation, Ohio: Ohio State University.

MUNSON, B. (forthcoming). Phonological pattern frequency and speech production in children and adults. *Journal of Speech, Language and Hearing Research*.

NORRIS, D.G. (1994) Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52, 189–234.

NORRIS, D.G., MCQUEEN, J.M. & CUTLER, A. (2000) Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* 23, 299–325.

OHALA, J. (1987) Experimental phonology. *Proceedings of the Berkeley Linguistic Society* 13, 207–222.

OHALA, M. & OHALA, J. (1987) 'Psycholinguistic probes of native speakers phonological knowledge', in W.U. DRESSLER (ed.) *Phonologica 1984*. Cambridge UK: Cambridge University Press.

PIERREHUMBERT, J. (1993) Dissimilarity in the Arabic Verbal Roots. *Proceedings of the 23rd Meeting of the Northeastern Linguistic Society, Graduate Student Association*. Amherst: University of Massachusetts.

PIERREHUMBERT, J. (1994) 'Syllable structure and word structure', in P.A. KEATING (ed.) *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form*, 168–188. Cambridge, UK: Cambridge University Press.

PIERREHUMBERT, J. (2000) 'What people know about sounds of language', in *Studies in the Linguistic Sciences* 29/2, 111–120. (Downloadable from http://www.ling.nwu.edu/~jbp).

PIERREHUMBERT, J. (2001) 'Exemplar dynamics: Word frequency, lenition, and contrast', in (eds) J. BYBEE & P. HOPPER *Frequency Effects and the Emergence of Linguistic Structure*, 137–157. Amsterdam: John Benjamins (Downloadable from http://www.ling.nwu.edu/~jbp).

PIERREHUMBERT, J. (in press) 'Why phonology is so coarse-grained', in J.M. MCQUEEN & A. CUTLER (eds) *Language and Cognitive Processes: Special Issue on: Spoken Word Access Processes*. (Downloadable from. http://www.ling.nwu.edu/~jbp).

PIERREHUMBERT, J., BECKMAN, M.E. & LADD, D.R. (2001) Conceptual foundations of phonology as a laboratory science, in N. BURTON-ROBERTS, P. CARR & G. DOCHERTY (eds) *Phonological Knowledge: Conceptual and Empirical Issues*, 273–304. Oxford, UK: Oxford University Press, (Downloadable from http://www.ling.nwu.edu/~jbp).

PINKER, S. (1999) *Words and Rules: The Ingredients of Language*. New York: Basic Books.

PLAUT, D.C. *et al*. (1996) Understanding normal and impaired reading: Computational principles in quasi-regular domains. *Psychological Review* 103, 56–115.

RAIMY, E. & VOGEL, I. (2000) *Compound and Phrasal Stress: A case of late acquisition*. Paper Delivered at the Annual Meeting of the Linguistic Society of America, Chicago, January 6–9.

SANKOFF, D. (1987) Variable rules, in U. AMMON, N. DITTMAR & K.J. MATTHEIER (eds) *Sociolinguistics: An International Handbook of the Science of Language and Society*, Vol. I, 984–997. Berlin: Walter de Gruyter.

SEIDENBERG, M.S. (1997) Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275, 1599–1604.

STERIADE, D. (2000) 'Paradigm uniformity and the phonetics–phonology boundary', in BROE & PIERREHUMBERT (2000) pp 313–334.

TESAR, B. & SMOLENSKY, P. (1998) Learnability in Optimality Theory. *Linguistic Inquiry* 29/2, 229–268. (See http://citeseer.nj.nec.com/tesar96learnability.html).

TREIMAN, R. *et al*. (2000) English speakers' sensitivity to phonotactic patterns, in BROE & PIERREHUMBERT (2000) pp 269–282.

VIHMAN, M. (1996) *Phonological Development: The Origins of Language in the Child*. Oxford, UK: Blackwell Publishers.

VITEVICH, M. & LUCE, P. (1998) When words compete: levels of processing in perception of spoken words. *Psychological Science* 9/4, 325–329.

VITEVITCH, M.S. *et al*. (1997) Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech* 40, 47–62.

VITEVITCH, M.S. *et al*. (1999) Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language* 68, 306–311.

ZURAW, K. (2000) *Patterned Exceptions in Phonology*. PhD Dissertation. Los Angeles, CA: UCLA. (Downloadable from http://www-rcf.usc.edu/~zuraw).