What is computational phonology?

Robert Daland

University of California, Los Angeles
e-mail: rdaland@humnet.ucla.edu

ABSTRACT: Computational phonology is not one thing. Rather, it is an umbrella term which may refer to work on formal language theory, computer-implemented models of cognitive processes, and corpus methods derived from the literature on natural language processing (NLP). This article gives an overview of these distinct areas, identifying commonalities and differences in the goals of each area, as well as highlighting recent results of interest. The overview is necessarily brief and subjective. Broadly speaking, it is argued that learning is a pervasive theme in these areas, but the core questions and concerns vary too much to define a coherent field. Computational phonologists are more united by a shared body of formal knowledge than they are by a shared sense of what the important questions are.

KEYWORDS: computational phonology

RESUMEN: *¿Qué es la fonología computacional?.-* La fonología computacional no representa un campo unitario, sino que es un término genérico que puede hacer referencia a obras sobre teorías de lenguajes formales; a modelos de procesos cognitivos implementados por ordenador; y a métodos de trabajo con corpus, derivados de la bibliografía sobre procesamiento del lenguaje natural (PLN). Este artículo ofrece una visión de conjunto de estas distintas áreas, identifica los puntos comunes y las diferencias en los objetivos de cada una, y pone de relieve algunos de los últimos resultados más relevantes. Esta visión de conjunto es necesariamente breve y subjetiva. En términos generales, se argumenta que el aprendizaje es un tema recurrente en estos ámbitos, pero las preguntas y los problemas centrales varían demasiado como para definir un área de estudio unitaria y coherente. Los fonólogos computacionales están unidos por un cúmulo común de conocimientos formales más que por un parecer compartido acerca de cuáles son las preguntas importantes.

PALABRAS CLAVE: fonología computacional

1. INTRODUCTION

What does it mean to be a scientific field of inquiry? Proceeding inductively, we might observe that well-established fields tend to exhibit the following properties:

(i) a core set of observable phenomena, which the field seeks to explain
(ii) a core set of research questions the field asks about those phenomena
(iii) a shared set of background knowledge that is in part specific to the field
(iv) a shared 'toolbox' of research methods used for gaining new knowledge

These properties exhibit a granularity of scale; within one field there may be sub-fields which ask more specific questions, assume greater amounts of shared knowledge than the field as a whole, and utilize a restricted set of methodologies. For example, linguistics is a rather wide field of inquiry; within this field there is a sub-field devoted to the study of syntax specifically. Because science is a dynamic and evolving enterprise, scientific fields exhibit the same kind of taxonomic structure as other evolutionary systems, such as species and languages –subfields may have sub-subfields of their own, and particular sub-fields may have more in common with a different field than the 'parent' field. For example, *psycholinguistics* can be considered a sub-field

of linguistics, but the research methods and the specialized knowledge specific to that field are arguably closer to the field of psychology. But, then, which of properties (i)-(iv) are essential to a field? The answer to this question will inform our answer to the question, "What is computational phonology?"

Some perspective on this question can be gained by considering the historical development of a field. Fields can occasionally form, or shift dramatically in character, with the emergence of a charismatic and persuasive thinker or a seminal publication. This was arguably the case in linguistics with Chomsky's review of B. F. Skinner's *Verbal Behavior* (1959) and other related publications (Chomsky, 1956). Fields may also stratify to the extent that it is worth considering them as two different fields. For instance, most of the scientific fields we know today have their roots in philosophy. Fields may coalesce by the identification of similar strands of thought in fields that were formerly separate; such is arguably the case with the field of cognitive science, or more specifically with psycholinguistics. In the case of newer, less-established fields, especially those which coalesced from multiple other fields, there is a much smaller core of shared, field-specific knowledge. Arguably, the codification of a shared body of field-specific knowledge is the *consequence* of establishing academic programs/departments for a given field, rather than a cause or necessary property of fieldhood. As for the research methods of a field, they are ever-changing. Methodology might be used to characterize a field at a particular historical moment, but most fields persist through several methodological turnovers. For example, the increase in computer resources over the last 50 years has revolutionized linguistic methodology, but the questions we ask now are arguably the same ones that Chomsky laid out in the 1950s: How do children learn language? Out of the space of logically imaginable linguistic patterns, why do many systematically not occur? To what extent can the occurrence/non-occurrence of linguistic patterns be explained by functional aspects of communication, and to what extent is it determined by properties of the cognitive systems(s) that process and represent language?

There is room for legitimate disagreement on this point, but for many researchers, a field revolves around a set of empirical phenomena, and a key set of research questions the field seeks to answer about those phenomena. By this standard, I will suggest that computational phonology is not really a single field. Rather, the phrase 'computational phonology' is used as an umbrella term for research which generally presupposes a shared, specific set of background knowledge and uses a common set of research methodologies, but often with radically diverging questions. I will make this case by surveying recent progress in four different subfields, all of which I or colleagues have identified as 'computational phonology'. We shall see there is a general emphasis on learning, and that all or most practitioners have a common background in corpus and finite-state methods, but

the sub-fields themselves differ quite radically in what the research questions are.

Prior to the survey, it is necessary to voice a caveat. The view of the field that I present is my own. I make no claim that the survey below is comprehensive, or unbiased; in fact, I avow that this review is strongly biased toward my own research interests, the readings I have done, and by informal conversations I have had with colleagues. I have surely omitted mention of a great deal of important and interesting work, either from time/space constraints or because I have not yet had the honor of being exposed to it. Still, as a multidisplinary researcher I hope that all readers will find something new within these pages, and I have aimed for the fairest, most scrupulous and scholarly tone for the works I was able to review here. Prior to the body of the paper I briefly review background material.

2. BACKGROUND

2.1. What is phonology?

I assume that the reader of this article has some background in formal linguistics, perhaps equivalent to a one-year undergraduate sequence covering phonetics, phonology, and other core areas. For example, I assume the reader is familiar with the concept of underlying representation (UR; also called lexical representation, or input) versus surface (SR; also called output), and the convention that URs are indicated with slashes // while SRs are indicated with brackets []; I assume knowledge of the terms 'segment', 'syllable', 'onset', 'coda', et cetera, and the International Phonetic Alphabet. Still, as I anticipate some readers will come from a computational background where the study of speech sounds is not emphasized, I will briefly describe here core concepts which figure prominently in the paper.

2.1.1. Markedness

Cross-linguistically some structural configurations appear to be dispreferred. For example, French has a complex process known as schwa deletion, in which the weak schwa vowel tends to delete, except if the deletion would create a triconsonantal cluster (Riggle & Wilson, 2005). Moreover, triconsonantal clusters do not appear in many languages, and tend to have a restricted distribution in languages that allow them at all. It appears as if French and many other phonologies are specifically avoiding this 'marked' configuration. The proper treatment of markedness is a core concern in phonological theory. What structural configurations are marked? How is markedness represented in the minds of speakers? How is markedness acquired –is it learned from phonetics, projected from the lexicon, or something else?

2.1.2. Alternations

Alternation is the name given to cases in which the same phonological entity appears with two or more forms. For example, compare my casual pronunciations of the English words *pentagon* and *pentagonal*:

(1)   *pentagon*   [ˈpʰɛ̃ ɾ̃ ə ˌgã n]

       pentagonal   [p ə̃ n ˈtʰ æ g ə̃ n -əɬ]

In (1), segment-to-segment identity is indicated by vertical alignment. Non-identical correspondents are vertically aligned, but indicated with a vertical bar or slash. Every corresponding vowel is different between these two forms, owing to the different position of stress. In addition, the medial coronal stop /t/ is aspirated in *pentagonal* because it precedes the stressed vowel, while it lenites to a flap in *pentagon* because it precedes an unstressed vowel (and additionally coalesces with the nasal to yield a nasalized flap). The proper treatment of alternations, wherein the 'same' phonological unit varies according to its context, is also a core concern of phonological theory.

2.1.3. Opacity

Opacity arises when the surface evidence for a phonological process is inconsistent. For example, Baković (2007) gives the following, well-known example from Yokuts Yawelmani:

(2)   UR   /ʔili:+l/
       Long High Vowel Lowering |ʔile:l|
       Closed Syllable Shortening |ʔilel|
       SR   [ʔilel]

Evidently, the Long High Vowel Lowering process serves to avoid long high vowels, a marked outcome which never appears on the surface in this language (even though many URs contain underlying long high vowels). The Closed Syllable Shortening process is similarly motivated by the observation that long vowels never co-occur with coda consonants. The 'problem' in (2) is that there is no reason for both processes to apply. Closed Syllable Shortening alone would avoid both marked structures, but Long High Vowel Lowering appears to apply anyways, gratuitously 'hiding' the underlying height of the vowel. Opacity, or at least certain types of opaque patterns, are believed to present a significant learning problem.

2.1.4. The Sound Pattern of English (SPE/Rules)

Chomsky and Halle (1968) proposed a phonological analysis of English using *string rewrite rules* of the form AXB → AYB, typically abbreviated X → Y / A__B and

read out loud as 'X goes to Y when it occurs after A and before B'. The formal mechanisms they introduced –including the treatment of segments as 'feature bundles', to which rules could refer, and language-specific rule orderings– became the dominant paradigm within the field of phonology for many years afterwards. Even as constraint-based formalisms have replaced *SPE*-style rules as the preferred vehicle for phonological analysis, many linguists still use rules as a convenient shorthand for describing phonological processes, e.g. in (2) above.

2.1.5. Optimality Theory

Optimality Theory, like *SPE*, defines the phonological grammar as a cognitive mechanism which implements the mapping from an input/UR to an output/SR, and may make reference to 'hidden' phonological structure such as metrical feet, syllables, etc. Unlike *SPE*, OT posits that there are multiple possible candidates for a given input, and there is a parallel computation to identify the optimal ('most harmonic') output candidate, rather than the serial/derivational process or ordered rule in *SPE*. Seminal works on OT (McCarthy & Prince, 1994; Prince & Smolensky, 1993, 2002, 2004; Smolensky & Legendre, 2006) define the core components of a broad class of constraint-based theories: there must be a component which proposes output candidates (GEN), a set of constraints (CON), and an evaluation/selection mechanism (EVAL) which chooses the winning candidate based on some language-specific prioritization of the constraints. Some authors use "OT" to refer to refer broadly to any such constraint-based theory of phonology. I will use "OT" to refer to the subclass of constraint-based theories with the "total ordering" evaluation method described in Prince and Smolensky (1993, 2002, 2004) and McCarthy and Prince (1994). That is, for the purposes of this article, "OT" means that constraint conflicts are resolved in favor of the highest-ranked constraint, regardless of whether the winning candidate incurs more violations of lower-ranked constraints than alternate candidates. (Constraint conflict arises for particular inputs when it impossible for an output to satisfy one constraint without violating another. An example is shown below in (3).

2.1.6. Harmonic Grammar

Later in the article, article, I will make frequent reference to MaxEntHG (Goldwater & Johnson, 2003; Hayes & Wilson, 2008), a probabilistic extension of Harmonic Grammar (Legendre, Miyata, & Smolensky, 1990; Smolensky & Legendre, 2006) in the log-linear framework. As the reader may not be familiar with Harmonic Grammar, I describe it very briefly here. Harmonic Grammar is a close variant of OT which differs in the evaluation procedure: constraints are weighted, rather than totally ordered, and the harmony

of a form is determined by the weighted sum of its constraint violation. As with OT, this is straightforwardly illustrated with a tableau; an example of word-final devoicing is shown in (3):

(3) /gad/	$Ident_{Vce}$[-son] wt = -1	*[-son,+vcd]$_{]PrWd}$ wt = -5	Harmony
[gad]		*	-1·0+-5·1=-5
☞ [gat]	*		-1·1+-5·0=-1

The UR /gad/ is given in the top left cell, while candidate SRs are listed below. Constraint names are given in the top row after the input; IDENT$_{VCE}$[-SON] penalizes obstruents for which the output voicing value does not match the underlying voicing specification, while *[-SON,+VCD]$_{]PRWD}$ penalizes voiced obstruents at the end of a word. Constraint violations are marked in the cells with an '*'. For inputs with underlyingly voiced final obstruents, it is impossible to satisfy both constraints at once; thus this is an example of constraint conflict. The constraint weights are listed directly underneath the constraints themselves, and are required to be nonpositive.[1] The final column indicates the *harmony* value of the output candidate, defined as the weighted sum of the constraint violations. As with OT, the most harmonic output candidate (or, equivalently, the least disharmonic) is selected as the winner; this is conventionally indicated with the "OT hand" ☞. In cases where only two constraint violations trade off against one another, Harmonic Grammar is equivalent to OT; however, the two theories make different predictions when a single constraint violation conflicts with multiple violations of a different constraint (*counting cumulativity*) or violations of multiple constraints (*ganging cumulativity*).

2.2. Probability

I assume the reader is familiar with elementary statistics and probability theory. For example, I assume the reader is familiar with the concept of *p*-value, *t*-test, and use of the binomial formula to calculate the probability of a series of coin tosses. I also assume the reader is familiar with exponentiation and the inverse operation of taking the logarithm. Below I describe the concept of odds, and briefly outline log-linear models.

2.2.1. Odds and log-odds

The odds of two events, sometimes written *a:b*, indicate the relative probability of the two events. For example, if the odds are 3:2 that Lucky Horse will win the race, it means that Lucky Horse is expected to win 3 times for every 2 times that Lucky Horse does not win. Odds can always be converted to probabilities and vice versa; for example 3:2 means that Lucky Horse will win 3 times out of 5 trials for a probability of 3/(3+2) = 0.6. Odds can be represented as single numbers by simple division, e.g. 3:2 = 3/2 = 1.5. Thus, when there are only two possibilities, an odds of 1.5 corresponds to a probability of 0.6. The *log-odds* of two events A and B is simply the logarithm of their odds, i.e. *log(a:b)*. (In general, I will assume the natural logarithm unless otherwise specified.) The log-odds has several intuitively attractive properties. It is zero when A and B are equiprobable, positive when A is more probable than B, and negative when A is less probable than B. Moreover, the greater the asymmetry in probability between A and B, the greater the magnitude of the log-odds. Finally, in many of the systems where log-odds are used, probability differences can be many orders of magnitude. The log operation makes the relative likelihood of these outcomes easier to grasp for normal readers.

2.2.2. Log-linear models

Log-linear models express the probability of input-outcome pairs in terms of some feature functions and associated weights. The *score* $H_{M(w)}$ of an input-outcome pair is the weighted sum of its feature values. The output of the model $M(\mathbf{w})$ is then determined by stipulating that *the probability of an input-outcome pair is proportional to the exponential of its score*. Formally, a log-linear model $M(\mathbf{w})$ consists of a vector of feature functions $\mathbf{f} = \{f_k\}$ and a relation GEN which gives the set of possible outcomes $\mathbf{y}_{ij}$ for each input $\mathbf{x}_i$. In addition, the vector $\mathbf{w}$ is a parameter of M, and represents the weights that are associated with the feature functions:

$$(4)\ Pr_{M(w)}(\mathbf{y}_{ij} \mid \mathbf{x}_i) = \exp(H_{M(w)}(\mathbf{x}_i, \mathbf{y}_{ij}))/Z(\mathbf{x}_i)$$

$$H_{M(w)}(\mathbf{x}_i, \mathbf{y}_{ij}) = \Sigma_k w_k \cdot f_k(\mathbf{x}_i, \mathbf{y}_{ij})$$

$$Z(\mathbf{x}_i) = \Sigma_{y[i,j']\in GEN(x[i])} \exp(H_{M(w)}(\mathbf{x}_i, \mathbf{y}_{ij'}))$$

Log-linear models have several attractive computational properties. One of them is that it is easy to interpret the *relative* probability of two different outputs: for a given input $\mathbf{x}_i$, the log-odds of outcome $\mathbf{y}_{ia}$ versus $\mathbf{y}_{ib}$ is simply the difference in their scores $H_{M(w)}(\mathbf{x}_i, \mathbf{y}_{ia})$–$H_{M(w)}(\mathbf{x}_i, \mathbf{y}_{ib})$. Another, especially important property is that for fixed $\mathbf{f}$ and GEN, only mild assumptions are

---

[1] Some authors instead require that weights be nonnegative. The formalism itself requires that the weights all be the same sign. (Otherwise, the theory would not exhibit harmonic bounding, and would lose the desirable typological restrictiveness that comes from an explicit theory of markedness.) I prefer negative weights, since this aligns intuitively with the definition of harmony: candidates with more constraint violations are less harmonic. As we shall see later, negative weights also aligns with the natural extension of Harmonic Grammar to a log-linear model.

needed to ensure that the probability of a dataset $\mathbf{X} = \{(\mathbf{x}_i, \mathbf{y}_{ij})_{i=1..N}\}$ is *convex* in the space of all possible weight vectors (Berger, Della Pietra, & Della Pietra, 1996). In more everyday language, this means two things. First, there is a unique 'best' weight vector ($\mathbf{w}_{max}$) which maximizes the likelihood of the observed data. Second, it is possible to *find* this unique best solution in a computationally efficient manner, using well-established numerical techniques like (conjugate) gradient ascent. As we will see later, log-linear models offer a natural probabilistic extension for Harmonic Grammar, which offers the exciting potential for a theory of phonological learning that is machine-implementable and testable on natural language data.

This completes the survey of background material. The next section begins the body of the paper. In that section, I briefly survey the field known as 'formal language theory', whence modern linguistics began.

3. FORMAL LANGUAGE THEORY

Formal language theory is an axiomatic, logical/mathematical approach to language. A 'language' is defined as a set of strings, often according to some process that generates the set. Researchers who work in this area are concerned with the classification of languages according to the 'complexity' of the process required to generate the language, as well as the assumptions needed to learn languages in the various classes identified. Two of the best-known concepts to have emerged from this line of research are the Chomsky-Schützenberger hierarchy (Chomsky, 1956) and the concept of identification in the limit (Gold, 1967), both of which will be briefly covered later. Two strands of work in this line of special relevance to phonology include comparisons of the expressive power of different phonological frameworks (e.g. Buccola & Sonderegger, 2013; Graf, 2010ab; Jardine, in press) and the elaboration of finite-state techniques which 'count' constraint violations for entire classes of strings, enabling efficient machine optimization (e.g. Eisner, 2002; Hayes & Wilson, 2008; Riggle, 2009).

As this material is somewhat technical and unlikely to be known to the average linguist, I begin with an overview of basic concepts. Furthermore, because the article aims to cover other topics besides just formal language theory, the overview is necessarily somewhat superficial; it is meant to describe the intuitions, the most common notation, and the most widely cited results. Readers who are already acquainted with this material may wish to skip directly to the *Framework comparison* subsection. Conversely, readers who wish to learn more are advised to peruse a source devoted to formal language theory: Heinz (2011ab) for phonology specifically, Stabler (2009) for a survey of formal language theory as it relates to natural language universals, or an introductory computer science textbook for the basics.

3.1. General overview

In formal language theory, 'language' does not refer to a shared linguistic code like English or Amharic or Tashliyt Berber. Rather, it is a formal object with precisely specified properties, which can be studied in a mathematical, axiomatic, logical fashion. Conventionally, formal language theory assumes an *alphabet* Σ and defines a *string* as an ordered sequence of elements from Σ. For example, if $\Sigma = \{a, b\}$ then $\sigma = ab$ is a (rather short) string over Σ. The set of all possible strings over Σ is denoted Σ^* (where * is called the Kleene star and has the conventionalized meaning of "0 or more repetitions"). Normally in formal language theory, a language L is defined as a subset of Σ^*. Note that the elements of the alphabet do not have any intrinsic meaning, or any internal structure; they are simply algebraic elements that are distinct from one another.

For example, we could define $\Sigma = \{C, V\}$ and $L = (CV)^+$ (where $^+$ means "1 or more repetitions"); the resulting set of strings would look to a phonologist like a strict CV language: {CV, CVCV, CVCVCV, ...}. But the formalism does not *know* that C means consonant and V means vowel in the same way that human speakers do. Humans know that vowels are characterized partially complementary articulatory and acoustic properties, as well as sequencing facts (e.g. words must begin with a C, every C must be followed by a V, V can end a word or be followed by a C). The formalism merely knows the sequencing facts, and that C is a different symbol than V. Indeed, the language $L = (ab)^+$ over $\Sigma = \{a, b\}$ has the same abstract structure as $L = (CV)+$ over $\Sigma = \{C, V\}$; from a formal language perspective, these are *notational variants*, meaning that they express the same pattern after a transparent, structure-preserving change in notation.

Interest in formal language theory is motivated by the assumption that natural languages can be mapped onto some particular class of formal languages (or vice versa), and that the properties of the formal language class will yield clear insights into how language is learned, represented and computed in the minds of speakers. For example, it is widely believed that syntax is mildly context-sensitive, while phonology is (sub)regular (e.g. Heinz, 2011ab; Stabler, 2009). We will unpack this assertion later. In the meantime, it must be acknowledged this idea of identifying languages with stringsets, and dividing them up into classes based upon certain properties, is a large assumption, whose full implications we do not have space to assess here. I will point out one implication here however: the literature on learning formal languages ('learnability') assumes that knowledge that is 'outside' the grammar is not brought to bear on grammar learning. For example, phonetic knowledge does not figure in the formal language learnability literature on phonology, just as semantic/pragmatic knowledge does not figure in the learnability literature on syntax. With this kind of caveat in mind, let us consider what formal language theorists mean by a language class.

3.2. The Chomsky Hierarchy

It may be helpful to begin with an example. Chomsky (1956) describes a way of generating strings that is now known as a phrase-structure grammar. Phrase-structure grammars are predicated on a system of "rewrite rules", in which one string is rewritten as another. Here is an example of an especially simple phrase-structure grammar:

(5) rewrite to nonterminals: $\underline{S} \to$ NP VP
  VP $\to$ V NP
  rewrite to terminals: V $\to$ likes
  NP $\to$ the boy
  NP $\to$ the dog

The $\underline{S}$ symbol is underlined to indicate that it is the unique start symbol. This grammar generates strings by beginning with the start symbol, and generating all possible outputs by applying any rule that can apply, at any time. For example, this grammar generates the string *the boy likes the boy* by rewriting '$\underline{S} \to$ NP VP', 'NP $\to$ the boy', 'VP $\to$ V NP', and 'NP $\to$ the boy' again. The sequence of rewrite operations, together with the final output of the derivation, has an elegant visual representation as a tree:

(6)

```
              S
         ┌────┴────┐
        NP         VP
         │      ┌───┴───┐
         │      V       NP
         │      │        │
      the boy likes   the boy
```
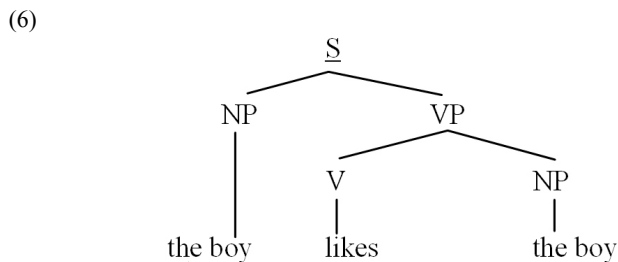
The grammar in (5) is simple enough to enumerate the entire language it generates: *the boy likes the boy, the boy likes the dog, the dog likes the boy, the dog likes the dog.*

More formally, a phrase-structure grammar G consists of a start symbol $\underline{S}$, a set of terminal symbols Σ, a set of *non*terminal symbols V (which must not share any symbols with Σ), and a collection of rewrite rules R, where each rewrite rule maps a sequence containing a nonterminal to a sequence of terminals and nonterminals. The language generated by such a grammar is defined as the set of strings generated by all derivations that terminate (i.e. strings containing only terminal symbols).

3.2.1. Context-sensitive languages

The set of languages that can be generated when the rewrite rules are only unrestricted to not increase the number of symbols is called the *context-sensitive* class. It is possible to define context-sensitive languages which are completely unlike natural languages, for example
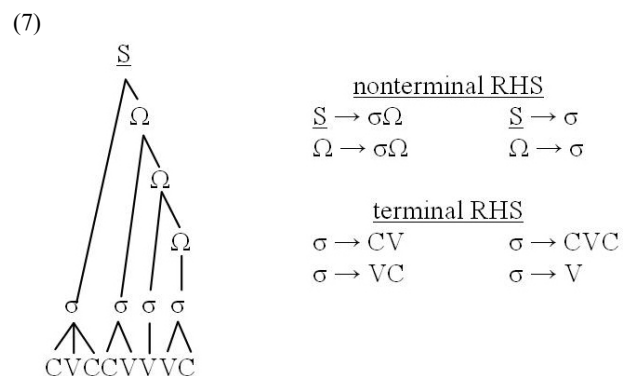
languages in which if the grammar generates a sentence $X = x_1 x_2 \dots x_n$, it also generates the mirror-image $X' = x_n x_{n-1} \dots x_1$. Natural languages exhibit certain kinds of regularities, such as constituency structure, which are not expected if rewrite rules are completely unrestricted. Therefore, the class of context-sensitive languages is 'too rich'; it does not explain the structural constraints that natural languages have.

3.2.2. Context-free languages

Chomsky (1959) defined the *context-free* class as the set of languages which can be generated by a grammar in which the left-hand side of every rewrite rules is a single nonterminal. In other words, the rewrite rules *substitute a unique nonterminal for something else* –crucially, without regard to what surrounds the nonterminal. Grammar (5) is an example: every rewrite rule contains a single nonterminal on the left-hand side. Intuitively, this means that the eventual output that corresponds to a nonterminal cannot 'look outside' the nonterminal itself. In other words, context-freeness imposes a type of *locality* restriction on how substrings may share dependencies. This is one means of enforcing constituent structure in context-free languages.

3.2.3. Regular languages

The *regular* languages are those that can be generated by rewrite rules in which the left-hand side consists of a single nonterminal, and the right-hand side may contain at most one nonterminal. Moreover, the nonterminals on the right-hand side of the rewrite rules must always be final in the rewrite string (in which case the language is called *right regular*), or must always be initial in the rewrite string (in which case the language is called *left regular*). Here is an example of a right regular grammar and a string that it generates:

(7)

```
    S
    │╲
    │ Ω
    │ │╲
    │ │ Ω
    │ │ │╲
    │ │ │ Ω
    │ │ │ │
    σ σ σ σ
   ╱╲ │ ╱╲
  CVCCVVVC
```

nonterminal RHS
$\underline{S} \to \sigma\Omega$ $\underline{S} \to \sigma$
$\Omega \to \sigma\Omega$ $\Omega \to \sigma$

terminal RHS
$\sigma \to CV$ $\sigma \to CVC$
$\sigma \to VC$ $\sigma \to V$

The symbols in (7) were chosen suggestively, to illustrate to readers how formal languages might encode structures and relations used in mainstream phonology.

Now we are in a position to understand the contrasting claims that "phonology is (sub)regular" while

"syntax is mildly context-sensitive". The former phrase expresses the belief that for every natural language L, there is a grammar G_L which can generate all and only the licit phonological strings of L, and G_L can be written as a regular grammar (possibly even as some proper subset of the regular languages). The latter phrase expresses the belief that this is not true for syntax, since there exist syntactic patterns which (it has been claimed) cannot be captured by regular rewrite rules, or even context-free rewrite rules. For example, Shieber (1985) gives the following Swiss German clause as an example of a cross-serial dependency:

(8) *...mer em Hans es   huus   hälfed aastriiche.*
   *...we Hans-DAT the house-ACC helped   paint.*

     We helped Hans paint the house.

Languages which admit of an arbitrary number of such dependencies are probably non-context free, and Shieber argues that Swiss German is just such a case.

3.3. Equivalency of Finite State Automata and Regular Languages

An overview of formal language theory would not be complete without mention of finite state machines (FSMs, also called FSAs for finite state automata). Practically speaking, an FSM is an alternative representation of a regular language. Historically, the two were conceived of separately, but the formal equivalence was noted and proved in early work. An FSM consists of a set of states, conventionally indicated with circles and an optional state label. In addition to the states, an FSM contains transitions between states, which must be labeled in most formulations. At least one state is designated as a start state, and at least one state is designated as an end state; these may be the same state. Conventionally, the start state is indicated with a thick circle, while other states are indicated with a single circle. Here are two examples:

(9)
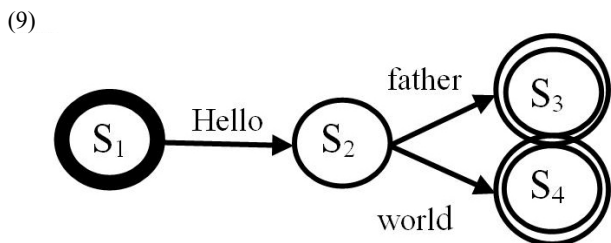
Finite state machines can be considered as generators or parsers, but either way, they describe the same set of strings. Example (9) describes exactly two strings: *Hello father*, and *Hello world*. Example (10) describes an infi-

nite number of strings, including *This is the cat, This is the cat that chased the rat, This is the cat that chased the rat that ate the cheese, This is the cat that ate the cheese that chased the rat*, etc. In generation mode, the FSM works by beginning at the start state. If it is at an end state, it may stop, having generated a complete string. If there are one or more transitions out of the current state, the machine selects one randomly and follows it, emitting a symbol along the way (the label on the transition). However, when there is only one transition out of a non-end-state, the machine must follow that unique transition. In parsing mode, the machine is said to consume symbols from an input string. It begins at the start state. When it receives the next symbol from the input string, it looks for a transition with a matching label. If there is a matching transition, it follows it and advances to the next input symbol. If there is not a matching label, the machine is said to reject the string. If the machine is in an end state when the input string is entirely consumed, the machine is said to accept the string; otherwise the machine rejects the string. In other words, the FSM accepts the string if and only if it can match every symbol in the input string with a transition and end up in an end state when the input is consumed.

(10)
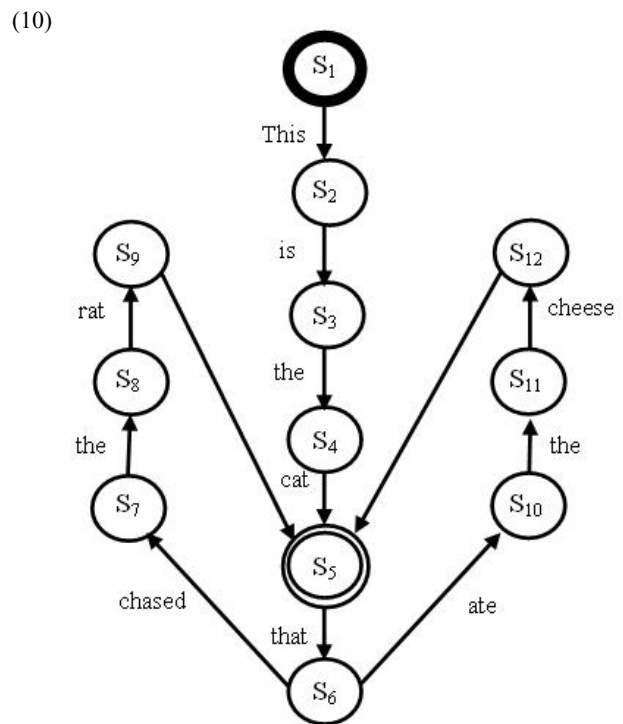
An example of a string that (10) does not accept is *This was the cat that chased the rat*. Initially, the machine is in the start state (S_1). The first input symbol, *This,* is presented and matches the label for the transition *leading* out of S_1 and into S_2, so *This* is consumed and the machine enters state S_2. Now the input symbol *was* is considered, but the only available transition label is

is, so the machine rejects the string. In formal language theory, accepting or rejecting a string is akin to offering a grammaticality judgment. Languages are defined as sets of strings, so in principle it is possible to make a binary judgment for every string, whether it is in the language. In the case of regular languages, there is guaranteed to be a finite state machine which can not only do this is principle, but can do so straightforwardly and efficiently in a computer implementation. For this reason, finite state methods have been applied throughout computer science, for both natural language processing and various other applications (such as programming language parsing and compilers).

It is worth noting here that there are alternative formulations of finite state machines. For example, it is possible to make the state labels correspond to symbols being generated/consumed, while the transitions are unlabelled. It is also possible to augment the transition and/or states with extra information, beyond the symbol being consumed/generated. Indeed, there is a great deal of work on this topic, which is omitted for space reasons.

A final type of finite state automata is known as a *finite state transducer* (FSTs). An FST is just like an FSM, except that it parses an input string *and* generates a corresponding output string. That is, the FST behaves just like a FSM in terms of parsing, but its transition labels have been augmented; the label consists of both the input symbol to match, and an output symbol to generate upon a successful match. Here is an example which implements an intervocalic lenition rule (d → ð / a__a):

(11)
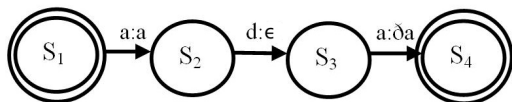
The symbol ∈ is a special symbol, conventionally used to indicate an empty output. The finite state transducer in (11) will first match an /a/ and output an [a]; then it will match a /d/ but output nothing, waiting to see if it gets another /a/. If it gets another /a/, it will then output the 'delayed' [ð] along with the [a]; otherwise, the FST will reject the string, indicating that the lenition rule does not apply to this input.

3.4. Identification in the limit, and other notions of learnability

Gold (1967) provided the first formalization of learnability for a formal language. In Gold's conception, the input to a learner is defined as a *text T* – an infinite sequence $(t_1, t_2, ...)$ of grammatical items from a *language L*, which is guaranteed to contain every item in *L* at least once, but not in any particular order. A *grammar G(L)* is defined as a finite representation that can generate all and only the strings of *L*. A *learner A* is defined as a function which accepts a finite subsequence

$T_n = (t_1, t_2, ..., t_n)$ from a text *T* and returns a hypothesized grammar. For example, $A(T_5)$ is the grammar that learner *A* would posit after hearing the first 5 sentences of *L* in text *T*. By feeding a learner *A* successively longer subsequences from an input text *T*, we obtain a sequence of posited grammars $A(T_1), A(T_2), ...$ A learner is said to identify L in the limit if for every text *T*, there is a finite amount of input *N* such that $A(T_N) = G(L)$, and $A(T_m) = A(T_N)$ for all $m > N$. In other words, the learner *A* is said to identify *L* in the limit if they are guaranteed to converge on a grammar that generates *L* in a finite amount of time.

Prior to presenting Gold's main result, it is worth considering how this framework compares with the child's learning situation. In the framework described above, a learner has access only to *positive evidence*, that is, only to sentences which are actually in the language. This is now referred to as *unsupervised learning*, since the learner does not have access to an external metric or 'objective function' which unambiguously indicates the nature of the solution to be learned. (Gold also considered *supervised learning*, in the form of an informant who presents both sentences from the language and sentences not from the language, while indicating which is which.) It is generally believed that children acquire the syntax of their language from positive evidence only, and tend to ignore the negative evidence they do get (R. Brown, 1973). On the other hand, Gold's framework does not allow for 'meaning', either the semantic meaning of the words that sentences hear, or the 'phonetic' meaning of the phonemes that make up those words. Moreover, Gold's notion of identification in the limit does not impose any constraints upon the input text, such as some kind of 'representativeness' criterion. That is, it is a safe bet that the words 'momma' or 'mother' appear in the first million words that every English-acquiring infant hears, but there is nothing in Gold's formulation which requires input texts to exhibit this kind of real-world distributional property. Thus, Gold's assumptions line up with the child's learning situation in one way, but differ from it in other ways.

There were two key results in Gold (1967). The first is that the class of regular languages is not identifiable in the limit. The second was that regular languages (and even higher classes in the Chomsky hierarchy) *are* learnable in the limit from an informant, i.e. supervised learning with positive and negative examples. Since phonology is believed to be (sub)regular, and syntax is believed to be at least context-free, and it is widely believed that children do eventually learn the correct grammar for their language, this result is interpreted by many theorists as proving that children possess innate constraints on the space of hypotheses that they consider as grammars for their language.

This conclusion does not actually follow from Gold's theorem. In general, one can only reason 'backwards' from a model to reality when one is confident that the model is an accurate portrayal of the reality it is modeling. That is, modeling results depend on a host of assumptions; they are akin to a logical proposition of the form, 'If A and B and C and D, then X'. We cannot conclude from the truth

of X that A and B and C and D are true. Moreover, we cannot conclude from the falsity of X that a particular assumption (e.g. C) is false; we can only conclude that *some* assumption is false. So from Gold's theorem, what we can conclude is quite limited. It *could* be that children are born with innate constraints on the grammars that they consider. But it also could be that the input is far more constrained than Gold assumes. It could be that children leverage multiple types of information (such as semantics and phonetics) in language acquisition, and that this provides extra constraints on the space of possible grammars. It could be that human grammars are not strictly comparable with the classes of string-generators that Gold considers. It is possible that humans do not actually converge a single final grammar state, and actually do update their grammars on the basis of new input throughout the lifespan. These possibilities are all compatible with Gold's theorem.

Close inspection of the proof for Gold's theorem reveals that it depends crucially on the order in which examples are presented. Gold shows that it is possible to construct a text which continually forces the learner to update their hypothesis, because the class of regular languages is rich enough that one can 'maliciously' deny crucial evidence to the learner ad infinitum. Valiant (1984) introduced a probabilistic framework for studying (machine) learning known as *probably approximately correct* (PAC). Abstracting away from the technical details, the key difference is that texts are required to be 'representative', in the sense that training examples must be drawn from a probability distribution, and the learner is counted as 'approximately correct' if its generalization errors on this distribution fall below an arbitrary threshold δ (which can be made as small as desired, as long as it is still greater than 0). A language class is said to be PAC-learnable if a learner can identify an 'approximately correct' language in the hypothesis space from a finite sample of the target language. It is *efficiently PAC-learnable* if there is an algorithm which is guaranteed to do this while requiring a number of examples that is polynomial in the size of the language. Kearns and Valiant (1994) show that regular languages are not efficiently PAC-learnable, while Li and Vitányi (1991) show that regular languages *are* efficiently PAC-learnable under the additional assumption that 'simple' examples are more likely to be drawn than complex ones (as assessed by a measure called Kolmogorov complexity).

Researchers have interpreted these learnability results in many ways. Some researchers believe that the way forward is to develop increasingly fine-grained specification of the assumptions and increasingly fine-grained classifications of the classes of languages. Some researchers believe that this kind of work simply has no bearing on the learning problem that children actually face. One generalization that many parties can agree to is that the learnability results offered so far are *fragile*, in the sense that seemingly small changes in the assumptions can result in large changes in the nature of the conclusion (while intuitively similar changes may also yield no meaningful difference). Thus, one way to view the careful work done by Gold, Valiant and others is as an ongoing attempt to characterize which assumptions actually matter for learnability.

There is a vast amount of work on formal language theory and learnability that cannot be surveyed here; I trust the presentation above was detailed enough to give the lay reader a sense of what formal language theory aims to accomplish. In the remainder of this section, I turn to two new lines of work in formal language theory with the potential to inform basic questions in phonology. The first concerns what might be called *framework comparison* –a methodology for comparing two distinct linguistic formalisms via the formal languages they generate. The second concerns the use of finite-state techniques for efficient implementations of constraint-based phonology, which I will refer to as *finite-state OT*.

3.5. Framework comparison

Modern linguistics has taken seriously the task of formalizing theoretical intuitions. From seminal works to the modern day, theorists are apt to propose new frameworks like *SPE* (Chomsky & Halle, 1968) and OT (Prince & Smolensky, 1993, 2002, 2004), or non-trivial departures from existing frameworks, such as autosegmental phonology (Goldsmith, 1976, 1990; McCarthy, 1981), Harmonic Serialism (McCarthy, 2008, 2011), and others. The formalist bent has paid off in theoretical precision: so long as the linguistic atoms and operations are specified, the reader of a paper can make new predictions from a theory which the original writer would agree with. This kind of precision enables rapid progress, and surely reduces the frequency and severity of fruitless debates in the field over misinterpretations of a theory. Still, as pointed out in Stabler (2009), the proliferation of theories does come with a cost. In many cases there are competing formalisms, but since the surface character of the explanation is so different, it is difficult to tell whether the theories actually make different predictions.

Formal language theory offers a way to directly compare the expressive and restrictive powers of two different frameworks. This kind of work is already well-established in the syntactic domain, as evident from the following quotation in a review by Stabler (2009):

> In the work of Joshi, Vijay-Shanker, and Weir (1991), Seki et al. (1991), and Vijay-Shanker and Weir (1994) four independently proposed grammar formalisms are shown to define exactly the same languages: a kind of head-based phrase structure grammars (HGs), combinatory categorial grammars (CCGs), tree adjoining grammar (TAGs), and linear indexed grammars (LIGs). Furthermore, this class of languages is included in an infinite hierarchy of languages that are defined by multiple context free grammars (MCFG), multiple component tree adjoining grammars (MCTAGs), linear context free rewrite systems (LCFRSs), and other systems. Later, it was shown a certain kind of "minimalist grammar" (MG), a formulation of the core mechanisms of Chomskian syntax –using

the operations merge, move, and a certain strict 'shortest move condition'– define exactly the same class of languages (Michaelis, 2001; Harkema, 2001; Michaelis, 1998). These classes of languages are positioned between the languages defined by context free grammars (CFGs) and the languages defined by context sensitive grammars (CSGs) like this.

The works cited by Stabler indicate that despite the surface differences between formal frameworks, they are sometimes "notational variants", in the deep sense that they describe the same set of languages. The details of these proofs are beyond the scope of this article, but the general nature of the argument is clear: provide a schema for translating one formalism into a particular kind of logic, which can be expressed as a formal language. Then do the same for the other formalism, and show that the two resulting formal languages are the same (or different) according to known properties of the formal language. In the view of the researchers who do this work, formal language theory has a certain potential to tell us what our formal mechanisms are actually buying for us.

Kaplan and Kay (1994) arguably supplied the first such example of this line of work in phonology. They proved that the rule-based rewrite system presented in *SPE* belongs to the class of regular languages, by embedding it in a class of logics known as Monadic Second Order (MSO) logics, known to be equivalent to the regular languages. More precisely, Kaplan and Kay claimed that *SPE* was regular even with 'cyclical rules' that are allowed to feed its own environment, as long as they are forbidden from feeding their own targets (for discussion and clarification see Kaplan & Kay, 1994). Potts & Pullum (2002) did essentially the same thing with OT, embedding a class of OT constraints into MSO. In addition, Potts & Pullum demonstrated that particular classes of OT constraints that had been proposed (e.g. align constraints) exceeded the power of regular languages, and in some cases they proposed regular alternatives.

Graf (2010ab) compared a formalism known as Government Phonology with *SPE*. For readers not already familiar with Government Phonology, Graf (2010a) very readably points out the vast surface differences between it and *SPE*:

> GP as defined in Kaye et al. (1985, 1990) and Kaye (2000) differs from SPE in that it uses privative features (features without values) rather than binary ones, assembles these features in operator-head pairs instead of feature matrices, builds its structures according to an elaborate syllable template, employs empty categories and allows all features to spread (just like tone features in autosegmental phonology). (p. 83)

Graf begins by translating each of these formalisms into a kind of propositional logic. Like Kaplan and Kay (1994), Graf embeds *SPE* in MSO. Graf goes on to show that if Government Phonology allows unbounded feature spreading, it can be embedded in MSO; if it allows only bounded spreading, it can be embedded

in a strictly less expressive logic. In other words, Graf argues that despite the many differences between these formalisms, the property that really matters is bounded vs. unbounded spreading, since with unbounded spreading the two theories can express the same languages.

Graf goes on to address the 'empirical bite' of this theory by asking whether any natural phonological phenomena do require unbounded feature spreading. He proposes two candidates –Sanskrit *nati* and Cairene stress assignment. According to Graf, the *nati* rule causes an underlying /n/ (the TARGET) to become retroflexed if it is the first postvocalic /n/ after a continuant retroflex consonant (/ɻ/ or /ʂ/; the TRIGGER), provided that no coronal intervenes between the trigger and target, that the nasal target is immediately followed by a nonliquid sonorant, and that there is no retroflex continuant in the string after the target. As for Cairene stress assignment, the rule is to stress the final if it is superheavy or the penult if it is heavy. If both the final syllable and the penult are light, the rule is to stress the penult or the antepenult, whichever is an even number of syllables from the closest preceding heavy syllable. This suggests the presence of an 'invisible' trochaic footing system, in which secondary stresses propagate in an iterative/bounded manner from the rightmost heavy to the penult or the antepenult. Of course, as Graf points out, the ability to analyze bounded/iterative spreading of an 'invisible' feature is empirically impossible to distinguish from unbounded spreading of a visible feature. Therefore, he proposes to ban bounded/iterative spreading of invisible features for the purposes of theory comparison. This suggests that unbounded feature spreading is required –an important theoretical claim.

Two other recent studies in this line of research are Jardine (in press) and Buccola and Sonderegger (2013). Jardine (in press), following directly from Graf (2010ab), asks whether autosegmental phonology belongs to the same class as *SPE*. Jardine does not give a complete answer to this question, owing to phenomena such as floating tones and dissociation rules. However, Jardine does show that MSO is expressive enough to cover the 'simple' phenomena that initially motivated autosegmental phonology, such as rightward feature-spreading. Buccola and Sonderegger address Canadian raising, an opaque phonological pattern in which allophonic variation is triggered by an underlying contrast that is erased at the surface:

(12) a. Raising before voiceless consonants

ride	/ɹaɪd/	[ɹaɪd]
write	/ɹaɪt/	[ɹʌɪt]

b. Foot-medial tapping

batter	/bætɚ/	[bæɾɚ]
badder	/bædɚ/	[bæɾɚ]

c. Interaction

ride	/ɹaɪɾɚ/	[ɹaɪd]
write	/ɹaɪɾɚ/	[ɹʌɪɾɚ]

Patterns like (12) are easily captured by a rules-based analysis in which the tapping rule applies after the Canadian raising rule. It is generally believed that such patterns cannot be accommodated by 'normal' OT, and a considerable body of work has been devoted to accommodating the theory to this type of pattern. What Buccola and Sonderegger show is that for any version of OT in which there is a single stratum (that is, one input representation and one selected output representation, with no intervening representational levels subject to competition), and in which faithfulness constraints assess the relationship only between an input segment and its output correspondence (i.e. without reference to its neighbors), the OT theory is strictly unable to account for the opacity pattern. They show this by translating OT constraints into finite-state transducers, in the manner proposed by Riggle (2004) and described in the next subsection. However, Buccola and Sonderegger also acknowledge that the highly related formalism of Harmonic Grammar (in which constraint competitions are resolved through linear combinations rather than strict domination) actually can accommodate cases like Canadian raising, without allowing for so-called positional faithfulness constraints. (Finally, there is an analysis in which the [ʌɪ]~[aɪ] contrast is treated as phonemic, although many linguists disprefer this, since it requires stipulating that [ʌɪ] is only licensed before coronal obstruents and [ɾ], and crucially fails to link this fact to the nearly complementary distribution of [aɪ].)

In summary, formal language theory has begun to deliver on the promise of framework comparison in phonology. If one is willing to accept the premise that a language is a set of strings, this kind of technically exacting work has the capacity to reveal surprising equivalences between formalisms, and to zoom in on key properties which distinguish expressivity. Still, it must be acknowledged that existing work seems to depend sensitively on details of the analysis which are not themselves rock-solid. For example, the claim that syntax is not context-free rests on phenomena like cross-serial dependencies, and more specifically on the claim that Swiss German allows an unbounded number of them. In practice, it is likely quite rare for natural usage to yield more than 1 crossing dependency. While Graf's (2010a) work does not strictly depend on whether unbounded feature spreading actually occurs in phonology, it does suggest that this is a critical distinction phonologists should attend to. However, as he acknowledges, the two putative cases he gives have been contentious in the literature. Buccola and Sonderegger (2013) discuss Canadian raising and more generally counterfeeding on environment (Baković, 2011) and seem to endorse a rules-based approach, but there are alternative analyses that do not require ad hoc modifications to existing theories.

In conclusion, formal language theory offers a rigorous, string-based and axiomatic approach to phonology as a formal system. Many researchers believe that this kind of logic- or model-based approach to phonology is the key to discovering what formal properties of our frameworks make for meaningful contrasts in empirical coverage and restrictiveness. Other researchers are uneasy with this approach, a feeling which Stabler (2009) aptly summarized thusly:

> But many linguists feel that even the strong claim that human languages are universally in the classes boxed in (1) is actually rather weak. They think this because, in terms of the sorts of things linguists describe in human languages, these computational claims tell us little about what human languages are like. (p. 203)

Looking back over the works reviewed above, it is clear that the formal language approach has relatively little to say about markedness, alternations, opacity, or many other core concerns of mainstream theoretical phonology.

For example, one of the most appealing aspects of constraint-based grammars is that they formally encode a substantive bias against marked structures, by directly including markedness constraint in the theory. Indeed, the success of OT in predicting the typology of syllable structures arises from the combination of an ONSET constraint (which punishes words that begin with a vowel) with a formal property called *harmonic bounding* (if candidate B is equal or worse on every dimension than candidate A, B can never win). OT thereby predicts the existence of languages which require words to begin with a consonant, and of languages which allow words to begin with a vowel, while correctly predicting the absence of languages which require words to begin with a vowel. But there is nothing about "regularity" which forces this property. Rather, it is part of the substantive content of the theory. Formal language theory simply has nothing to say about it.

In any case, it is clear that the formal language theory approach to framework comparison has just begun to affect phonology. There will be more of this work in the near future, not less. The eventual theoretical impact of this line of work cannot be determined yet, and is likely to depend on the extent to which theorists engage with well-established natural language data.

3.6. Finite-state OT

Mainstream phonological theory has undergone a paradigm shift with the innovation of constraint-based theories such as Optimality Theory (McCarthy & Prince, 1994; Prince & Smolensky, 1993, 2002, 2004). It was Ellison (1994) who first proposed a finite-state implementation of OT. The essence of the proposal was to construct an individual FST for each constraint. For example, with particular representational assumptions, the constraint *CODA can be encoded with (13):

(13)
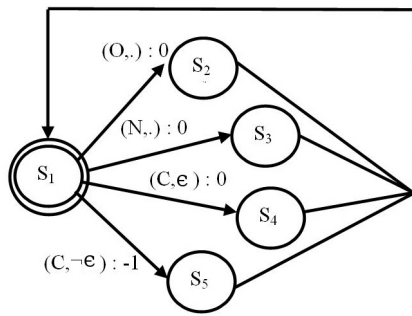
In (13), the input is coded as pairs of syllable slots and segmental material, with *O* indicating an onset position, *N* the nucleus, *C* a coda, ε the empty string (a syllable position that is not filled), and ¬ε any nonempty string (a syllable position that *is* filled). Thus, for example, when the syllabified form [al.qal.am.u] is run through the FST in (13), it is represented as in (14a), and the output is as in (14b):

(14)

a.	O	N	C	O	N	C	O	N	C
	ε	a	l	q	a	l	m	u	ε
b.	0	0	-1	0	0	-1	0	0	0

In other words, the input string is transduced to a string of constraint violations, whose sum indicates the number of constraint violations for the candidate as a negative integer. Moreover, by constructing a regular expression which generates all possible syllabifications of /alqalmu/ and performing an operation known as intersection (also called the product), one can obtain the constraint violations for every possible syllabification. The advantage of doing this with finite-state methods is that they are amenable to memory- and operation-efficient computer implementation; in fact, standard finite-state libraries have been developed for most major computer programming languages.

Subsequent work has elaborated on this conception in various ways, although the core idea of writing constraints as FSTs has remained. For example, Karttunen (1998) proposed to compose constraints according to their ranking in a particular language with *lenient composition*, which efficiently removes candidates from the computation as soon as they become suboptimal, while allowing candidates to violate high-ranked constraints when there is no better competitor. Frank and Satta (1998) study the generative power of OT, and conclude that it is regular only if individual constraints can assign at most an *n*-ary distinction in well-formedness for some finite *n*. For example, the ALIGN family of constraints, which might penalize an element according to its (potentially unbounded) distance from the edge of a word, is suspect by these criteria.

Finite-state OT is particularly exciting to me because of its potential for the study of *learning*. The key ideas can be traced to a variety of papers. Goldwater and

Johnson (2003) first noticed that Harmonic Grammars could be extended to log-linear (maximum entropy) models, simply by treating constraints as the feature functions. Berger et al. (1996) proved that under mild assumptions the likelihood function of log-linear models is convex in the weight space, which means that there is a unique maximum and it can be found efficiently using the *gradient* (the vector of derivatives with respect to each weight). Berger et al. further observed that the gradient can be calculated as **O–E**, where O_i is the observed violation count for constraint f_i in the training data, and E_i is the expected violation count. Eisner (2002, *et seq.*) and Riggle (2004, 2009) extended the finite-state conception of constraints with a special product operation that tracks the violation vector for an entire grammar, along with the vector's (log-)probability, using an algebraic structure referred to as a 'violation semiring'. The violation semiring construction offers computationally efficient computation of the weighted violation vectors for any regular class of strings. Therefore, it can be used to calculate the expected violation count **E** when that value is well-defined. Together, these results imply that a machine-implemented log-linear grammatical model can feasibly be trained. Hayes and Wilson (2008) actually implemented such a model in Java, and have been producing interesting work with it in subsequent papers. I will return to this model in the Cognitive Modeling section.

Heinz and colleagues have applied finite-state techniques to the acquisition of phonology. For example, Heinz (2007) treats the acquisition of long-distance phonological patterns (such as sibilant harmony, vowel harmony, and stress assignment) using finite-state learning. Heinz observes that all such long-distance phenomena exhibit a property he calls *neighborhood distinctness*, a property which enforces certain kinds of generalization, and which falls out naturally from applying a 'state merging' operation during construction of the FSM. Later work by Heinz considers learning various classes of subregular languages, often directly motivated by particular phenomena such as vowel harmony, and sometimes with proofs of identifiablility in the limit (Heinz, 2010; Heinz & Koirala, 2010; Heinz & Lai, 2013).

Formal language theory has developed a large body of axiomatic results on classes of 'languages', defined as stringsets generated intentionally by some finite, compact generative mechanisms. Work on this topic is generally concerned with 'learnability', which is typically formulated at an abstract, algebraic level. For example, a class of languages is identifiable in limit if an optimal learning algorithm can be guaranteed to converge upon the correct language in the class given an arbitrary sample of some size. Recent work on this topic has illustrated surprising insights on the expressive equivalence of formal frameworks with very different surface characteristics, and has provided powerful tools for implementing constraint-based phonology in computationally efficient finite state machines.

4. NATURAL LANGUAGE PROCESSING AND AUTOMATIC SPEECH RECOGNITION (NLP/ASR)

> Every time I fire a linguist, the performance of the speech recognizer goes up. --Fred Jelinek (in Hirschberg, 1998)
> There are three kinds of lies: lies, damn lies, and statistics. --Benjamin Disraeli (Twain, 2006, p. 471)

Computational phonology is generally used to refer to basic research. However, there is extensive overlap with the fields of Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), since all three deal with computations involving (representations of) speech sounds. Despite the overlap, there is a certain tension between the goals of the scientist and the goals of engineers who wish to apply the science to solve real-world problems, as revealed in Jelinek's oft-repeated quip, above. This review will not address cutting-edge work in NLP or ASR, since 'computational phonology' is not generally used to describe this kind of work. Still, current computational work owes a huge debt to NLP and ASR for the application of statistical methods to natural language. I will briefly describe two concepts which originated from NLP/ASR but which have spread to computational linguistic in general.

4.1. Zipfian distributions

It seems trivial, almost to the point of banality, to observe that some things happen more than others; for example, some words are repeated more frequently than others. However, the nature of the distribution can have powerful consequences for language acquisition and processing. It turns out that the variation in word frequencies is not completely random; it follows what has come to be known as a Zipfian distribution (Zipf, 1935, 1949). This means that a small number of items have a large frequency, and a large number of items have a small frequency. It is also sometimes informally described as 'most events are rare'.

Zipfian distributions are found at every level of linguistic structure. Baayen (2001) considers the implications of this fact for morphology. An essential point is that for any natural language text, the probability of encountering a new item never drops to zero. Therefore, a functioning model of language use must always allow for unseen items. The reader might be surprised to learn how much research does not provide for this. For example, the best-known and most-successful model of word recognition, TRACE (Elman & McClelland, 1985), does not have any explicit mechanism for handling so-called Out-of-Vocabulary (OoV) items. Daland and Pierrehumbert (2011) found that even segmental diphones (a consonant or vowel, followed by another consonant or vowel) exhibit a Zipfian distribution in English. Daland and Pierrehumbert go on to show that an English listener gets enough input in one day to approximate the frequency distribution over (frequent) diphones, yet might still encounter new pairs of speech sounds throughout their life. The enormous range of variation in frequencies has important but sometimes underappreciated implications for how learners might acquire phonology.

4.2. Statistical models

As noted above, the goal of many NLP and ASR researchers is to build language technologies that work, rather than focusing on the cognitive principles that underlie language use. Of course, those two goals are not mutually exclusive, but they are not identical either. In fact, the general experience of the NLP and ASR community has been that 'dumb' models with lots of training data perform better than 'smart' models with less training data:

> I don't know how many of you work in IT have had this experience, but it's really awfully depressing to spend a year working on an interesting research idea and then discover you can get a bigger BLEU score increase by, say, doubling the size of your language model training data. I see a couple of nodding heads. --Phillip Resnick (in P. Brown & Mercer, 2013)
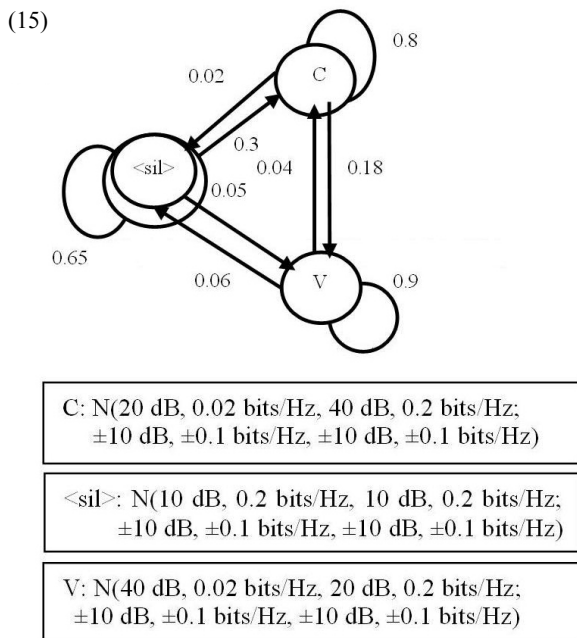
An example of a 'dumb' model in syntax is the Markov/*n*-gram models that Chomsky (1956) attacked as insufficient to explain various long-distance phenomena. From the perspective of NLP/ASR researchers, linguistic theory is good to the extent that it is useful and necessary for building systems that work:

> It's not that we were against the use of linguistics theory, linguistic rules, or linguistic intuition. We just didn't know any linguistics. We knew how to build statistical models from very large quantities of data, and that was pretty much the only arrow in our quiver. We took an engineering approach and were perfectly happy to do whatever it took to make progress. In fact, soon after we began to translate some sentences with our crude word-based model, we realized the need to introduce some linguistics into those models... We replaced the words with morphs, and included some naïve syntactic transformation to handle things like questions, modifier position, complex verb tenses and the like... Now this is not the type of syntactic or morphological analysis that sets the linguist's heart aflutter, but it dramatically reduces vocabulary sizes and in turn improves the quality of the EM parameter estimates... From our point of view, it was not linguistics versus statistics; we saw linguistics and statistics fitting together synergistically. --Peter Brown (in P. Brown & Mercer, 2013)

A crucial contribution of NLP/ASR has been the insight that probabilistic approach to language modeling is necessary for developing real-world applications. Arguably, it is also inspiring a revolution in how we conceptualize language acquisition, or at least phonological acquisition.

This community has also developed machine-learning techniques that enable efficient estimation of model

parameters. For example, commercial ASR technologies like Nuance Dragon rely on an acoustic model which relies on a 'dumb' Hidden Markov Model (HMM). An HMM is a close relative of a probabilistic FSM, with two key differences. First, the states themselves are latent variables (in the sense that the model builder posits that they exist, and they condition the model's output, but their parameters/relationships to other model components are learned during training). Second, emission of a string is not directly associated state transitions; rather, each state is associated with a probability distribution over observations. The acoustic observations are a time series $\{o^t\}^{t=1..M}$ where each o^t is some kind of vector, typically generated by some kind of spectral decomposition of overlapping time frames from the waveform. For example, a simple HMM is shown in (15):

(15)

C: N(20 dB, 0.02 bits/Hz, 40 dB, 0.2 bits/Hz;
  ±10 dB, ±0.1 bits/Hz, ±10 dB, ±0.1 bits/Hz)

<sil>: N(10 dB, 0.2 bits/Hz, 10 dB, 0.2 bits/Hz;
  ±10 dB, ±0.1 bits/Hz, ±10 dB, ±0.1 bits/Hz)

V: N(40 dB, 0.02 bits/Hz, 20 dB, 0.2 bits/Hz;
  ±10 dB, ±0.1 bits/Hz, ±10 dB, ±0.1 bits/Hz)

In this case, the task is to parse an acoustic sequence by labeling each discrete time frame as belonging to one of the categories 'C', 'V', or '<sil>' (silence). The acoustic observations have four dimensions, representing the absolute magnitude of the signal in the band 2000-5000 Hz, the entropy per Hertz in this band (a measure of aperiodicity), the absolute magnitude of the signal in the band 5000-10000 Hz, and the entropy per Hertz in this band. (Note that this type of acoustic representation is quite different from what is used in commercial applications. A concrete example is given to help the reader conceptualize an HMM. The parameters in (15) were not generated from actual speech data; they are included only for concreteness.) The solid lines represent state transitions, and the numbers represent the associated probabilities. Self-

transitions take up the bulk of the probability in each case since normally the same vowel/consonant is spread over many observation frames. The 'emission probability' boxes characterize the likelihood of emitting the current observation o^t given the posited state s^t using a multi-dimensional normal distribution. For example, the 'C' label is associated with relatively lower amplitude and less periodicity than vowels in the 2-5 kHz band, and relatively higher amplitude but still less periodicity than vowels in the 5-10 kHz band.

Much of the early, seminal work in these fields focused on developing dynamic programming techniques to train these models efficiently from limited or very large amounts of training data. Especially well-known are the Viterbi algorithm for finding the most likely sequence of states given an observation sequence (Viterbi, 1967), and the Baum-Welch (or forward-backward) algorithm for finding the unknown parameters of an HMM (Baum & Petrie, 1966; Jelinek, Bahl, & Mercer, 1975). These algorithms, or modest adaptations/generalizations of them, are still used in most or many NLP papers published today, as well as in the finite-state OT methods described earlier and elaborated in more detail in later sections.

The discussion of NLP/ASR is necessarily brief. As emphasized throughout this discussion, NLP has made significant contributions to what now might be called computational phonology, although in practice NLP is interested in engineering applications (such as ASR) and is normally considered a separate field. The use of statistical models has transformed cognitive modeling in phonology, to which I turn next.

5. COGNITIVE MODELING

The advent of statistical models in NLP offered up new avenues for more cognitively minded researchers. Early examples of this include the work of the Parallel Distributed Processing group, who formulated the TRACE model of speech perception (Elman & McClelland, 1985) as well as a hotly-contested single-route model of past tense formation (Rumelhart & McClelland, 1986). The 'connectionist' approach they employed, emphasizing so-called Artificial Neural Networks (ANNs), has largely been abandoned in contemporary cognitive science, for reasons too complex to discuss here. Nonetheless, the PDP group deserves credit for ushering in a new era in cognitive science by attempting to explicitly link (psycho-)linguistic theories with human behavioral data.

5.1. Phonotactic and phonological learning

The bulk of cognitive computational modeling of phonology that this author is aware of is concentrated in the areas of phonotactic and phonological learning. There are two key messages that this literature suggests to me. The first is that a constraint-based approach to

phonological learning makes sense from a range of standpoints. The second is that a stochastic approach to phonological variation makes sense from a range of standpoints.

5.1.1. Factoring the learning problem

As nicely set forth in Hayes (2004), a constraint-based approach makes sense of the empirical data we see on phonological development. More specifically, Hayes (2004) reviews a range of studies suggesting that infants acquire significant aspects of the phonotactics of their language by 9-11 months of age, while there is no or little evidence of unambiguously phonological alternations until 15-24 months of age. In a constraint-based framework, this pattern can be captured by a theory in which markedness constraints are learned early. While Hayes (2004) does not claim that infants have no command of faithfulness constraints, it seems intuitively plausible that it is easier to learn about which surface structures do and do not occur (phonotactics) than it is to also learn about non-transparent relationships between UR and SR.

5.1.2. Learnability proofs for constraints

Although it is in principle possible to reason about acquisition within *SPE*-style rules, the nature of the OT formalism has evidently been more amenable to formal analysis. The advent of OT was followed in short order by learning algorithms, and formal proofs of their efficacy. For example, Tesar and Smolensky (2000) summarize a large body of earlier work treating the phonological acquisition problem from the perspective of OT. One aspect of the learning problem is learning the production grammar –the component which maps underlying representations to fully specified surface representations. They give a formal proof of the 'correctness' of an algorithm they refer to as Error-Driven Constraint Demotion (EDCD), which solves this problem. That is, if the learner is given correct underlying forms and correct surface forms from an OT grammar with constraints $\mathbf{C} = \{C_k\}$, EDCD probably converges to the correct total ordering over $\mathbf{C}$ which generated the learning data. Of course, the learning problem for infants is more difficult –they must infer not only the grammar, but the underlying forms and the correct surface forms (including hidden structure). Tesar and Smolensky describe the process of assigning a fully specified surface representation to an observable form as Robust Interpretive Parsing (RIP; although Boersma, 2003, points out this could simply be called perception). Tesar and Smolensky further propose Lexicon Optimization, the assumption that when multiple input forms map to the same hypothesized surface representation, the most faithful UR is selected. They show in a series of simulations that this combination (EDCD+RIP+Lexicon Opti-

mization) correctly learns a significant majority of stress patterns in a factorial typology, although there were cases in which the learner got 'stuck', failing to converge on any correct grammar.

The adoption of scalar-valued weights has opened up additional analytic possibilities in constraint-based learning. For example, Potts, Pater, Jesney, Bhatt, and Becker (2010) showed that the simplex algorithm could be used to identify weights for a Harmonic Grammar. This provides a learnability proof for Harmonic Grammar that is entirely analogous to the correctness proof of Tesar and Smolensky's EDCD for OT, except that Potts et al. employ a pre-existing mathematical approach with a well-established pedigree. In a series of papers, Magri (2012, in press) analyses the phonotactic learning problem using a scalar-valued variant of OT in which the winning input-output candidate is determined by a total ordering of constraints, which is projected from underlying scalar-valued constraint weights. Magri gives bounds under which the use of scalar weights and error-driven re-weighting is sufficient to render learning algorithms tolerant to noise (i.e. occasional data points which violate the grammar). However, Magri's work generally deals with the grammar as a function, meaning that an input must be mapped to the same output on every occasion. Boersma and colleagues have shown that a stochastic approach provides graceful handling not only of noise, but of free variation. For example, Boersma applied stochastic gradient ascent to a probabilistic variant of OT (some readers may know this as the Gradual Learning Algorithm). Boersma and Hayes (2001) tested this algorithm on a number of empirical phenomena, finding that it was able to handle not only exceptional data points, but to accurately model genuine free variation. A more comprehensive review of this topic is given in Section 4 of Coetzee & Pater (2011).

5.1.3. Stochastic phonology

The work of Pierrehumbert and colleagues reflects some of the advantages of adopting a statistical perspective in the study of phonology and phonological acquisition. For example, Pierrehumbert (1994) conducted a study of the triconsonantal clusters observed word-medially in English. As a crude first pass, she proposed that the expected occurrences of a medial cluster in monomorphemes could be determined compositionally from the probabilities of generating the cluster from a syllable coda and a following syllable onset, e.g. $E[lfr] = |L| \cdot \Pr(l_{]\sigma}) \cdot \Pr(_{[\sigma} fr)$ where $|L|$ is the size of the monomorphemic lexicon. Pierrehumbert found that of the 8708 potentially grammatical medial clusters that could be generated in this way, only 50 were actually attested monomorphemically. Naively, one might imagine this means there is a lot of work for linguistic theory to do, explaining why so many possible events don't occur. However, Pierehumbert pointed out, over 8500 of these 8708 clusters had an expected frequency

below 1. In other words, a proper linguistic explanation was only needed for the 150 or so triconsonantal clusters which had expected frequencies well above 1, but observed frequencies of 0. 'Chance' alone was enough to explain the absence of most unattested clusters, alleviating the burden on linguists.

Coleman and Pierrehumbert (1997) further elaborated this idea by formalizing a syllable parser as a probabilistic context-free grammar (a PCFG is a CFG like in example (6), but with probabilities attached to the rewrite rules). They added prosodic features to distinguish stressed from unstressed syllables, as well as initial versus noninitial and final versus nonfinal syllables. Coleman and Pierrehumbert validated their model against human judgments from a nonce-word acceptability task. They found that the aggregate acceptability of their nonwords was almost perfectly correlated with the log-probability their model assigned, a finding that has since been replicated with numerous other probabilistic models (Daland & Pierrehumbert, 2011). In addition to comparing the model output to behavioral data, Pierrehumbert and colleagues' work represents an early instance of a critical aspect of computational cognitive modeling –specifying a meaningful baseline, against which the utility of a particular formal device can be measured.

Another domain in which a stochastic approach has had some success is in non-deterministic morphophonology. As mentioned above, the PDP group proposed an influential connectionist model of past tense production in English (Rumelhart & McClelland, 1986). This paper was very polarizing, since it suggested that both regular and 'irregular' morphophonology could be explained by a single, analogical system. A number of researchers, including Pinker and Marcus, proposed a dual-route model in which regular morphology is calculated by a rule-based grammar, while 'irregular' morphology is calculated by an analogical system. Owing to the heated rhetoric surrounding the issue and the number of papers written on this topic (Albright & Hayes, 2003; Daugherty & Seidenberg, 1992; Marcus, 1995; Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1996; Pinker & Prince, 1988; Plunkett & Marchman, 1991; to name just a few), it has become known as the Past Tense Wars. Although there is not space to review this fascinating literature, it is mentioned here because cognitive computational modeling played such a prominent role in the debate –formal models were implemented in computer programs, which generated data that was then compared to child and/or adult production. Partly as a result of researchers' commitments to actual implemented models, a number of important discoveries were made. These included the observation that minority inflectional patterns can be marginally productive (e.g. *spling* → *splung*), the discovery of output-oriented processes (e.g. irregulars like *burnt* share surface commonalities with regularly inflected items, in this case the presence of a word-final coronal stop that is not present in the verb stem), and the discovery of 'islands of reliability' not only in irregularly inflected patterns but also in regular forms (for further discussion see Albright & Hayes, 2003).

5.1.4. Constraint-based stochastic phonology

Following the research program of Hayes (2004), and the insight of Goldwater and Johnson (2003) that Harmonic Grammar can be naturally extended to the log-linear framework, Hayes & Wilson (2008) describe and implement a phonotactic learner that is supplied with a proto-lexicon (a list of wordforms) and a phonological feature set. The feature set defines a set of natural classes, following mainstream phonological theory. The software then considers grammars consisting of 'n-gram constraints', e.g. the bigram constraint '*[-son,+vcd][-son,-vcd]' might prohibit a sequence of obstruents O_1O_2 in which O_1 is voiced while O_2 is voiceless. For a given set of constraints, the software uses the finite-state methods of Riggle (2004, 2009) to rapidly determine the optimal weights. The grammar is built and pruned iteratively, by selecting new constraints from a very large hypothesis space according to various search heuristics, and then retaining those constraints which pass a complexity-penalized statistical criterion for improving the model fit to the training data. Hayes & Wilson (2008) demonstrate that the grammars learned by the model exhibit various empirically desirable properties. For example, when trained on onset clusters in the English lexicon, it assigns gradient well-formedness scores to legal and unattested onset clusters, which correlate quite tightly with the aggregate judgments of schoolchildren on the same onsets as reported in the body of the paper. Further computational work studying this model's predictions for sonority sequencing is given in Daland and Pierrehumbert (2011) and Hayes (2011). Hayes and White (2013) use the model as a baseline to test for 'phonetic naturalness' effects in learning, i.e. whether two putative constraints which receive equal support from the lexicon, but differ in the extent of phonetic motivation, are treated equally by adult English speakers in rating novel forms.

The work of Jarosz (2006, 2013) has concentrated particularly upon the problem of learning underlying representations in stochastic constraint-based phonology. For example, Jarosz (2013) contains a careful analysis of why Robust Interpretive Parsing (Tesar & Smolensky, 2000) fails in particular cases; among other things, Jarosz concludes that encoding a probability distribution across outputs allows the learner to recover from the 'traps' that caused Tesar and Smolensky's algorithm (which was cast in categorical, non-stochastic OT) to fail.

This moment is a very exciting one in the theory of phonological acquisition. The fieldwide shift to constraint-based theories has opened up multiple new lines of attack on the acquisition problem. As Hayes (2004) pointed out, the constraint-based approach is compatible

with the developmental trajectory that is actually observed, under the interpretation that children set the relative prioritization of markedness constraints rather early in development. Nearly all of the papers reviewed in this section represent significant insights onto the acquisition problem, that would not have been possible under *SPE*-style rules. While there are no doubt additional subtleties in this approach that have not been discovered, the rather rapid progress that has been made in the last 10 years on phonological acquisition in particular arguably outstrips the progress that had been made in the preceding 30-40 years during which *SPE*-style rules were the dominant phonological framework.

One part of what has made this progress possible is that the constraint-based approach lends itself naturally to problem representations that are similar, and adaptable to, problem representations in machine learning. The more that linguistic problems can be represented like problems in other scientific fields, the more we linguists are able to leverage the powerful computational tools that have been developed to solve them, such as maximum entropy models (Goldwater & Johnson, 2003; Hayes & Wilson, 2008; Jarosz, 2013). At the same time, the adoption of machine learning methods promises to help focus phonological theory on the substantive components which it adds, over and above theory-innocent machine learning methods. For example, Hayes repeatedly makes the point that a kitchen-sink approach to constraints fails with toy languages and otherwise successful learning algorithms (Hayes, 2004; Hayes & White, 2013). Analogously, it is common lore amongst theoretical phonologists that a successful OT analysis can be sunk by the wrong constraint, and this holds equally true in a computational setting where some of the candidate enumeration and scoring is done rigorously by the computer.

We can expect further, rapid progress on this domain in particular; the author is in communication with a number of scholars doing new and interesting things on this topic at this very moment. In the next subsection, we turn to another area where computational modeling has had a significant impact on rapid progress, word segmentation.

5.2. Word segmentation

Word segmentation is the perceptual process whereby listeners parse the speech stream into word-sized units. As evident from listening to speech in an unfamiliar language, many words are not followed by a silence or other language-general auditory boundary cue. However, fluent and normally-hearing listeners epiphenomenally report the sensation of hearing discrete words during speech perception, except under the most challenging listening conditions. Word segmentation refers to the cognitive process or processes that have applied between the auditory level and the listener's percept, of discrete words in a sequence.

One of the earliest computational approaches to word segmentation was the seminal TRACE model of speech perception, published by the already-mentioned PDP research group (Elman & McClelland, 1985). In this model, the listener is equipped with a bank of phonological features, a phoneme (or allophone) inventory, and an inventory of words. The 'auditory input' is represented as a time-varying vector of feature values. The model is a specific instance of a general class of models, quite popular in the psycholinguistic literature, known as 'spreading activation': the perceptual information from the 'bottom' (in this case, auditory featural) level percolates up to 'higher' levels (phonemes, and then words), and in some cases 'top-down' information also percolates downward. As a result, the 'output' of the model is a time-varying vector of word activations. The model is deemed to have successfully parsed a sentence if at the end of the sentence, all of the sentence's words are highly activated, and no other words are highly activated.

As Strauss, Harris, and Magnuson (2007) write:

> Although TRACE was introduced 20 years ago, it continues to be vital in current work in speech perception and SWR. Despite well-known limitations (acknowledged in the original 1986 article and discussed below), TRACE is still the best available model, with the broadest and deepest coverage of the literature... TRACE has proved extremely flexible and continues to spur new research and provide a means for theory testing. For example, it has provided remarkably good fits to eyetracking data from recent studies of the time course of lexical activation and competition (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson & Tanenhaus, 2001), including subtle effects of subphonemic stimulus manipulations (Dahan, Magnuson, Tanenhaus, & Hogan, 2001). (p. 20)

TRACE is mentioned here because, among other things, it has been claimed to account for word segmentation. The idea is that if you recognize the words themselves, the epiphenomenal percept of word segmentation has been explained. However, as hinted above, TRACE is not necessarily a viable model of acquisition. In particular, the model can only recognize words in its lexicon; no model-internal means is available for processing novel words and adding them to the lexicon. As is generally acknowledged in the literature on the acquisition of word segmentation, this is an essential aspect of the larger problem, since experimental evidence suggests that infants are able to segment previously unknown words, and indeed, this is the majority of new words that are learned (for argumentation see Daland & Pierrehumbert, 2011, and Goldwater, Griffiths, & Johnson, 2009).

Subsequent computational research on this topic employed corpus studies in combination with connectionist modeling (Aslin, Woodward, LaMendola, & Bever, 1996; Cairns, Shillcock, Chater, & Levy, 1997; Christiansen, Allen, & Seidenberg, 1998; Elman, 1990),

with the promising result that relatively simple neural network models could predict word boundaries *without* necessarily recognizing the neighboring words. However, owing to the well-known difficulties with interpreting the internal representations of connectionist networks, this line of research stalled shortly after the initial wave, essentially because it proved impossible to reason from the modeling results to how infants actually solved the problem. Although this is a more general issue with modeling research, it proved especially acute here because it was not even possible to determine how the models solved the problem.

Nonetheless, the finding that prelexical segmentation was computationally practical had important consequences. Experimental evidence began pouring in around this time for *phonotactic* segmentation, meaning segmentation based on (knowledge of) likely, unlikely but permissible, and impermissible sequences within and across prosodic units such as words (e.g. Jusczyk, Hohne, & Baumann, 1999; Jusczyk, Houston, & Newsome, 1999; Mattys & Jusczyk, 2001; Saffran, Aslin, & Newport, 1996; for a more comprehensive review see Daland & Pierrehumbert, 2011). The experimental evidence shows quite clearly that infants can and do extract new wordforms from the speech stream, even from 'difficult' positions such as phrase-medially when there are good phonotactic cues.

This prompted a wave of computational models which attempted to solve the segmentation problem using only phonotactic knowledge. Early instances include Xanthos (2004) and Fleck (2008), who used utterance boundary information to infer lexical phonotactic properties, as originally suggested by Aslin, Woodward, LaMendola, and Bever (1996). A probabilistically rigorous bootstrapping model was formulated and tested in Daland and Pierrehumbert (2011) using *diphones*, sequences of two segments; in English, individual diphones typically have positional distributions that are highly skewed toward being either word-internal, or word-spanning, so that this phonotactic cue is an excellent one for word segmentation. Daland and Pierrehumbert advocate for a phonotactic approach to word segmentation because phonotactic segmentation becomes efficacious as soon as infants possess the necessary phonetic experience, around 9 months, consistent with the developmental evidence. Moreover, Daland and Pierrehumbert show that the phonotactic approach is robust to conversational reduction processes that occur in English. For example, it is well-known that word-final coronal stops are often deleted in conversational English; Daland and Pierrehumbert show that this kind of process causes only a modest decrement to their phonotactic model, but has rather more drastic effects on *lexical* models which use wordform recognition to do word segmentation (since current-generation lexical models assume the surface pronunciation of a wordform is its canonical and only form, their distributional assumptions are violated by speech containing pronunciation variation). Adriaans and Kager (2010) propose an analogous model in the framework of OT, which induces segmentation constraints from featural co-occurrence information.

The phonotactic approach has not panned out as well as its proponents originally hoped, however. As the empirical coverage widened to other languages, it became clear that phonotactic approaches always worked best for English (vs. Korean: Daland & Zuraw, 2013; vs. Spanish and Arabic: Fleck, 2008; vs. Japanese: Fourtassi, Börschinger, Johnson, & Dupoux, 2013; et alia). Moreover, the assumption (based on maternal questionnaires; Dale & Fenson, 1996) that 9-month-old infants barely knew any words was contradicted by experimental evidence (e.g. Mandel, Jusczyk, & Pisoni, 1995) suggesting that infants knew some word*forms* as early as 4-6 months, even if they were not necessarily aware of the corresponding meanings.

In the meantime, the phonotactic approach to modeling word segmentation was overshadowed by the Bayesian, lexical approach developed by Goldwater, Johnson, and colleagues. This approach, which had its roots in the computational models of Batchelder (2002) and Brent and Cartwright (1996), returns to the view of word segmentation as an epiphenomenon of word recognition popularized in TRACE, but departs from TRACE in various ways. Most crucially, the models included means to add previously unencountered wordforms to its lexicon ('learn new words'); also, Brent and Cartwright (1996) defined an explicit and probabilistic mathematical objective which their model was supposed to maximize. Thus, Brent and Cartwright advocated for framing the segmentation problem at Marr's computational level ('What is the mathematical characterization of the function that humans optimize?') rather than the algorithmic level ('How do humans *find* the optimal solution for the function that they are optimizing?'). Goldwater, Griffiths, and Johnson (2009) extend the early work of Brent and Cartwright to a more general setting, factoring the learning problem so as to enable efficient optimization, reframing the objective in a Bayesian setting (rather than the related, but more restricted Minimal Description Length approach used by Brent and Cartwright; for discussion and analysis see Goldwater, 2006), and extending the data model so as to be both more powerful and more flexible. For example, Goldwater et al. (2009) show that better segmentation is predicted if infants attend to dependencies between words, a prediction that was retroactively confirmed by an experimental study showing that 6-month-olds use their own names to segment the following word (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005).

Numerous authors have followed up on Goldwater and colleagues' seminal work. For example, Blanchard, Heinz, and Golinkoff (2010) adapt the model in Goldwater et al. (2009) by including an incremental *n*-gram phonotactic model, whose parameters are discovered during the learning process. They found a significant but very modest gain in performance, suggesting that much of the problem-solving power of Goldwater's

model is actually located in the prior distribution (owing to reasons of space, I am unable to describe this model in more detail here; the reader is encouraged to consult the original paper for clear exposition). Pearl & colleagues have experimented with the idea that 'adding performance back in' to computational-level models can yield more psycholinguistically valid (and sometimes more accurate) performance, by incorporating limited short-term memory and/or long-term forgetting into Goldwater-like models (Pearl, Goldwater, & Steyvers, 2011; Phillips & Pearl, 2012). Lignos (2012) presents an incremental model with a slightly different objective than in Goldwater et al. (2009); an innovation is the use of a lexical filter which prevents low-confidence words from being incorporated into the model's lexicon. A variety of lexical filters have been used in previous work, including especially the constraint that a word must contain a vowel (Brent & Cartwright, 1996) or that it must have a certain minimal frequency (Daland & Pierrehumbert, 2011; see Ch. 5 of Daland, 2009, for modeling, analysis, and discussion of 'error snowballs' and Pearl et al., 2011, for argumentation that memory limitations help prevent error snowballs by forgetting early misparses).

The rapid, intense progress that has taken place in our understanding of word segmentation acquisition has been driven by an interplay and dialogue between a variety of research traditions, most notably developmental psycholinguists (Jusczyk, Mattys, Morgan, Saffran, etc.) and cognitive computational modelers (Daland, Goldwater, Johnson, Pearl, etc.), as well as researchers who are able to mix these methodologies (Aslin, Kager, Swingley, etc.). This is, in the author's humble opinion, a wonderful thing, and it is to be hoped that this example spreads to other domains.

More generally, the impact of cognitive modeling cannot be understated in linguistic theory and in cognitive science more generally. The interaction between domain-general and domain-specific representations and learning algorithms is a topic of perennial interest, and computational modeling has and continues to shed new light on the complexities. Modeling has in some cases clearly ruled out hypotheses as to cognitive processes that seemed a priori quite plausible; while in other cases it has shown that two formalisms which might naively be supposed to make completely divergent predictions actually offer statistically indistinguishable explanations for the very same data set (e.g. Jarosz, 2013). Just as with formal language theory for framework comparison, it is safe to predict that there will be more of this work in the future, not less. In the next and final content section of this review I turn briefly to the topic of corpus studies.

6. CORPUS STUDIES

A corpus study is any study in which the central data consists of a 'corpus' –a body of text representing some aspect of language use– and the central analysis consists of counting elements in the text and doing statistical comparisons. Corpus studies flourished in the early days of the CHILDES database (MacWhinney, 2000), an early crowdsourced project in which (usually orthographic) child-related corpora were assembled together under the auspices of a single research group. For example, much of the early work on morphological acquisition focused on order-of-morpheme acquisition, e.g. comparing the time and frequency of -*ing*, -*ed*, and other English functional morphemes (R. Brown, 1973).

Owing to the orthographic coding of most corpora, and the phonologically non-transparent nature of English (the analysis language for most corpus-based research to date), the bulk of corpus work has focused on morphology and syntax rather than phonology. Nonetheless, there is a significant body of corpus work in phonology. I will limit the review to a few examples, as much of this work is of a similar character.

Two studies which address phenomena of interest to theoretical phonology were done by Zuraw and colleagues. Zuraw (2006) collected a corpus of Tagalog loanwords using Internet blogs. Loanwords were desirable for this study since the research question pertained to the productivity of intervocalic tapping, and the productivity of phonological patterns from high-frequency native items is confounded with lexicalization. Using this corpus, Zuraw examined how morphological status interacts with a variable phonological process; she found interesting differences between the prefix+stem and stem+enclitic cases, which there is not space to discuss here. In a conceptually similar study, Hayes, Zuraw, Siptár, and Londe (2009) investigate the vowel harmony pattern of Hungarian, which is largely categorical, but exhibits variation in particular cases (notably, when an initial back vowel is followed by one or more 'neutral' vowels, which do not undergo acoustically obvious harmony processes themselves). Hayes et al. (2009) note several 'phonetically unnatural' aspects of the harmony system which appear, at least statistically, to not be due to chance alone (for example, associations between consonant place and vowel height that condition the application rate). They go on to assess the productivity of these 'unnatural' patterns, and compare them to the productivity of 'natural' patterns with similar statistical support, finding that Hungarian native speakers exhibit knowledge of both, but apparently exhibit more productivity for the 'natural' patterns (see the paper for details).

Larsen and Heinz (2012) present a corpus study, also of vowel harmony, but in Korean, and particularly in its onomatopoetic sub-lexicon. Their analysis confirms some aspects of previous accounts of this sub-lexicon, but add nuances, e.g. that the harmony class of a vowel may depend on its position in the word. Daland (2013) presents a corpus study of adult- versus child- directed speech, in which he compares the relative frequency of different segmental classes. Daland argues against the claim that adults tailor the segmental frequencies in their

child-directed speech, by showing that the moment-to-moment variation in segmental frequencies dwarfs the putative aggregate differences that had been reported in previous research.

In all of these corpus studies, researchers take an existing corpus (or create one) and then analyze it and compare the counts against the predictions of some existing phonological theory or account. Corpus studies are relatively easy to conduct and replicate once the corpus has been created, so they are an appealing methodology. However, it is the norm to supplement corpus studies with additional computational studies and/or experimentation, so as to provide converging evidence. There are many corpus studies that could have been reviewed here, and I selected a mere handful to illustrate the 'flavor' of this style of research. (A number of corpus studies were also reviewed in the cognitive modeling section earlier). This style of research is reviewed here, at least briefly, because it is considered to be 'computational phonology' by many researchers, including specialists on language acquisition.

7. SUMMARY AND CONCLUSIONS

In this paper I have reviewed a number of sub-fields which I or close colleagues consider to be 'computational phonology'. I began with formal language theory as it is specifically applied to phonology. After reviewing the fundamentals, I discussed recent theoretical work of interest, including the use of equivalencies between formal languages and logics to compare formal frameworks (like *SPE* and OT), as well as the application of finite-state methods for efficient optimization of large-scale constraint-based models. Next, I briefly discussed the influence of NLP/ASR (Natural Language Processing and Automatic Speech Recognition) on computational phonology; although those fields are not considered computational phonology, cognitive scientists owe a huge debt to these fields for introducing and demonstrating the utility of probabilistic models for natural language problems. In the section of the paper that corresponds the most closely to my own research interests, I discussed cognitive computational modeling in general, and focused in particular on computational approaches to phonological and phonotactic acquisition, as well as the acquisition of word segmentation by infants and children. Finally, I very briefly discussed corpus studies; there is a long tradition in corpus work and it is a very general methodology, so I only gave a few examples to illustrate what it can and cannot do.

Stepping back from the many and important details that go into making any one particular study, it is time to revisit the question with which this article began: What is computational phonology? Let us begin with what is common. As claimed in the introduction, many or most of the works reviewed above draw upon a common foundation of formal language theory. For example, some of the most exciting work on cognitive

modeling of phonological acquisition makes use of finite-state OT (Hayes & Wilson, 2008). Similarly, most of the work on computational phonology relies on a shared body of methodological knowledge about corpus linguistics. For example, it is nearly always necessary to preprocess a corpus for one's particular research needs. Moreover, the Natural Language Processing field has repeatedly and forcefully demonstrated the dangers of overfitting; it is now received wisdom in this field that generalization must be assessed by testing on a different data set than the model was trained on (except in certain cases of unsupervised learning). Nearly all of the work reviewed above in cognitive computational modeling deals either with a corpus of phonological data, or with behavioral results from a 'corpus' of stimuli, or both. Finally, the bulk of the studies reviewed here deal specifically with first language acquisition (although, to be fair, that partially reflects the author's interests, in addition to the inherent biases of the field). This is quite a bit of shared knowledge and methodological commonality. However, if we examine the research questions that each subfield asks, despite the fact that there is a general preoccupation with language acquisition, we still see a greater amount of variation than is, I think, common for a coherent field.

Within formal language theory, the pursuit is really not of empirical phenomena that do or don't occur in natural languages; rather, the goal is to understand and elucidate the formal relationships between various formal models of 'language'. This subfield has largely resisted probabilistic approaches, and it has concentrated on *formal* restrictions on the generative capacity of formal models (such as regular versus context-free), at the expense of *substantive* restrictions (such as the implicational universal that words with consonant onsets are strictly less marked than onsetless words). A large amount of work in this field is devoted to acquisition, but it tends to proceed in a proof-based or algorithmic manner, asking if learning algorithm A is guaranteed to learn every language L in a given class. The psychological plausibility of the learning assumptions is not always a very important concern to such researchers; rather they are interested in the mathematical and logical relationships between A and L.

Within Natural Language Processing (NLP), the goal is to solve real-world engineering problems, often ones in which money can be made. For example, it is worthy and important to translate documents from resource-rich languages like English to high information-demand languages (such as Mandarin Chinese). It is also worthy and important to translate documents from languages whose speakers produce goods and technologies (like Mandarin Chinese) to languages whose speakers consume goods and technologies (like English). Translators work slowly and must be paid a considerable amount of money; there is a lot of money to be made and saved in developing good machine translation. In this sort of application, the formal properties of a model are of interest only insofar as they impact the ultimate perfor-

mance of the system as a whole. There are of course researchers whose interests span both NLP and more basic science, including researchers who believe that understanding the way humans do language may result in better NLP, and soon. Nonetheless, the field as a whole is oriented toward developing and applying statistical models which solve 'real-world' problems. There are many and interesting problems in this field, which this author is too distant from to review in the detail they deserve here. It is quite clear, however, that the types of problems this field is concerned with are quite different than the rather abstract questions that preoccupy formal language theorists.

In cognitive computational modeling, the goal is more specifically to elucidate how humans actually do some particular linguistic task. This is related to, but crucially different from, the formal language theory approach. At the risk of oversimplifying considerably, one might put it this way: formal language theory asks, "What does model X do?"; cognitive modelers ask, "Do humans do it like model X?". That is, in this field, computational researchers are concerned much more with psychological plausibility, and less with the abstract structure of the problem space. It is no surprise, then, that computational research in this field responds and is responded to more tightly with developmental research on language acquisition.

My goal, in reviewing these different subfields, is not to claim that one is superior to another. Rather, it has been to illustrate the rich tapestry of human thought that falls under the broad umbrella term 'computational phonology'. There are strands that connect each of these subfields, even as the core concerns differ from researcher to researcher and subfield to subfield. Computational phonology is getting bigger and bigger, and fragmenting more with each passing year. But, too, we are learning more and more.

REFERENCES

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language, 62*, 311-331. http://dx.doi.org/10.1016/j.jml.2009.11.007

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental Study. *Cognition, 90*, 119-161. http://dx.doi.org/10.1016/S0010-0277(03)00146-X

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117-134). Mahwah, NJ: Erlbaum.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, Netherlands: Kluwer Academic. http://dx.doi.org/10.1007/978-94-010-0844-0

Baković, E. (2007). A revised typology of opaque generalisations. *Phonology, 24*, 217-259. http://dx.doi.org/10.1017/S0952675707001194

Baković, E. (2011). Opacity and ordering. In J. Goldsmith, J. Riggle & A. C. L. Yu (Eds.), *The handbook of phonological theory* (2nd ed.). Oxford, UK: Wiley-Blackwell. http://dx.doi.org/10.1002/9781444343069.ch2

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*, 167-206. http://dx.doi.org/10.1016/S0010-0277(02)00002-1

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 37*(6), 1554-1563. http://dx.doi.org/10.1214/aoms/1177699147

Berger, A., Della Pietra, S., & Della Pietra, V. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics, 22*(1), 39-71.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language, 37*, 487-511. http://dx.doi.org/10.1017/S030500090999050X

Boersma, P. (2003). [Review of the book *Learnability in Optimality Theory*, by B. Tesar & P. Smolensky]. *Phonology, 20*, 436-446. http://dx.doi.org/10.1017/S0952675704230111

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry, 32*, 45-86. http://dx.doi.org/10.1162/002438901554586

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science, 16*, 298-304. http://dx.doi.org/10.1111/j.0956-7976.2005.01531.x

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93-125. http://dx.doi.org/10.1016/S0010-0277(96)00719-6

Brown, P., & Mercer, R. (2013). Twenty years of Bitext [Transcription and slides]. Invited talk. *EMNLP workshop Twenty years of Bitext*. Seattle, WA. Retrieved from: http://cs.jhu.edu/~post/bitext/

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Buccola, B., & Sonderegger, M. (2013). *On the expressivity of Optimality Theory versus rules: An application to opaque patterns*. Refereed presentation presented at the meeting Phonology 2013, Umass Amherst, 09 November 2013.

Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33*, 111-153. http://dx.doi.org/10.1006/cogp.1997.0649

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*, 113-124. http://dx.doi.org/10.1109/TIT.1956.1056813

Chomsky, N. (1959). [Review of the book *Verbal Behavior*, by B. F. Skinner]. *Language, 35*(1), 26-58. http://dx.doi.org/10.2307/411334

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*(2-3), 221-268. http://dx.doi.org/10.1080/016909698386528

Coetzee, A. W., & Pater, J. (2011). The place of variation in phonological theory. In J. Goldsmith, J. Riggle & A. C. L. Yu (Eds.), *The handbook of phonological theory* (2nd ed.). Oxford, UK: Wiley-Blackwell. http://dx.doi.org/10.1002/9781444343069.ch13

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *3rd Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop, 12 July 1997* (pp. 49-56). Somerset NJ: Association for Computational Linguistics.

Daland, R. (2009). *Word segmentation, word recognition, and word learning: A computational model of first language acquisition* (unpublished doctoral dissertation). Northwestern University, IL. Retrieved from: http://www.linguistics.northwestern.edu/docs/dissertations/dalandDissertation.pdf

Daland, R. (2013). Variation in child-directed speech: A case study of manner class frequencies. *Journal of Child Language, 40*(5), 1091-1122. http://dx.doi.org/doi:10.1017/S0305000912000372

Daland, R., & Pierrehumbert, J.B. (2011). Learning diphone-based segmentation. *Cognitive Science, 35*(1), 119-155. http://dx.doi.org/10.1111/j.1551-6709.2010.01160.x

Daland, R., & Zuraw, K. (2013). *Does Korean defeat phonotactic word segmentation?* Short paper presented at the 51st Annual Meeting of the Association for Computational Linguistics in Sofia, Bulgaria, August 4-9, 2013.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125-127. http://dx.doi.org/10.3758/BF03203646

Daugherty, K., & Seidenberg, M. S. (1992). Rules or connections? The past tense revisited. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 259-264). Hillsadle, NJ: Erlbaum.

Eisner, J. (2002). Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 1-8). East Stroudsburg, PA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1073083.1073085

Ellison, M. T. (1994). Phonological derivation in optimality theory. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)* (Vol. 2, pp. 1007-1013). Kyoto, Japan: Association for Computational Linguistics. http://dx.doi.org/10.3115/991250.991312

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211. http://dx.doi.org/10.1207/s15516709cog1402_1

Elman, J. L., & McClelland, J. L. (1985). An architecture for parallel processing in speech recognition: The TRACE model. In M. R. Schroeder (Ed.), *Speech recognition* (pp. 6-35). Gottingen, Germany: Biblioteca Phonetica.

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 130-138). Madison, WI: Omnipress.

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 1-10). Sofia, Bulgaria, August 8, 2013.

Frank, R., & Satta, G. (1998). Optimality theory and the computational complexity of constraint violability. *Computational Linguistics, 24*, 307-315.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*(5), 447-474. http://dx.doi.org/10.1016/S0019-9958(67)91165-5

Goldsmith, J. (1976). *Autosegmental phonology*. Doctoral dissertation, MIT, MA.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology*. Oxford, UK: Basil Blackwell.

Goldwater, S. (2006). *Nonparametric Bayesian models of lexical acquisition* (unpublished doctoral dissertation). Brown University, RI. Retrieved from: http://homepages.inf.ed.ac.uk/sgwater/papers/thesis_1spc.pdf

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory* (pp. 113-122). Stockholm University, Sweden.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*(1), 21-54. http://dx.doi.org/10.1016/j.cognition.2009.03.008

Graf, T. (2010a). Comparing incomparable frameworks: A model theoretic approach to phonology. *University of Pennsylvania Working Papers in Linguistics, 16*(1), art. 10. Retrieved from: http://repository.upenn.edu/pwpl/vol16/iss1/10

Graf, T. (2010b). Formal parameters of phonology: From Government Phonology to SPE. In T. Icard & R. Muskens (Eds.), *Interfaces: Explorations in logic, language and computation*, *Lecture Notes in Artificial Intelligence 6211* (pp. 72-86). Berlin, Germany: Springer.

Hayes, B. (2004). Phonological acquisition in Optimality Theory: the early stages. In R. Kager, J. Pater & W. Zonneveld (Eds.), *Fixing priorities: Constraints in phonological acquisition* (pp. 158-203). Cambridge University Press.

Hayes, B. (2011). Interpreting sonority-projection experiments: The role of phonotactic modeling. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 11-Hong Kong)* (pp. 835-838) Hong Kong, PRC.

Hayes, B., Kie, Z., Siptár, P., & Londe, Z. C. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language, 85*(4), 822-863. http://dx.doi.org/10.1353/lan.0.0169

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry, 44*(1), 45-75. http://dx.doi.org/10.1162/LING_a_00119

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*(3), 379-440. http://dx.doi.org/10.1162/ling.2008.39.3.379

Heinz, J. (2007). *The Inductive Learning of Phonotactic Patterns* (doctoral dissertation), University of California, Los Angeles.

Heinz, J. (2010). String extension learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics in Uppsala, Sweden* (pp. 897-906).

Heinz, J. (2011a). Computational Phonology - Part I: Foundations. *Language and Linguistics Compass, 5*(4), 140-152. http://dx.doi.org/10.1111/j.1749-818X.2011.00269.x

Heinz, J. (2011b). Computational Phonology - Part II: Grammars, Learning, and the Future. *Language and Linguistics Compass, 5*(4), 153-168. http://dx.doi.org/10.1111/j.1749-818X.2011.00268.x

Heinz, J., & Koirala, C. (2010). Maximum likelihood estimation of feature-based distributions. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, 28-37,* Uppsala, Sweden.

Heinz, J., & Lai, R. (2013). Vowel harmony and subsequentiality. In A. Kornai & M. Kuhlmann (Eds.), *Proceedings of the 13th Meeting on Mathematics of Language,* Sofia, Bulgaria.

Hirschberg, J. (1998). 'Every time I fire a linguist, my performance goes up', and other myths of the statistical natural language processing revolution. Invited talk. *15th National Conference on Artificial Intelligence, Madison, WI.*

Jardine, A. (in press). Logic and the Generative Power of Autosegmental Phonology. In *Supplemental Proceedings of Phonology 2013.* Retrieved from: https://sites.google.com/site/adamajardine/research-interests

Jarosz, G. (2006). *Rich lexicons and restrictive grammars - Maximum likelihood learning inOptimality Theory* (doctoral dissertation). Johns Hopkins University. Retrieved from Rutgers Optimality Archive No. 884.

Jarosz, G. (2013). Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology, 30*, 27-71. http://dx.doi.org/10.1017/S0952675713000031

Jelinek, F., Bahl, L., & Mercer, R. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions of Information Theory, 21*(3), 250-256. http://dx.doi.org/10.1109/TIT.1975.1055384

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics, 61,* 1465-1476. http://dx.doi.org/10.3758/BF03213111

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*(3-4), 159-207. http://dx.doi.org/10.1006/cogp.1999.0716

Kaplan, R.M., & Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics, 20*(3), 331-379.

Karttunen, L. (1998). The proper treatment of optimality theory in computational phonology. In *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing* (1-12). Ankara, Turkey: Association for Computational Linguistics.

Kearns, M., & Valiant, L. (1994). Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM, 41*(1), 67-95. http://dx.doi.org/10.1145/174644.174647

Larsen, D., & Heinz, J. (2012). Neutral vowels in sound-symbolic vowel harmony in Korean. *Phonology, 29,* 433-464. http://dx.doi.org/10.1017/S095267571200022X

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: An application. In *Proceedings of the twelfth annual conference of the Cognitive Science Society* (pp. 884-891). Cambridge, MA: Lawrence Erlbaum.

Li, M., & Vitányi, P. M. B. (1991). Learning simple concepts under simple distributions. *SIAM Journal of Computing, 20*, 911-935. http://dx.doi.org/10.1137/0220056

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics, 30,* April 13-15, 2012.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Volume I: Transcription format and programs. Volume II: The database*. Mahwah, NJ: Lawrence Erlbaum.

Magri, G. (2012). Constraint promotion: not only convergent but also efficient. In *CLS 48: Proceedings of the 48th annual conference of the Chicago Linguistic Society*, Chicago, IL.

Magri, G. (in press). Error-driven and batch models of the acquisition of phonotactics: David defeats Goliath. In *Phonology 2013: Proceedings of the 2013 Phonology Conference*, November 8-10, 2013, Amherst, MA.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science, 6*, 315-318. http://dx.doi.org/10.1111/j.1467-9280.1995.tb00517.x

Marcus, G. F. (1995). The acquisition of the English past tense in children and multi-layered conectionist networks. *Cognition, 56*, 271-279. http://dx.doi.org/10.1016/0010-0277(94)00656-6

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1996). German inflection: The exception that proves the rule. *Cognitive Psychology, 29*, 189-256. http://dx.doi.org/10.1006/cogp.1995.1015

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91-121. http://dx.doi.org/10.1016/S0010-0277(00)00109-8

McCarthy, J.J. (1981). A prosodic theory of non-concatenative morphology. *Linguistic Inquiry, 12*(3), 373-418.

McCarthy, J.J. (2008). The gradual path to cluster simplification. *Phonology, 25*, 271-319. http://dx.doi.org/10.1017/S0952675708001486

McCarthy, J.J. (2011). Autosegmental spreading in Optimality Theory. In J. Goldsmith, A. E. Hume & L. Wetzels (Eds.), *Tones and Features* (pp. 195-222). Berlin, Germany: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110246223.195

McCarthy, J.J., & Prince, A. (1994). The emergence of the unmarked: Optimality in prosodic morphology. In *Proceedings of the North East Linguistics Society 24*. Amherst, MA.

Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online Learning Mechanisms for Bayesian Models of Word Segmentation. *Research on Language and Computation, 8*(2-3), 107-132. http://dx.doi.org/10.1007/s11168-011-9074-5

Phillips, L. & Pearl, L. (2012). Syllable-based Bayesian inference: A (more) plausible model of word segmentation. *Workshop on Psychocomputational Models of Human Language Acquisition.* Portland, OR.

Pierrehumbert, J. (1994). Syllable structure and word structure: a study of triconsonantal clusters in English. In P. Keating (Ed.), *Papers in laboratory phonology III: Phonological structure and phonetic form* (pp. 168-188). Cambridge, UK: Cambridge University Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition. *Cognition, 28*(1-2), 73-193. http://dx.doi.org/10.1016/0010-0277(88)90032-7

Plunkett, K. & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition, 38*, 43-102. http://dx.doi.org/10.1016/0010-0277(91)90022-V

Potts, C., Pater, J., Jesney, K., Bhatt, R., & Becker, M. (2010). Harmonic Grammar with linear programming: From linear systems to linguistic typology. *Phonology, 27*, 77-117. http://dx.doi.org/10.1017/S0952675710000047

Potts, C., & Pullum, G. K. (2002). Model theory and the content of OT constraints. *Phonology, 19*, 361-393.

Prince, A., & Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar. Technical Report, RUCSS, Rutgers University, New Brunswick, NJ. Published in 2004 by Blackwell.

Prince, A., & Smolensky, P. (2002). Optimality Theory: Constraint interaction in generative grammar. Retrieved from: roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell.

Riggle, J. (2004). *Generation, recognition, and learning in finite state Optimality Theory* (doctoral dissertation). UCLA, CA.

Riggle, J. (2009). Violation semirings in Optimality Theory. *Research on Language and Computation, 7*(1), 1-12. http://dx.doi.org/10.1007/s11168-009-9063-0

Riggle, J., & Wilson, C. (2005). Local optionality. In *Proceedings of NELS 35*. Amherst, MA: GLSA.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216-271). Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928. http://dx.doi.org/10.1126/science.274.5294.1926

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy, 8*(3), 333-343, http://dx.doi.org/10.1007/BF00630917

Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: From neural computation to Optimality-Theoretic grammar* (Vol. 1: Cognitive Arquitecture, pp. xvii-563. Vol. 2: Linguistic and Philosophical Implications, pp. xvii-611). Cambridge, MA: MIT Press.

Stabler, E. (2009). Computational models of language universals. In M. H. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (Rev. ed., pp. 200-223). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195305432.003.0010

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE : A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, 39*(1), 19-30. http://dx.doi.org/10.3758/BF03192840

Tesar, B., & Smolensky, P. (2000) *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.

Twain, M. (2006). Chapters from My Autobiography -- XX. *North American Review DCXVIII*. Project Gutenberg. Retrieved from: http://www.gutenberg.org/files/19987/19987-h/19987-h.htm (original work published in 1906).

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM 27*, 1134-1142. http://dx.doi.org/10.1145/1968.1972

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*(2), 260-269. http://dx.doi.org/10.1109/TIT.1967.1054010

Xanthos, A. (2004). Combining utterance-boundary and predictability approaches to speech segmentation. In W. G. Sakas (Ed.), *Proceedings of the first workshop on psycho-computational models of language acquisition at COLING 2004* (pp. 93-100). Geneva, Switzerland.

Zipf, G. K. (1935). *The Psychobiology of Language*. Boston, MA: Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.

Zuraw, K. (2006). Using the web as a phonological corpus: a case study from Tagalog. In *EACL-2006: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics/Proceedings of the 2nd International Workshop on Web As Corpus* (pp. 59-66). http://dx.doi.org/10.3115/1628297.1628306