Phonology and perception: A cognitive scientist's perspective

James L. McClelland

As a researcher who has long been interested in the perception, use, and acquisition of language, the title of this volume, *Phonology in Perception*, already piques my interest. Closer examination reveals an exciting development: A diverse group of researchers grounded deeply in the discipline of linguistics are grappling with details of the actual human processing of language, something that would have been almost unthinkable just a few years ago. Every chapter speaks to issues of processing and learning about spoken language and refers to data from experimental psycholinguistics. These developments lend hope to the idea that the distinction between linguistic and psychological approaches to language will gradually fade away, replaced by an interdisciplinary investigation of language, encompassing the structure, use and acquisition of language and even language change. The remarks I make below are offered in the spirit of hastening this integration.

The authors of the various chapters raise a number of issues and questions, either explicitly or implicitly, that lie at the heart of debates within psychological as well as linguistic circles.

- Is language special or does it reflect the operation of domain general principles?
- What is built in, and what is learned, about language?
- Can phonology be treated separately from the sensory and motor processes that are required for overt communication?
- What are the different levels of representation of language, in what form is information represented within each, and how are they interrelated?
- How formal can/should our system of representing language information be? What is the actual status of any such formalization?
- In what form should regularities of language be captured? What is the status of such constructs as rules, constraints, and preferences?
- How can we best capture gradient aspects of language, and do such gradient aspects belong within linguistic theory?

In what follows I will comment briefly on each of these issues. Before I start I would like to make two more general observations.

First, the book exudes a refreshing openness to a very broad range of alternative approaches. As in any field, each author has a particular viewpoint and a particular argument to make in support of that viewpoint; yet for the most part the authors seem open to and interested in the ideas and insights that emerge from the approaches of others. This seems a far cry from an earlier day when clashes of perspectives appeared to be framed in starker, sometimes almost doctrinal terms – and it also seems a very healthy development. Each perspective has its strengths, and it is by seeing these strengths in juxtaposition to each other that we have the greatest chance of being able to find ways to combine the best of each into an ultimately more satisfying synthesis.

Second, the book reflects broad currents within the field, visible both in other work on phonology as well as many other aspects of language use, processing, and learning. Sticking close to phonology, one case in point is Joan Bybee's book on *Phonology and Language Use* (Bybee 2001). Though not focused on the role of perception per se, Bybee argues there for an approach in which phonology is shaped by the use of language, and reflects processes and principles of a general cognitive nature. Other relevant work clearly bridging the fields of linguistics and psycholinguistics includes, for example, the work of Janet Pierrehumbert and others who take an exemplar approach to phonological and lexical representation (Pierrehumbert 2001).

Now on to the issues! Rather than summarize or evaluate arguments made by other authors in this book, I will simply present findings and viewpoints coming from my own background as a cognitive scientist interested in many aspects of cognition, including language.

– Is language special or does it reflect the operation of domain general principles?

This question raged for years within the psychologistic community, with strong proponents for the view that language was special, both at the level of language as a whole (c.f. Fodor 1983) and speech as a specific aspect of language (Liberman 1996).

My own view on this question is that language reflects the operation of domain general mechanisms subject to the particular constraints imposed by the task of linguistic communication – a position that appears to be quite close to the natural phonology position described in the chapter by Balas (this volume). This view has been supported over the years by the joint success of two closely related models, one of context effects in visual letter perception and one of context effects in speech perception. The first of these, the interactive activa-

tion model of letter perception (McClelland and Rumelhart 1981), addressed a phenomenon known as the word superiority effect – the finding that we see letters better when they fit together with their neighbors to spell a word than when they occur in isolation or in a jumbled array of unrelated letters. The model embodied a few simple principles – that when we perceive we rely on graded representations (activations, similar to probability estimates), that activation depends on the propagation of activation via weighted connections (whose values correspond approximately to subjective estimates of conditional probabilities), and that activation spreads, both from the stimulus 'up' and from higher-level representations 'down'. The ideas in this model draw their initial inspiration from properties of neurons, which of course provide the substrate for all aspects of human cognition, and they are closely related to ideas that suffused a number of neural network models proposed as solutions to 'constraint satisfaction problems' that arise in a wide range of domains, including visual scene recognition as well as printed and spoken language perception. Indeed, Jeff Elman immediately recognized the relevance of these ideas for understanding a wide range of findings in the perception of speech sounds, leading us to formulate the TRACE Model of speech perception relying on the same principles (McClelland and Elman 1986).

It is important to note that the TRACE model is not the same as the interactive activation (IA) model. In the IA model, we treat the printed word as arriving at the senses all at once, while in TRACE the speech stream unfolds sequentially over time. This required an elaboration of the architecture of the IA model in a direction that makes the TRACE model somewhat specialized for the processing of speech. A host of issues arise in the case of speech perception that do not arise in the perception of printed words – the ephemeral nature of speech, the absence of word boundaries in the speech stream, and effects of coarticulation are three differences between speech and print. The differences in architecture may reflect the structuring role played by the differences in the task demands of speech perception and visual letter perception, rather than innately pre-specified differences in the neural machinery of speech perception. The work of Sur demonstrating that auditory cortex takes on properties of visual cortex if it receives visual instead of auditory input supports a strong role for experience.

It may be worth noting that the models mentioned here can now be seen as early instantiations of probabilistic models that are enjoying wide popularity today, extending to all aspects of human and machine intelligence including natural language processing. Yet, whenever such models are used, there is always some question of domain-specificity, since a model for any particular domain will always include a set of units, and an arrangement of these units,

that is to some extent domain-specific. We will consider this issue jointly with the second question I raised at the outset.

– What is built in, and what is learned, about spoken language?

The question of what is specific of language, and to what extent whatever is specific must be innately specified, has of course been central in all aspects of linguistic and psycholinguistic inquiry, including in phonology. The issue also comes up in the context of the models mentioned above. In the interactive activation model, there is a structured arrangement of units corresponding to hypotheses about a presented visual stimulus at three levels: a feature level, a letter level, and a word level. In the TRACE model, there are also the same three levels, but, of course, different speech-specific features and a phonemic level in place of the letter level.

Given that written language is not ubiquitous and has only been in use for less than 4000 years, and given the differences in the world's orthographies, it never seemed sensible to assume that the particular feature or letter units needed for perception of written words in any particular language could have been innately pre-specified. Rather, it seems likely that such units came into use as a result of a socio-cultural process working in interaction with available technology for written communication, and that adaptation of the perceptual and motor systems of a child learning how to read and write is largely a matter of learning.

We face what has often been viewed as a very different situation with spoken language, in that, first, spoken language has been with us for much longer than written language, and, second, there are evident commonalities across languages at both the featural and phonological levels. These points, taken together with the fact that the featural and phonological characteristics we see across human languages are not widely exploited in the communication systems used by other species, seem like strong points in favor of the view that somehow the basic building blocks of speech are 'special' to human spoken language and arise as a result of evolution rather than learning.

I consider it important to try to understand how language might have special characteristics that are not built in, or at least not built-in as such. To be sure, there are special characteristics of the human vocal tract that make it better suited to spoken language production than the vocal tracts of other organisms, and these characteristics are clearly given to us by evolution rather than produced in response to experience. Even here, however, the ability of parrots to mimic speech places limits on just how special or unique we should see the elements of human speech production to be.

Within relatively broad constraints established by what we can produce and the effects of alterations in production on perception, a range of perspectives remain viable regarding the extent to which we need to see the units of speech production and perception as innately given. One that I myself find particularly congenial is the idea that the phonological systems found in the world's languages might reflect an optimization over several constraints. (1) Messages should be as easy as possible to produce (2) their characteristics should be perceptually salient and (3) different messages should be mutually distinct from one another. These ideas were introduced by Lindblom and colleagues (e.g. Lindblom, MacNeilage and Studdert-Kennedy 1984) and are being actively pursued by other phonologists (Flemming [1995] 2002; Boersma and Hamann 2008). The suggestion is that these simple principles, together with the physical characteristics of the articulators and the consequences for the sounds that they can produce, could explain the emergence of phonological systems consisting of a largely combinatorial system of phonemes built around contrasts such as manner and place of articulation. Given just a little in the way of an innately predetermined ability to produce the relevant repertoire of gestures, the rest can be left to the same forces that shape the world's orthographies: a socio-cultural process working in interaction with available technology for spoken communication, with the adaptation of perceptual and motor systems within the individual child learning to understand and speak being very largely a matter of learning.

– Can phonology be treated separately from the sensory and motor processes that are required for overt communication?

This issue lies at the heart of the present volume and, perhaps, could be a defining issue for the future investigation of phonology: The general theme of the book is essentially that we will ultimately reap important rewards if we allow the sensory and motor processes involved in the perception and production of speech to affect our thinking about the structure of phonology. To me, as an outsider to the field, the idea that this issue was one that required any discussion comes as quite a shock. True, speech perception researchers once made quite a big deal out of the idea that there was a special 'speech mode' of perception quite distinct from perception of non-speech (c.f. Liberman 1996), but even these researchers treated speech as organized around the recovery of the underlying articulatory gestures that, they believed, were what the perceptual processing of speech aimed to uncover. Thus, the notion that the discipline of phonology might, within certain branches of Optimality Theory at least, be construed as the study of a completely abstract system of essentially arbitrary

constraints seems to me strange and foreign. Luckily, this is not the position taken by the authors of the articles in this book, and so from that point of view, perhaps little more need be said about it. On the other hand, to the extent that the issue is alive at all as a differentiating feature of contemporary perspectives on phonology, perhaps some of the evidence from the field of speech perception that points forcefully toward a role of specifically auditory factors in speech perception is worth a brief mention.

There is now a very large literature that shows how characteristics thought at one time to be special to the perception of speech also arise in non-speech contexts. As one case in point, the categorical perception of the distinction between /b/ and /p/ was once thought to be a special characteristic of the speech mode. But as early as the 1970's, researchers noted that a similar tendency toward categorical perception occurs with the distinction between plucked and bowed violin sounds (Cutting, Rosner and Foard 1976). Other work showed that such distinctions tended to be perceived categorically by non-human animals (Kuhl and Miller 1975). The particular contrasts used in particular languages appear to be influenced by properties of the acoustic signal, but do vary from language to language (Kuhl 1991). It is now widely noted that the tendency for speech perception to be categorical is more marked for consonants than vowels, and it turns out that there is a parallel tendency among non-speech sounds, such that the tendency toward categorical perception is far greater among sounds marked by rapid transitions or abrupt changes, and weaker in perception of sounds distinguished by their steady state characteristics (Mirman, Holt and McClelland 2004). The data are consistent with the view that an intra-linguistic contrast (between consonants and vowels) has a non-linguistic basis, grounded in a distinction in processing between transient and steady-state signals.

A further and perhaps even more telling set of findings relates to the cross-influence of non-speech stimuli on the perception of speech. A key phenomenon taken at first as a sign of the special speech mode of processing was the finding of compensation for co-articulation. A following /l/ pulls a preceding stop forward and a following /r/ tends to push it back. Perceivers compensate for this, tending to perceive an ambiguous sound falling about half way between /d/ and /g/ as a /g/ when followed by an /l/ but as a /d/ when followed by an /r/. Strikingly, however, the same effect can be obtained by following the ambiguous sound by a tone stimulus (Wade and Holt 2005) that is not perceived as speech but that contains frequencies matching those of the third formant onset frequency of /l/ (relatively high) or /r/ (relatively lower). The phenomenon is explained by the authors by assuming that perceptual systems use neighboring frequencies as reference points. Frequencies below a

context reference will be heard as relatively lower (more /d/ like) and frequencies above a context reference will be heard as relatively higher (more /g/ like) (Lotto and Kluender 1998). Thus the perception of the category of a spoken language sound appears to be highly dependent on general purpose auditory processing mechanisms.

– What are the different levels of representation of spoken word forms, in what form is information represented within each, and how are they interrelated?

These are among the central questions of this book, and certainly they are the focus of the introductory chapter by Boersma and Hamann (this volume). They are also very complex questions, and the answers are clearly not independent.

Several models reviewed in the introductory chapter propose an underlying form, a surface form, and two phonetic forms, an auditory phonetic form and an articulatory phonetic form. There, the motivation for considering these different forms arises in the context of capturing phonological phenomena, particularly those that may depend on aspects of perception. Here I will discuss these issues from the point of view of the processing mechanisms involved in perception and production.

It seems uncontroversial enough to think that most utterances arise because speakers have something in mind to say; so there must be some intended communicative content; and when they speak, they produce a sequence of muscular contractions driving the articulators. Although it is possible to imagine otherwise, most theories do posit that the intended communicative content is first translated into some sort of underlying representation capturing aspects of the intended articulation (e.g. the sequence of abstract phonological segments contained in the message, generally embellished with stress and structure markings), which is then further transformed to produce the overt muscle contractions and resulting trajectories of the articulators. So, we have at least three representations: The intended message, the underlying representation of articulatory content, and the actual sequence of muscle contractions and movement patterns in the articulators.

Proceeding toward the receiving side, it seems uncontroversial to state that the process of articulation gives rise to an auditory waveform. Perceptible visual cues are also produced and are known to play a role in speech perception; it seems likely that such cues will ultimately play a role in explaining some phenomena in phonology, but I will not consider them further here. The auditory waveform gives rise to internal processes within the listener, which appear to involve formation of both a perceptual representation – what the listener thinks

s/he heard – and a conceptual representation, roughly, the message the perceiver construes the speaker to have intended to communicate. On the perceiving side, then, there would appear to be three forms: the auditory waveform, the percept, and the message-as-received.

So far of course I have said very little, attempting to be as neutral as possible. To say more than this is to begin to specify one particular theory; a theory specifying, for example, the actual form and structure of intended communicative content; the form and structure of a percept or of the underlying representation of articulatory content; and the aspects of articulation and resulting auditory waveform that are relevant to speech perception. It is apparent that these matters are far from settled.

For example, there are many models of human language processing that posit the existence of lexical entries or lexical units that are supposed to mediate between intended communicative content and the percept on the input side and the underlying representation of the intended articulation on the output side. The TRACE model of speech perception is an example of a model that contains such units on the perceptual side, and Levelt and his collaborators (Levelt, Roelofs and Meyer 1999) have proposed that speech production involves an essential stage of lexical unit (lemma) selection as an intermediary between meaning and speech.

While these early models contained processing units corresponding to conventional lexical and phonological units (words and phonemes), more recent models employ learned distributed representations. One such model is shown in Fig. 1 (it is similar to models implemented in Dilkina, McClelland and Plaut 2008 and Rogers *et al.* 2004). The model illustrates the approach my colleagues and I take to characterizing the mechanisms involved in understanding and speaking, as well as in perceiving objects and events in the world and then taking action. Pools of units correspond to different types of perceptual representations – percepts of printed or spoken words, or of actual objects – as well as plans for actions of different types, including producing the correct spoken word for a presented object (saying ball in the example). The model includes an integrative layer that receives projections from and sends projections to the different perceptual and output layers respectively. The figure also includes a direct pathway from the pool corresponding to the speech percept to the pool corresponding to the articulatory plan (Hickok and Poeppel 2003) as well as pathways from a visual word-form representation to both the speech percept and articulatory plan representations (Mechelli *et al.* 2005). Additional unlabeled pools of units are included to represent other possible inputs (haptic, olfactory) and other possible types of output (possibly including emotional responses, for example).

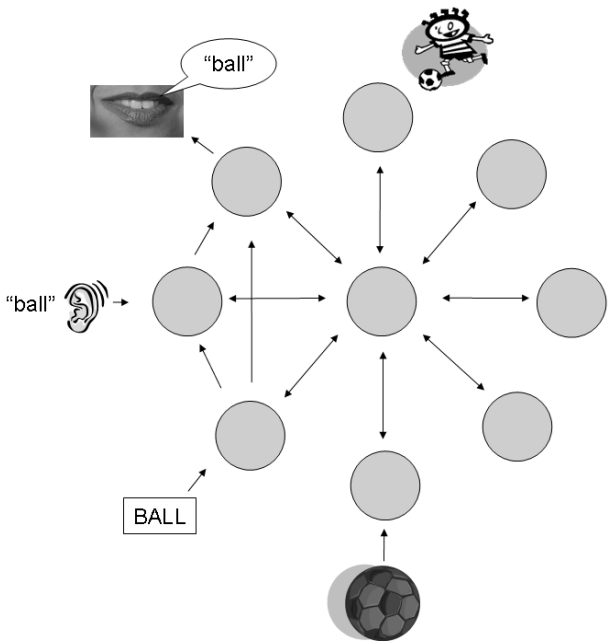

Figure 1. A schematic rendition of the distributed connectionist model of semantic and lexical knowledge, extending a similar figure in Dilkina, McClelland and Plaut (2008). Each circle stands for a pool of neuron-like processing units over which patterns of activation represent some aspect of experience with objects or words. The number and functions of all of the pools of units involved are not known, but the cognitive neuroscience literature assigns specific brain areas that correspond to some of the pools of units in the model. As one example, there is a 'visual word form area' corresponding to the pool that represents the pattern corresponding to the visual form of a word (lower right pool of units in the figure). Other pools are associated with auditory word forms and articulatory word forms. Still other pools are associated with representations of the visual forms of objects, the actions we take on objects, etc. A single integrative layer is shown in the middle of the figure, with bi-directional arrows to each of the other pools. According to the theory (Rogers *et al.* 2004; Dilkina, McClelland and Plaut 2008), bi-directional connections from each of the surrounding pools to the integrative pool in the middle allow input arising in any of the pools to give rise to the corresponding output on any of the other pools. Thus on hearing the word 'ball' activation would propagate from the auditory representation layer to the integrative layer, and from there to all the other layers, allowing the network then to pronounce the word and visualize its spelling, and also to imagine the object and the action one might take upon it, among other things.

Existing implemented models of this type (e.g., Dilkina, McClelland and Plaut 2008) learn to map from patterns on the visual and spoken input layers to appropriate output layers, and do so without employing individual processing units corresponding either to individual words or to individual concepts. Instead, the models rely on learned distributed representations that mediate between the different input and outputs, and that are acquired through a neural network training algorithm. Each item develops its own learned distributed representation over the integrative layer that mediates between all of the different types of information about both words and objects (the same representation mediates representations of the spelling, sound, and articulation of the word 'ball' and the associated conceptual knowledge of what balls look like, how they move, how we interact with them, etc.). At first all items rely on highly overlapping patterns of activation, but as learning proceeds these become differentiated, increasing distinctiveness but not completely eliminating overlap. While these learned distributed representations function like concepts or lexical entries in some ways, they are graded, distributed representations whose patterns of overlap reflect similarity relations. As such they show tendencies to generalize and to degrade gracefully under damage in ways that are not intrinsic to models containing discrete units or entries for individual lexical items.

Models of this type have been highly successful in accounting for the effects of a neurological disorder thought to affect the brain analog of the integrative layer – the anterior temporal cortex. Among the findings is the fact that patients with this disorder lose specific information about concepts as well as specific information about words, while still preserving more general knowledge about words and objects. Patients still know what typical objects look like, and make errors that "typicalize" exceptional properties of objects (drawing, for example, a human-like ear in place of an elephant's ear when drawing a picture of an elephant). Patients also still know typical spelling-sound correspondences, and typicalize exceptional aspects of word's spellings, and our models do the same (McClelland, Rogers, Patterson, Dilkina and Lambon Ralph 2008). No discrete lexical or conceptual units are employed in capturing both correct normal performance as well as the effects of brain damage.

For simplicity, implemented versions of our models use one dedicated unit to stand for each phoneme in the phonological input pool, a unit for each letter in the word form pool, and a unit for each phoneme in the speech percept pool, and another unit for each phoneme in the speech output pool. In this respect they are like many other models and psychological theories that contain explicit phoneme units. In our case, however, we view this, not as a representation of reality but as a simplification. Just as words and concepts need not be represented by individual dedicated units, so also even phonemes and gra-

phemes might not really be represented by such units either. Because hearers appear to be sensitive to auditory detail and speakers produce the same 'phoneme' in different ways that are often lexical-item-specific (e.g., the silence and following burst release associated with the /t/ in 'softly' are briefer that those associated with the /t/ in 'swiftly'; Hay 2001) it has been suggested that spoken language representation might contain far more articulatory or auditory detail than is naturally captured by thinking that a word's phonological form is represented as a string of discrete units. Indeed, models that map from raw acoustic input via an intermediate layer of learned distributed representations onto some sort of meaning-like representation have been developed (Kaidel, Zevin, Kluender and Seidenberg 2003), and we plan to incorporate such learned distributed reprsentations in future implementations of the model shown in figure 1.

With these efforts as context, the idea that human mental processing of language may involve neither lexical nor phonological units in any kind of explicit form becomes more and more of a possibility. Within this context, we can ask, just what is the status of the different levels of representation postulated in linguistic theories of phonology?

The proposal that arises from a distributed connectionist perspective is to view the units and levels found in linguistic theory as useful approximations that serve to succinctly characterize clusters of material that is similar in some respect rather than strictly identical. For example, we use the symbol [p] to represent a wide range of slightly different articulatory gestures that share several properties and have similar acoustic consequences. To distinguish useful subsets of these we use additional markings, for example, to distinguish aspirated and unaspirated variants. We recognize a regularity within this class of sounds, which is that aspiration tends to be reduced or absent when /p/'s follow /s/'s but to be present to a greater degree and more often when /p/'s occur in word-initial position. These are useful descriptive statements even if in fact aspiration is a matter of degree, and even if there is overlap in the frequency distributions for different degrees of aspiration in the different types of contexts.

In light of the above, I often find debates in linguistics about the relative merits of different formalisms for capturing regularities to be unnecessary. In fact it is my belief that no such formalism will ever really do full justice, and that there are many with considerable utility.

- How formal can/should our system of representing language information be? What is the actual status of any such formalization?
- In what form should regularities of language be captured? What is the status of such constructs as rules, constraints, and preferences?
- How can we best capture gradient aspects of language, and do graded strength parameters have a role in linguistic theory?

The three issues above seem intimately intertwined, and I've already begun to indicate the general nature of my own preferred answer. Although these questions do not come up overtly in most of the papers in this volume, the chapter by Balas (this volume) does raise them explicitly in her contrast between natural phonology and OT. We have, on the one hand, within OT, a seeming commitment to a program quite similar in some ways to Chomsky's program in syntax, seeking a very abstract and formal characterization of the principles of phonological structure. As characterized by Balas, 'classic' OT is treated as a purely formal system, devoid of sensory-motor content, stipulating a set of universal constraints that govern phonological forms and different only between languages in how the constraints are ranked. OT then invites us, we are told, to see learning as a matter of establishing constraint rankings, a task that should be simpler, than, say, learning exactly how the constraints should be structured or formulated. On the other hand, natural phonology is cast as a framework within which very general pressures – e.g. to keep messages short and simple but also distinct – operate in conjunction with characteristics of the articulatory apparatus of speech, the ways in which articulation shapes sound, and the ways in which sound is processed by mechanisms of auditory processing to shape the characteristics of phonological forms.

To me, it is clear that both approaches have their virtues, especially when viewed as ways of helping to channel researcher's thinking toward insights into the nature and structure of natural languages. As an outsider to the field of phonology, particularly to the full and by now very complex literature on OT, it is difficult to have a definitive take on the prospects for the OT program in the form stated above. However, from my own research in one circumscribed sub-area in phonology -- the rimes found in English word forms – it seems to me that the search for a simple list of universal constraints can take us part, but not all the way, toward a characterization of the details of phonological structure. I therefore see OT as being a useful formalism, but one that should be viewed as providing only an approximate characterization of the real underlying nature of phonological structure.

As an illustration of these points, let us consider the data in Fig. 2. The figure displays a partial ordering of each of several different rime types occurring

in monomorphemic, monosyllabic word forms in the CELEX English corpus (Baayen, Piepenbrock and van Rijn 1993). The figure encompasses all of the rime types in the corpus containing a short (V) or long (VV) vowel, and a single stop consonant plus no more than one pre-stop consonant – a nasal, an l, or an alveolar fricative. The numbers written next to each rime in the figure indicate the average per-vowel[1] occurrence rates in the corpus of monosyllabic, monomorphemic English word lemmas of each type. Thus, for example, for the form Vt there are 113 such lemmas summing over the 5 short vowels included in the corpus, producing an average per vowel occurrence rate of 22.6 for this rime type.

Within these rime types, several very general principles seem to hold. Four very simple, and arguably[2] universal constraints – keep it short, simple, coronal, and unvoiced – do a good job of capturing ordinal relationships among the occurrence rates of the different types of rimes listed in the figure. There is a preference for short relative to long vowels. There is a preference for simpler forms – those without the added consonant – compared to their more complex counterparts. There is a preference for coronal relative to non-coronal stop consonants, and a preference for unvoiced relative to voiced stops. The constraints are represented in the figure by placing forms that violate a given constraint below those that adhere to it, and connecting members of the same minimal pair – a pair of forms that are the same except that one violates exactly one more constraint than the other – with an upward arrow. Where the arrow is solid, the data are consistent with the constraint, in that either (i) the form at the top of the arrow has a greater occurrence rate than the form at the bottom of the arrow or (ii) neither form occurs at all (see the figure caption for more details). Of the 140 minimal pairs encompassed by the figure, there are 135 where the occurrence rates are consistent with the constraints, and only 5 case that are inconsistent, indicating strong overall consistency with the four simple constraints.

---

1. The CELEX English corpus uses Southern British English pronunciations. The counts exclude forms containing relative rare vowels of each of the short and long types. Five short and 10 long vowels are included. See McClelland and Vander Wyk (2006) for more details.
2. I say 'arguably' here to make it clear that any claim of universality will require more detailed specification of the constraints. In particular, the context in which these universals apply must be specified. A preference for relatively simpler forms seems likely to operate generally. A preference for unvoiced relative to voiced and coronal as opposed to non-coronal articulation in codas of monomorphic monosyllabic word forms appears widespread, as does a preference for short over long vowels in forms that contain stop consonants in the coda. Some or all of these preferences may be at work in other contexts, but may be overridden in other contexts by counter-veiling factors.

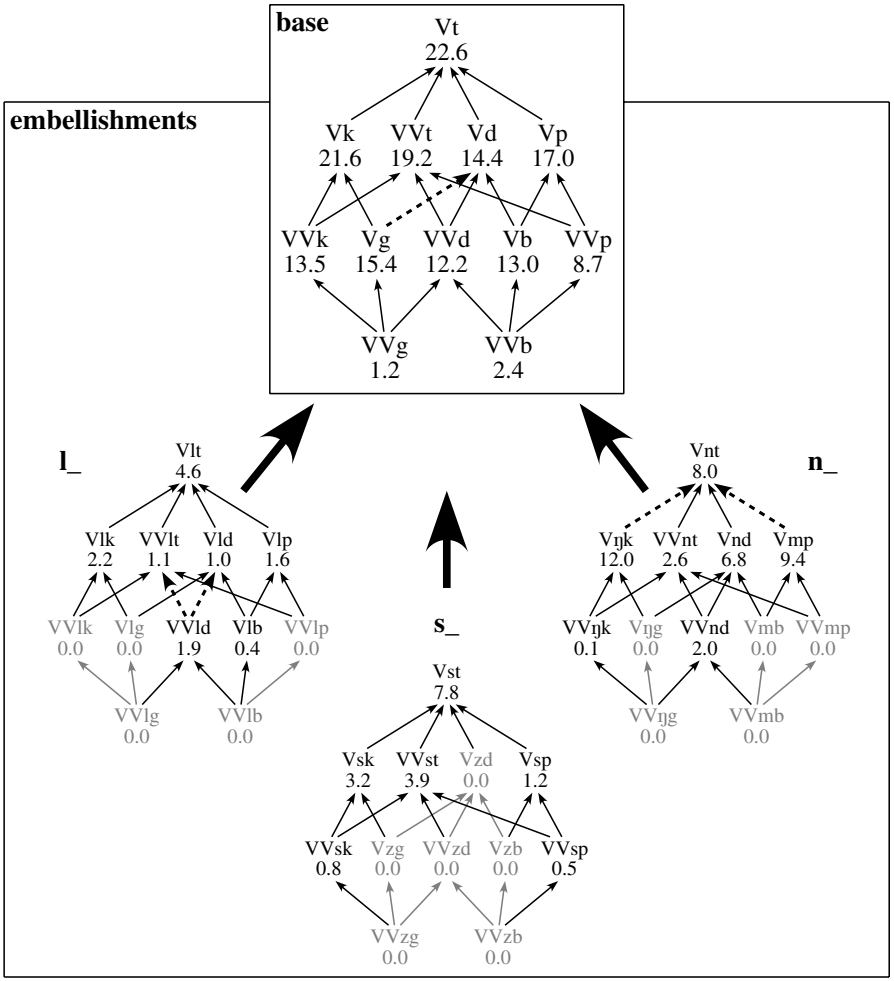

Figure 2. Partial ordering graph showing graded influences of several constraints. The base subgraph in the small box at the top shows the rime types containing a single coda stop consonant which may be either voiced or unvoiced, coronal or non-coronal, and is preceded by a short (V) or long (VV) vowel. Numbers shown are the average per-vowel occurrence rates in the corpus, as described in the text. Arrows indicate dominance relationships according to the four constraints described in the text. Solid arrows are used in cases where the occurrence rates are consistent with the constraints, and dotted arrows indicate cases where the occurrence rates violate the constraints. The other three sub-graphs within the larger box

Clearly, these data at least provide evidence for each of these four constraints, and leaving them simply as abstract principles might be viewed as a good, first-order summary of this data. Here we characterize these constraints in the simplest possible form:
*VV (disprefer long vowels)
*X (disprefer added segments of any type)
*Voi (disprefer voicing of coda obstruents)
*NC (disprefer non-coronals)

One the one hand, we could see these constraints as compatible with an OT approach, in that they are very abstractly formulated and possibly universal. On the other hand, it is also possible to view some of these constraints as so general that they are not really specific to language. One possibility is that, at least in part, all of these constraints reflect a pressure to keep word forms shorter in duration. Of course, forms containing fewer segments and short vowels rather than long vowels do take less time to articulate. Somewhat less obvious is the finding that violations of Voicing and Coronality are also associated with longer spoken word form durations (Vander Wyk and McClelland in preparation). Thus, it may be that all these constraints reflect a very general preference for shorter word-forms, a constraint that does not seem on the face of it to require an appeal to a construct such as Universal Grammar.

On the other hand, these constraints, without further details, do not provide a full account of the data. If we wish to explain in more detail exactly which forms do occur and which forms do not, or if we wish to address the occurrence rates quantitatively, we will need to specify additional information. Here we seem to pass beyond what is ordinarily offered in the abstract framework of Optimality Theory. Even just to address whether a particular rime type does or does not occur, we already run into difficulty, if we try to rely on the standard constraint ranking logic of OT. If *Voi outranks Faithfulness, then no voiced coda obstruents should occur, but if Faithfulness outranks *Voi, then coda

---

indicate corresponding data for cases where the rime contains a pre-stop /l/, pre-stop /s/, or pre-stop nasal segment. The solid arrows from each of these three sub-graphs to the sub-graph in the box indicates that in every case, the presence of the pre-stop segment reduces the occurrence rate of each form in the subordinate sub-graph, compared the corresponding base form in the base subgraph. As one example, the occurrence rate of Vst, a form in the pre-stop /s/ subgraph, is less than the occurrence rate of Vt, a form in the base subgraph. Adapted with permission from McClelland and Vander Wyk (2006).

obstruents should occur without penalty. Clearly, none of the abstract constraints under consideration individually outranks faithfulness, since forms violating each of these constraints do occur in the language. However, when a rime type violates several of the constraints, it may well not occur in the language. Some form of constraint cumulation appears to be in order, violating a principle employed in standard versions of OT. Simply specifying that up to two violations are allowed but that a third is not[3] might capture some of the data, but some forms that violate three constraints do occur (VVnd as in *find* violates *VV, *X, and *Voi) and some that violate three constraints do not (Vmb violates *X, *Voi, and *NC, and there are no words containing this rime type – the b in bomb, for example, is not pronounced).

One way to go beyond the limits of standard OT is to stay with the idea of very abstract constraints, but to return to the approach taken in Harmony Theory, the predecessor of OT, which relied on weighted parameters and a quantitative rule for combining the weights. Appeals to graded constraints are, of course, quite common in phonological research (e.g., Harris 1994), including work undertaken within the OT framework (Boersma 1998; Burzio 2000), and quite a lot of formal work is now being undertaking using some form of graded constraint representation (e.g., Hayes and Wilson 2008). In addition to the notion that constraints have continuous-valued weights, it will be useful to allow continuous variation in the degree to which a particular constraint is violated by a particular word form. Allowing the total extent of constraint violation to be given by the product of a continuous-valued weight specifying the importance of the constraint times a continuous-valued score specifying the degree of the violation should simplify, for example, the analysis of many of the phenomena reviewed in the chapter on cue constraints by Boersma (this volume).

In McClelland and Vander Wyk (2006), we proposed an extremely simple version of this idea, in which the underlying constraint violation score (CVS) associated with a form is a simple linear function of the set of constraints that it violates:

$$\mathrm{CVS}_i = \Sigma_j \, C_{ij} \, w_j$$

Here the subscript i indexes different rime types, and the subscript j indexes constraints. w_j refers to the strength or weight of constraint j, and C_{ij} takes the value 1 if the rime type violates constraint j, and is 0 otherwise. Smaller CVS values are associated with 'better' forms.

3. Even this appears to violate the standard version of OT, in which counting of violations is not allowed (Prince and Smolensky 2004).

To relate this formula to the actual occurrence rates of forms in English, we found that the following function provided a better fit to the data in the table than other formulations:

$$S_i = B - \mathrm{CVS}_j \; ; \quad R_i = [S_i]+$$

The expression simply states that the strength (S_i), or tendency to occur, of a particular rime type is equal to a (positive) baseline occurrence rate B, less the constraint violation score. The notation $R_i = [S_i]+$ indicates that the predicted occurrence rate of the form, R_i, is simply equal to S_i if the value of S_i is greater than 0; otherwise, R_i is equal to 0. Note that B itself already reflects constraints operating on the simplest form included (Vt). The remaining constraints cumulated in the *CVS* for a given item include one for Long relative to Short vowels, and one for voiced relative to unvoiced obstruents in the coda. There are additional constraints corresponding to penalties for non-coronal articulation and for adding additional segments over and above the vowel and one stop consonant. In fitting the data, we found that some types of added segments appeared to exert a greater cost than others, and different ways of being non-coronal also appeared to vary in cost. Thus, we found it useful to include a separate constraint violation weight for each type of pre-stop coda consonant (pre-stop /s/, prestop /l/, prestop nasal) and a separate constraint violation weight for each for the two types of non-coronal stops (velars and labials). The version of this model that we used to fit the data in along with some additional data not shown had ten[4] numeric parameters (the baseline B plus nine constraint weights), and accounted for 85% of the variance in the observed occurrence rates. It also correctly predicted that 38 of the 40 rime types that do occur would occur, and only incorrectly predicted that 4 rime types would occur that do not occur. All of the mispredictions were relatively small in magnitude (i.e., the 4 forms predicted to occur that do not occur were predicted to occur with low rates). This level of success in our model supports the view that it may be worthwhile to consider integrating graded constraints in a more thoroughgoing way into phonological theory, and to treat what has become the standard version of OT as a simplification that may be useful for some purposes.

While inclusion of graded constraint weights and graded degrees of constraint violation should help, even this may not be enough to account for all the subtleties in the real data. Even in the data summarized in Fig. 2, there are a few deviations in the partial ordering predictions that would still be unexplained. As one example: the type VVld occurs less frequently than Vld, even

[4] In fact our fits used a data set including forms containing post-stop /t/ and post-stop /s/, requiring 1 more weight for each of these two types of added segments, for a total of 10.

though the former has a long vowel. Furthermore, there appear to be some constraint interactions: We find that non-coronal place of articulation interacts strongly both with consonant voicing and with the presence of a nasal segment (see Fig. 3). How are these additional features of the data to be explained?

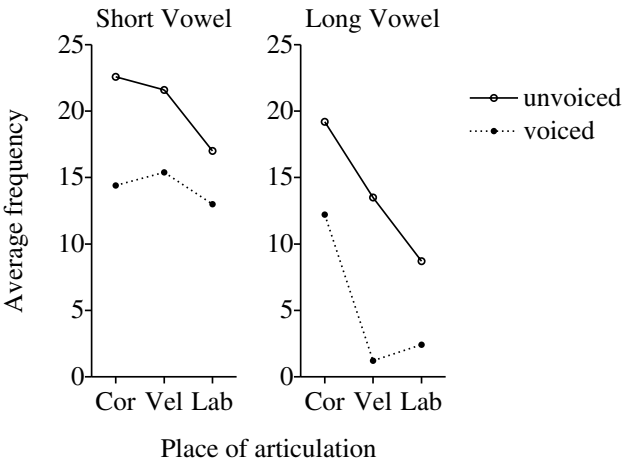

Figure 3. Average per vowel occurrence rates for the six forms in the base subgraph of Fig. 2. Each data point represents a rime type, consisting of a single vowel, which may be short or long and a single stop consonant, which may be unvoiced or voiced, and which may have a coronal (cor) velar (vel) or labial (lab) articulation. It is evident that the constraint against non-coronals is greater when the vowel is long, and may be amplified further when the consonant is voiced. Adapted with permission from McClelland and Vander Wyk (2006).

A number of possible explanations can be envisioned. As suggested by natural phonology, some of these may well involve details of interactions between the actual gestures required to produce adjacent segments and/or effects of attempts to combine such gestures on perceptibility. For example, the gesture required to produce an /l/ may interact with the gestures required to produce neighboring vowel segments in ways that make some long vowels more compatible with a following /l/ than some short vowels, or may shift the perceived quality of the preceding vowel. Another possible type of explanation may revolve around the idea that the distribution of word forms in the language is a solution to an optimization problem, in which the distribution of rime types in the language is thought of as a compromise solution, influenced both by simplicity as well as perceptual distinctiveness of the resulting word forms.

The point of reviewing these ideas here has been to suggest that the succinct statement of abstract constraints, as in some versions of Optimality Theory, should be viewed, not as a matter of fundamental theoretical principle, but as a matter of simplicity that allows a good approximate description of the facts of phonological structure in a very compact and straightforward form. A full understanding will require appeals to the actual magnitudes of particular specific instances of constraints, as well as appeals to particular details of articulation. On this view, OT may be a useful notational framework that facilitates understanding in some cases, but it should not be viewed as the one true way to characterize phonological structure. I would, in fact, suggest that the search for the 'true' abstract framework for capturing phonological (or any other aspect of linguistic) structure may no longer be the best path toward a fuller understanding. We should continue the effort to provide useful ways of summarizing facts about language structure, but view these as essentially descriptive activities, without seeing alternative approaches as in fundamental opposition to each other.

References

Baayen, R. Harald, Richard Piepenbrock, and Hedderik van Rijn
1993 *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balas, Anna
this volume Why can Poles hear "Sprite", but not "Coca-Cola"? A Natural Phonological account.

Boersma, Paul
1998 *Functional Phonology: Formalizing the Interactions Between Articulatory and Perceptual Drives*. Ph.D. dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.
this volume Phonological perception as an interplay between structural and cue constraints.

Boersma, Paul, and Silke Hamann
2008 The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25: 217–270.
this volume Introduction: Models of phonology in perception.

Burzio, Luigi
2000 Missing players: Phonology and the past-tense debate. In: Kleanthes K. Grohmann and Caro Struijke (eds.), *University of Maryland Working Papers in Linguistics* 10: 73–112.

Bybee, Joan L.
2001 *Phonology and Language Use*. Cambridge: Cambridge University Press.
Cutting, James E., Burton S. Rosner and Christopher F. Foard
1976 Perceptual categories for musiclike sounds: Implications for theories of speech perception. *The Quarterly Journal of Experimental Psychology* 28: 361–378.
Dilkina, Katia, James L. McClelland and David C. Plaut
2008 A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology* 25: 136–164.
Flemming, Edward
1995 *Auditory Representations in Phonology*. Ph.D. dissertation, University of California, Los Angeles. Published London and New York: Routledge [2002].
Fodor, Jerry A.
1983 *Modularity of Mind*. Cambridge, Mass.: MIT Press.
Harris, John
1994 *English Sound Structure*. Oxford: Blackwell.
Hay, Jennifer
2001 Lexical frequency in morphology: Is everything relative? *Linguistics* 39: 1041–1070.
Hayes, Bruce, and Colin Wilson
2008 A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379–440.
Hickok, Gregory, and David Poeppel
2004 Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92: 67–99.
Keidel, James L., Jason D. Zevin, Keith R. Kluender, and Mark S. Seidenberg
2003 Modeling the role of native language knowledge in perceiving nonnative speech contrasts. In: Marie-Josep Solé, Daniel Recasens and Joaquin Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*, 2221–2224. Barcelona.
Kuhl, Patricia K.
1991 Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50: 93–107.
Kuhl, Patricia K., and James D. Miller
1975 Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science* 190: 69–72.
Levelt, Willem J. M., Ardi Roelofs, and Antje S. Meyer
1999 A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22: 1–38.
Liberman, Alvin M.
1996 *Speech: A Special Code*. Cambridge, MA: MIT Press.

Lindblom, Björn, Peter F. MacNeilage, and Michael Studdert-Kennedy
1984 Self-organizing processes and the explanation of phonological universals. In: Brian Butterworth, Bernard Comrie and Östen Dahl (eds.), *Explanations for Language Universals*, 181–203. Berlin: Mouton.

Lotto, Andrew J., and Keith R. Kluender
1998 General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics* 60: 602–619.

McClelland, James L., and Jeffrey L. Elman
1986 The TRACE Model of Speech Perception. *Cognitive Psychology* 18: 1–86.

McClelland, James L., Timothy T. Rogers, Karalyn Patterson, Katia N. Dilkina, and Matthew R. Lambon Ralph
in press Semantic cognition: Its nature, its development, and its neural basis. In: Michael Gazzaniga (ed.), *The Cognitive Neurosciences IV*. Boston, MA: MIT Press.

McClelland, James L., and David E. Rumelhart
1981 An interactive activation model of context effects in letter perception. Part 2: An account of basic findings. *Psychological Review* 88: 375–407.

McClelland, James L., and Brent C. Vander Wyk
2006 *Graded Constraints in English Word Forms.* Working manuscript, Department of Psychology, Carnegie Mellon University.

Mechelli, Andrea, Jennifer T. Crinion, Steven Long, Karl J. Friston, Matthew R. Lambon Ralph, Karalyn Patterson, James L. McClelland, and Cathy J. Price
2005 Dissociating reading processes on the basis of neuronal interactions. *Journal of Cognitive Neuroscience* 17: 1753–1765.

Mirman, Daniel, Lori L. Holt, and James L. McClelland
2004 Categorization and discrimination of non-speech sounds: Differences between steady-state and rapidly-changing acoustic cues. *Journal of the Acoustical Society of America* 116: 1198–1207.

Pierrehumbert, Janet
2001 Exemplar dynamics: Word frequency, lenition and contrast. In: Joan L. Bybee and Paul Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, 137–157. Amsterdam: John Benjamins.

Prince, Alan, and Paul Smolensky
2004 *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell.

Rogers, Timothy T., Matthew R. Lambon Ralph, Peter Garrard, Sasha Bozeat, James L. McClelland, John R. Hodges, and Karalyn Patterson
2004 The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review* 111: 205–235.

Vander Wyk, Brent C., and James L. McClelland
in preparation Constraints affecting occurrence rates of English word forms also affect spoken word duration.
Wade, Travis, and Lori L. Holt
2005 Effects of later-occurring non-linguistic sounds on speech categorization. *Journal of the Acoustical Society of America* 118: 1701–1710.