History of Phonology: Learnability

Jeffrey Heinz and Jonathan Rawski
Dept. of Linguistics and Institute for Advanced Computational Science
Stony Brook University

To appear in *The Oxford Handbook of the History of Phonology, B. Elan Dresher and Harry van der Hulst (eds.)*

1 ♫ What a short strange trip it's been

It is strange to write a chapter about the history of a subfield like learnability *in phonology*. On the one hand, it is arguably less than forty years old — so how much history can there be? On the other hand, it is strange because many of the ideas—with a few remarkable and important exceptions—that dominate contemporary approaches to phonological learning have come to phonology from outside linguistics. Given the central role language continues to play in Artificial Intelligence (AI) and Machine Learning (ML)—both young fields themselves—one might expect that linguists working on programs that learn linguistic generalizations from data to be leading developments in these other areas instead of borrowing from them. So we also wonder: How much phonological learning originates in theoretical linguistics?

Despite this strangeness, it is also a crucial moment to step back and take a historical view of developments in phonological learning due to the recent attention to ML and AI. This attention pervades the scientific press (LeCun, Bengio, & Hinton, 2015; Hutson, 2018), in addition to national and international newspapers and magazines (Lewis-Kraus, 2016; Greene, 2017; Hofstadter, 2018; Beard, 2018; Marcus & Davis, 2018). This focus has allowed issues fundamental to phonology and linguistics as a whole to reemerge—innateness vs experience, empirical coverage vs scientific insight and understanding, and domain-general vs task-specific learning.

At the same time, there is a yearning within ML for the rigor and insight that linguistic theory has sought to provide. To quote a recent ML conference plenary talk, there is "an anguish in the field" because it has become "alchemy" (Rahimi & Recht, 2017). Douglas Hofstadter, as quoted by Somers (2013), noted about black box successes in learning games, "Why conquer a task if there's no insight to be had from the victory? Okay, Deep Blue plays very good chess — so what? Does that tell you something about how we play chess? No. Does it tell you about how Kasparov envisions, understands a chessboard?"

Linguistic issues, including phonological ones, are right at the center of this, and in fact have always been, going back to the dawn of the study of computation. As we will discuss, the role of language and learning has always been of prime concern to computer science and its applications in addition to any definition of artificial intelligence. It is our belief

that phonologists should actively contribute to this rich tradition, rather than merely being beneficiaries of it at best and bystanders at worst.

To do an exhaustive overview of all the different facets of phonological learning would be, in a word, exhausting. For one, things can get muddy terminologically. One might distinguish 'language learning' from 'language acquisition' from 'learnability'. Scientists may obtain an empirical description of the acquisitional stages a group of infants goes through, which may be distinct from an algorithmically worked out theory of the acquisition of linguistic competence. They may also program software to simulate aspects of acquisition from corpora. All such facets of learning have been presented, and all have made their mark within phonology, some directly, some indirectly. There are many overviews which have described the state and practice of learning models and learnability in phonology specifically (Tesar, 2007; Albright & Hayes, 2011; Heinz & Riggle, 2011; Jarosz, 2019) and readers interested in comprehensive reviews and comparisons of existing approaches are directed to these sources.

Our goal in this chapter is instead to identify and trace overarching themes and tensions in the history of phonological learning. Doing so will require that we situate the enterprise in the larger history of science, of AI and ML, and well as within the more narrow domain of linguistics, since many learnability results first emerged through studying syntactic systems. Inevitably, the topics we do and do not include are a reflection of what we think to be important developments in the field. In an active research area and with such a plethora of work, this is unavoidable; as famed historian Howard Zinn put it, "you can't be neutral on a moving train." While our goal is not to advocate for certain views over others, inevitably some opinions will emerge. It is our conviction that in a field whose youth means it is full of promise and progress, we should embrace a diversity of viewpoints and let a thousand flowers bloom, as it were.

We will highlight several foundational problems and tensions in phonological learnability, as well as key results on learning rule systems, parameter setting, and constraint-based systems. We will overview the wealth of ideas that have been borrowed from domains outside linguistics, which have seen an explosion in the years surrounding the turn of the century.

2 Learning and the computational mind

2.1 Language and computation

The birth of generative grammar followed a period of intense upheaval in the scientific and mathematical world. Out of these crises was born the theory of computation, which in turn transformed language and psychology, spurring the cognitive revolution. This marriage of many fields produced two new twin disciplines, not identical, yet not completely distinct: Artificial Intelligence and Cognitive Science. Language was crucial to the development of both. Conversely, both disciplines had an immense effect on the early course of linguistic theory. Generalization from experience—learning—became a central issue. While each discipline approached the question of learning in its own way, the notion of minds—in whichever physical form—as computational entities shaped the modern field of linguistics.

Tomalin (2006) writes that the two-pronged approach that was a major feature of generative grammar (grammars as theories of languages, plus theories about the structure of grammars themselves) has its roots in techniques developed for work on mathematics and logic, which then bled into the philosophy of science and later linguistics. These include Cauchy's (1821) methods for solidifying the calculus, Whitehead & Russell (1912)'s axiomatic method in *Principia Mathematica*, the Hilbert program (1928) to prove the consistency and foundational security of of mathematics, Carnap's proposals for logically reconstructing science in an experience (see Carnap (1928)), among others. Tomalin also notes that these styles of analysis can be found in linguistics since Bloomfield (1926)'s axiomatic approach to a general linguistic theory.

Early generative grammar was in fact a *computational* theory, and as such owed much to the development of mathematical theories of computation and computability. One model of computation was Church's (1932) lambda calculus. Another was Turing's (1937) model of computation which introduced a universal, abstract machine, now called the Turing Machine. Turing (1937) also showed that any computation performed by the lambda calculus can be performed by a Turing Machine and vice versa. Later other models of computation such as string rewriting systems (Thue, 1914) and Post's (1936) model were also shown to be Turing equivalent. These models could arguably compute anything that can be reasonably said to be computed by a human 'computer' following a fixed set of instructions.

Consequently, Kleene (1952) transformed these results into the **computability thesis**: there is an objective notion of effective computability independent of a particular formalization. What has come to be known as The Church-Turing thesis states that a function on the positive integers is effectively calculable if and only if it is computable. More broadly it is the thesis that anything computable can be computed with a Turing machine or its equivalents like the lambda calculus.

The question of whether computing machines can think was on the minds of many and this question was inextricably linked to language. Turing's (1950) imitation game for example, proposed that a machine could be said to think if a human could not reliably distinguish *conversations* between a human and a machine from conversations between two humans. Even those who have come to reject the Turing Test for detecting thinking machines offer replacements that still feature *language* prominently (Levesque, 2017).

This new field of AI pursued by John McCarthy, Marvin Minsky, Allen Newell, Herb Simon, and others in the following years, is today called Good Old-Fashioned AI (GOFAI). These researchers emphasized a kind of intelligence based not on learning from massive amounts of data, but on common sense. As John McCarthy put it in his influential 1958 paper wherein he coined the term "Artificial Intelligence",

> a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows... Our ultimate objective is to make programs that learn from their experience as effectively as humans do.

And later:

> The only way we know of expressing abstractions ... is in language. That is why we have decided to program a system which reasons verbally.

This last point is echoed by Steedman (2008), who identifies the importance of the field of computational linguistics as owing to the fact that "Human knowledge is expressed in language." McCarthy's points are highlighted by Levesque (2017) as emblematic of the centrality of language to computation, intelligence, and learning. At the same time, Levesque is quick to draw a distinction between AI and the field of cognitive science:

> What is the difference between cognitive science and AI? While there are strong connections and overlap between the two, the main difference is that cognitive science is the interdisciplinary study of people as cognitive beings, whereas AI is the study of intelligent behavior achieved through computational means. The analogy I like is the difference between studying flying animals and studying flight. Before the advent of aircraft, the only large-scale flying objects were animals like birds and bats. Cognitive science is like an interdisciplinary study of these flying animals, whereas AI is more like the study of what is sufficient for flight. Obviously, if one wants to understand flying animals, it helps to know something about flight in general; similarly, if one wants to understand the principles of flight, it helps to know something about the animals that fly. But the two areas have quite different objectives and methods.

In these terms, the study of learnability falls distinctly within the AI net. However, many central ideas in learnability often developed in parallel in both AI and cognitive science, chiefly the importance of *structured hypothesis spaces*. Many of these insights drew from European ethologists, whose work was becoming more known in the early 20th century. Eric Lenneberg, a crucial influence on the biological study of language, championed the idea of structural limits in learning since the early 1950s. Decades of work culminated in his 1967 landmark "Biological Foundations of Language" where he noted that "there is no possible way in which we could think of a device, natural or artificial, that is freed from all structural information" (p.394). Compare this to a later statement by Gleitman (1990), that "the trouble is that an observer who notices *everything* can learn *nothing*, for there is no end of categories known and constructable to describe a situation." Lenneberg understood the importance of hypothesis *classes* for learning, noting that "within the limits set, however, there are infinitely many variations possible. Thus the outer form of languages may vary with relatively great freedom, whereas the underlying type remains constant" (p.374).

These deep connections between computation, knowledge, and learnability prominently feature *language*. They are important and foreshadow another synthesis, this time regarding how natural languages are learned and acquired by children.

2.2 The linguistic synthesis

The early work of Noam Chomsky drew on the best of these exploratory ideas about computation, language, knowledge, and learnability. Chomsky catalyzed the study of learnability in linguistics by placing it as a central pillar of linguistic theory. Chomsky established that any sufficient theory of language must not only adequately describe linguistically significant generalizations, but also account for the ability of a child to learn these generalizations and infer the underlying linguistic forms despite sparse and underspecified data, the so-called *Poverty of the Stimulus*. Chomsky states the goal as follows:

> To learn a language, then, the child must have a method for devising an appropriate grammar, given primary linguistic data. As a precondition for language learning, he must possess, first, a linguistic theory that specifies the form of the grammar of a possible human language, and, second, a strategy for selecting a grammar of the appropriate form that is compatible with the primary linguistic data. As a long-range task for general linguistics, we might set the problem of developing an account of this innate linguistic theory that provides the basis for language learning. (Note that we are again using the term "theory" — in this case "theory of language" rather than "theory of a particular language" — with a systematic ambiguity, to refer both to the child's innate predisposition to learn a language of a certain type and to the linguist's account of this.) To the extent that a linguistic theory succeeds in selecting a descriptively adequate grammar on the basis of primary linguistic data, we can say that it meets the condition of explanatory adequacy. (Chomsky, 1965, pp.23-24):

Chomsky's other early work relevant to learnability is his contribution to the theory of formal languages, most famously "Three models for the description of language" (Chomsky, 1956). There he attempted to situate grammatical string patterns characterizing English syntax within the theory of recursively enumerable functions, and demonstrated the need for a certain power of machine to generate such a set. The three machines defined there—Finite-State Markov Processes, Phrase Structure Grammars, and Transformational Grammars—do not correspond directly to the classes in the Chomsky hierarchy, which appeared more or less in its present form in (Chomsky, 1959)(see Fig.1). They are theories for the explanation of linguistic phenomena rather than precise mathematical models. Chomsky stressed the idea of generative grammar: not just diagramming utterances in a language but actually providing a mechanism for generating all and only the sentences of the language. Extensions of this work have shown that natural language patterns carve out particular niches in the Chomsky hierarchy, and emphasize the importance of considering finer classes of formal grammars and the languages they generate when characterizing natural language (for more see Jäger & Rogers (2012))

Chomsky has been very clear on the importance of formal languages and grammars for linguistic theory and explanatory adequacy. In a well-known debate on language and learning (Piattelli-Palmarini, 1980), he states,

> that is exactly what generative grammar has been concerned with for twenty-five years: the whole complicated array of structures beginning, let's say, with finite-state automata, various types of context-free or context-sensitive grammars, and various subdivisions of these theories of transformational grammars—these are all theories of proliferating systems of structures designed for the problem of trying to locate this particular structure, language, within that system. So there can't be any controversy about the legitimacy of that attempt; in fact that is what all the work in formal linguistics has been about for a certain number of years.

Chomsky's synthesis showed that the importance of linguistic theory in the development of AI and cognitive science could not be ignored. Linguists concerned with any theoretical

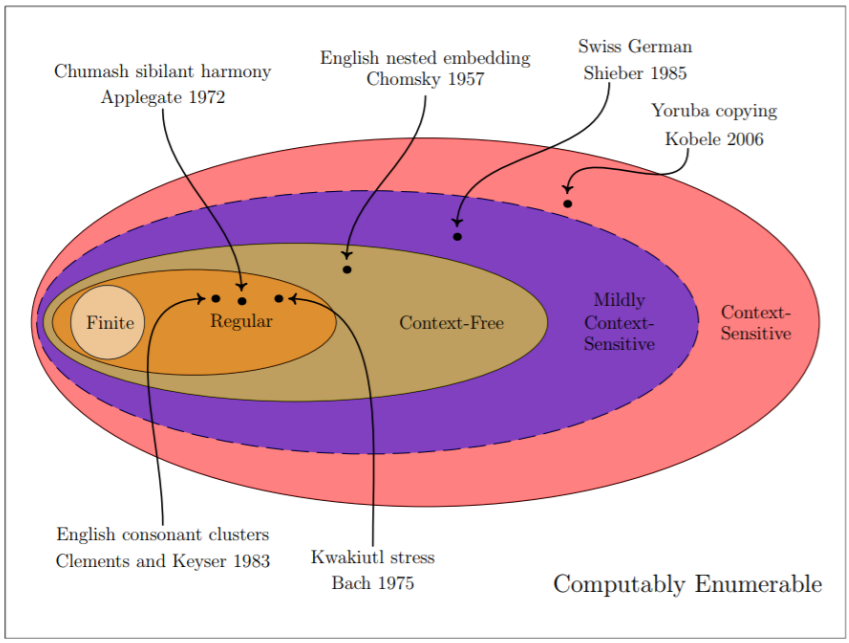

Figure 1: Natural Language Patterns within the Chomsky Hierarchy

concept had to consider the learnability of their grammars, and AI researchers concerned with learning had to consider *language*-learning, and the linguistic grammars and representations that go along with it. These considerations caused significant transformation in both fields, and the subsequent enthusiasm surrounding them quickly led to a plethora of learning frameworks and results, both in phonology and more generally.

3 Borrowing outranks invention

Given the central role language plays in cognitive science and artificial intelligence, it is not unreasonable to think that linguists may be leading scientists in the development and understanding of machine learning algorithms. However it appears that the rule is to borrow ideas, and the exception is to develop new ones.

Consider the numerous approaches to learning morpho-phonological grammars and generalizations. There is research drawing from connectionism (Rumelhart & McClelland, 1986; Goldsmith, 1994), decision trees (Ling, 1993), information theory (Goldsmith & Riggle, 2012), minimum description length (Goldsmith, 2001; Rasin & Katzir, 2016), Bayesian methods (Goldwater & Johnson, 2004; Cotterell, Peng, & Eisner, 2015), linear programming (Potts, Pater, Jesney, Bhatt, & Becker, 2010), statistical inference methods such as maximum likelihood (Jarosz, 2006) and maximum entropy (Goldwater & Johnson, 2003; Hayes & Wilson, 2008), grammatical inference (Gildea & Jurafsky, 1996; Heinz, 2010*a*; Chandlee, Eyraud, & Heinz, 2014), the Kullback-Leibler dissimilarity measure (Peperkamp, Calvez, Nadal, & Dupoux, 2006; Calamaro & Jarosz, 2015), and Markov logic networks (Vu *et al.*, 2018). This is just a small sampling of the work that has been done. Again, for more comprehensive reviews readers are referred to (Tesar, 2007; Albright & Hayes, 2011; Heinz &

Riggle, 2011; Jarosz, 2019). All of these methods were initially developed outside of linguistics, although some, such as grammatical inference, can trace their origin to the fundamental problem of language acquisition (de la Higuera, 2010). The work of the phonologist studying learning has thus primarily been to figure out how to synthesize and apply methods developed elsewhere to learning problems where the targets are phonological grammars.

Of course there have been original studies on learning phonological grammars. These studies occure across grammatical formalisms including the the principles and parameters framework (Dresher & Kaye, 1990), rule learning (Johnson, 1984; Albright & Hayes, 2003; Simpson, 2010), morpho-phonological paradigm learning (Hulden, Forsberg, & Ahlberg, 2014) and Optimality Theory (Tesar & Smolensky, 1998, 2000; Tesar, 1995, 1998*b*,*a*). Tesar and Smolensky's Recursive Constraint Demotion (RCD) algorithm is perhaps the best example (Tesar & Smolensky, 1998, 2000). Tesar and Smolensky (1996: 26) show that the structure that ranked constraints give to the hypothesis space guarantees that RCD will converge to a consistent grammar with a polynomial mistake bound (unlike a brute-force enumeration). This work inspired much subsequent research in the learning of phonological grammars as scholars studied variants of RCD (Magri, 2009, 2013), learning in stochastic versions of OT (Boersma, 1997; Boersma & Hayes, 2001), and the learning of other constraint-based formalisms referenced above.

To be clear, we are not saying good ideas in neighboring fields should be ignored. Instead when we consider the history of learning ideas in phonology, we are making the following observations. First, that the original learning ideas within phonology are relatively few. Second, these ideas, as good as they are, do not appear to have impact beyond the field of phonology. Third, many ideas from machine learning, statistical inference, and artificial intelligence are borrowed into phonology, often after being imported into neighboring fields such as natural language processing (as was the case with maximum entropy and statistical inference more generally), cognitive science (Bayes), or information theory (Minimum Description Length).

Given this history we can expect a deluge of papers using "deep learning" (Goodfellow, Bengio, & Courville, 2016; Goldberg, 2017) to enter the phonological literature in the forthcoming years. Indeed, they are already prominently featured, for example, at the Morphological Re-inflection Challenge (Cotterell *et al.*, 2016, 2017) organized by the Association for Computational Linguistics Special Interest Group in Computational Phonetics, Phonology and Morphology (ACL-SIGMORPHON).

4 A Short Guide to Hard Problems

Throughout its history, learnability has emphasized the *nature* of learning, what is sufficient for learning, under which conditions, and for which criteria of success. Computational learning theory provides definitions of what it means to learn and asks, under those definitions: What can be learned, how so, and why? Determining which definitions are "correct" or "best" for a given scenario are major issues.

We want to emphasize the import of learnability results for all branches of linguistics, independent yet characteristic of all terminology, theories, and perspectives. Specifically, learning theory provides non-trivial conditions of explanatory adequacy on theories of natural

language. Osherson & Weinstein (1983, p. 37) put the claim this way:

> For a class of languages to be the natural languages, the class must be learnable by children on the basis of the kind of linguistic exposure typically afforded the young. Call this the learnability condition on the class of natural languages. Formal learning theory is an attempt to deploy precise versions of the learnability condition in the evaluation of theories of natural language. In the present context, such a theory will specify (a) the kind of linguistic input available to children, (b) the process by which children convert that experience into successive hypotheses about the input language, and (c) the criteria for "internalization of a language" to which children ultimately conform. From (a)-(c) it should be possible to deduce (d) the class of languages that can be internalized in the sense of (c) by the learning mechanism specified in (b) operating on linguistic input of the kind characterized in (a). Such a theory is correct only if (d) contains exactly the natural languages.

What makes a class of languages learnable? In this section, we review some of the most important problems and frameworks to come out of formal learning theory. These results have permeated almost every theory of phonological learning, and indeed every problem of learning related to language structure.

4.1 The Logical Problem of Language Acquisition

What is called the logical problem of language acquisition is the general problem of generalizing beyond one's experience in the context of language-learning. Pinker (2004) explains.

> The problem, as with all problems of induction, is that an infinite number of generalizations are consistent with any finite sample of data. Many curves can be drawn through a set of points, many laws are consistent with a set of observations, and many grammars are consistent with a set of sentences. Therefore any learner who correctly induces a function, theory, or grammar must respect prior ('innate') constraints on its hypothesis space; the data alone are insufficient. This is a logical point which cannot be denied by any theory, nativist, empiricist, behaviourist, connectionist, constructivist, or emergentist (Quine, 1969). For the behaviourists, the innate constraints reside in the generalization gradients and response classes. For the connectionists, they reside in the features defining the units and the topology of the networks. For Chomskyans, they reside in categories, operations, and principles.

Thus, any debate that exists regards the *character* of the innate constraints required for learning, not the *existence* of such innate constraints themselves.

This logical point emerges concretely in formal learnability results. For example, in the Identification in the Limit frameworks (Gold, 1967), the learner is said to identify a class of languages if it identifies in the limit every member language of the class when there are no limits on the learner's computational resources or time, and the input is assumed to be a finitely long initial portion of an infinitely long noise-free data stream called a 'text'. In this

framework, there is provably no learning algorithm that can learn every finite language and any one infinite language from positive examples. In the Probably Approximately Correct (Valiant, 1984) learning frameworks, not even the class of finite languages is learnable. Under these learning definitions, in order to learn linguistic generalizations like the kinds found in natural languages, the innate constraints must demarcate classes of formal languages which do not include every finite language. This is what we expect: given some finite sets of experiences, learners infer grammars which generate languages that *necessarily go beyond* this experience.

Thus the first point is that an explanatorily adequate linguistic theory will not only identify the nature of the innate constraints, but show how they, along with linguistic data, lead to the desired grammars. In their quest for explanatory adequacy, many linguists and phonologists usually assume the innate constraints are whatever the current theory of Universal Grammar (UG) is, and take their task as establishing a program that takes linguistic data and the hypothesis space given by UG and returns the target phonological grammar. The current theory of UG is usually one that has been determined on the basis of traditional lingustic analysis, in which learnability principles have played no part. As we will see, it is only recently that phonologists have begun to consider hypotheses spaces primarily governed by learning considerations instead of traditional linguistic analysis.

A second point is that mathematical definitions of learnability provide a concrete way to study what kinds of concepts and linguistic patterns are learnable with what kinds of innate constraints and with what kinds of data.

4.1.1 The subset problem and tell-tale sets

The Subset Problem was first formulated for phonology in (Dell, 1981), and has been widely discussed since (see (Berwick, 1985; Hale & Reiss, 2000) and many others). In a nutshell the problem refers to the situation which may be encountered in the course of learning where the learner's current grammatical hypothesis H is an overgeneralization of the target grammar G. As such H accepts strictly more than G so mathematically $L(G) \subseteq L(H)$ (see Figure 2). The problem is that no positive evidence can ever deter the learner from H since every positive data point consistent with G is also consistent with H. In these cases, it is all too easy for learning algorithms to arrive at grammars that classify the observed data as legal, while failing to classify the illegal forms as such.

If negative evidence (data points not consistent with G) were available, then such examples would be enough to push the learner away from hypothesis H. Indeed, Gold's 1967 learning result that the entire class of computable formal languages is learnable from positive and negative data uses negative evidence in precisely this way. So the absence of negative evidence makes the learning problem harder and the subset problem is one concrete instance of this difficulty.

Gold's work inspired an extremely productive research direction in computational learning theory. Working within this framework, Angluin (1980) defined a benchmark for necessary and sufficient structure in a class of formal languages. If every language L in a class contains a finite set S, where no other language L' in the class is simultaneously a superset of S and proper subset of L, then this hypothesis space is sufficiently structured such that identification in the limit from positive data can succeed. She calls such a finite set S a

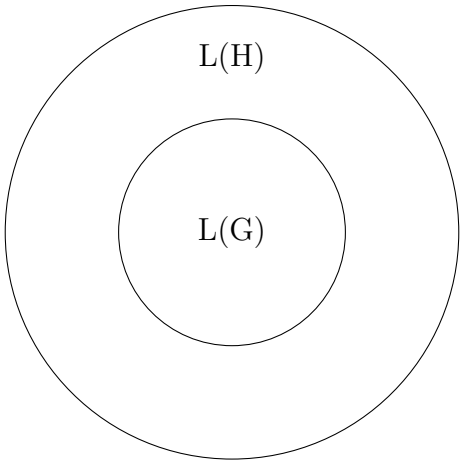

Figure 2: The Subset Principle

tell-tale set, and the above property of hypothesis spaces is the *tell-tale property*.

The tell-tale property is necessary and sufficient for learning, because a learner who guesses L after exposure to its tell-tale set is guaranteed to have hypothesized *the smallest language* in the class consistent with the data sample. In other words, they are guaranteed to never overgeneralize in the way described above. Conversely, if a learner always guesses the smallest language in the class consistent with the positive data sample, this learner is guaranteed to with a large enough sample (one that includes the tell-tale set). Characterizing the tell-tale sets of a hypothesis space, and more generally, the nature of the finite experience a leaner needs to generalize correctly to the patterns in a hypothesis space, is one of the important lessons of learning theory for linguistics. In fact, many learning algorithms for classes of formal languages learnable from positive data work this way (Angluin, 1982; Garcia, Vidal, & Oncina, 1990; Heinz, 2008), including ones arguably relevant to natural language phonology (Heinz, 2010*b*). What makes these classes learnable is that they are structured and organized in a natural way (Heinz, Kasprzik, & Kötzing, 2012).

Albright & Hayes (2002, 2003) proposed a model of phonological rule learning based on *minimal generalization*, which likewise should circumvent the subset problem (see also (Albright, 2009)). The idea is that as structural changes and their contexts are observed, they are generalized to the smallest natural classes that include them.

Many have investigated the subset problem in Optimality Theory (Hayes, 2004; Jarosz, 2006; Jesney & Tessier, 2011; Prince & Tesar, 2004; Tessier, 2009). Magri (2013) investigates the subset problem for Optimality Theory by asking how easy or hard is it to find a constraint ranking consistent with the data that generates a smallest language. An important early result in Optimality Theory discussed in more detail later was that there is an efficient method to find a constraint ranking with the data (Tesar & Smolensky, 2000), though it may not necessarily generate a smallest language. Magri finds that the OT subset problem is intractable in general and concludes that the "subset problem thus needs to be restricted to plausible typologies, and solution algorithms need to take advantage of the additional structure brought about by these typological restrictions." This lesson echoes the one from formal language theory that the class of formal languages—which corresponds to the typo-

logical language space—must have the right kind of structure to be learned from positive data.

4.2 The credit problem

As we have seen, a natural approach to learning is based on error-correction. If the observed sample of data is not consistent with a current hypothesis H, it is abandoned in favor a hypothesis that is consistent with the data. If the grammar generating a smallest language consistent with the data can be identified then this is a natural choice. However, it is not always the case that there is one smallest language in the class consistent with the data seen so far, nor is it always the case that a given positive sample of data includes a tell-tale set. This typically occurs when the grammatical space has been determined by descriptive linguistic considerations to the exception of learnability ones. In these cases, there is a puzzle: what hypothesis should H be replaced with?

The credit (or blame) problem (Clark, 1989) exacerbates this puzzle because covert structure prevents the learner from directly observing the source of this error in H (Dresher, 1999). When learning phonological mappings to underlying forms, errors can arise from hypothesizing the wrong lexical representation or the wrong phonological mapping, but the learner must still somehow determine which is the source of the error. For example, metrical footing is not present in the signal, so if a learner notices an error, which aspects of the grammar must it assign credit to? It is not immediately clear.

This is not unlike the problem determining which, of several possible sources account for given observations. Here is a concrete case. Say a learner observes a trisyllabic word with stress on the medial syllable. Should the learner hypothesize left-aligned iambs or right-aligned trochees?

Several learning proposals in phonology explicitly engage the credit problem, notably Dresher & Kaye (1990); Tesar & Smolensky (2000); Jarosz (2006, 2013) and (Tesar, 2014). Jarosz (2019) makes the argument that statistical inference enables inferences about sources of error/blame (see also Nazarov & Jarosz (2017) on learning interdependent parameters for stress). Many have also observed that the credit problem can be solved by deterministic grammatical models. Dresher & Kaye (1990) explain.

> While there are obvious computational advantanges to deterministic parsing, Berwick (1985) argues that it also aids acquisition. Thus, suppose the parser encounters a sentence containing a new structure unknown to it; the parse will fail. The parser will then attempt to modify its grammar so as to accommodate the new data; but to do that successfully, it is necessary to know where the failure occurred. A nondeterministic parser, which routinely can make use of unlimited backtracking, will characteristically fail backwards through the whole sentence, undoing correct as well as incorrect substructures. The same is not true of a deterministic parser; as Berwick and Weinberg point out (1984, p. 231), "from the standpoint of learning, the effect of determinism and the restriction that rules refer only to bounded context is to pinpoint errors to a 'local radius' about the point at which the error is detected," The ability to keep problems local aids in the learnability of grammars.

(Dresher & Kaye, 1990, pp.161-162)

A similar point is made by Heinz, de la Higuera, & van Zaanen (2015). One concrete example comes from the class of finite-state transducers. If this class includes non-determinstic transducers then no formal learning results are known to exist and furthermore in some learning frameworks (Identification in the Limit from Positive Data and PAC), it is provable that no such learning algorithms exist. On the other hand if the class of finite-state transducers is limited to deterministic ones then the algorithm OSTIA is guaranteed to succeed (Oncina, García, & Vidal, 1993). Similar results hold for classes of deterministic versus non-deterministic finite-state probability distributions over strings (de la Higuera, 2010).

5 Tensions, contemporary and ancient

5.1 Typology and learnability in grammar design

Most linguists agree that linguistic theories should satisfy a learnability condition, so that the set of grammars possible in an adequate linguistic theory must be such that any of them are acquirable by children on the basis of the kind of linguistic experience they typically get. However, many linguists tend to dismiss the claim that formal learning theory can provide non-trivial adequacy conditions on linguistic theories. Chomsky (1965), in summarizing his subchapter on "Linguistic theory and language learning," captures the tension between typological coverage and learnability considerations:

> The real problem is that of developing a hypothesis about initial structure that is sufficiently rich to account for acquisition of language, yet not so rich as to be inconsistent with the known diversity of language. (p. 58)

Linguistic typology is central to language learnability. Linguistic theory proposes a class of possible human languages, which is effectively the class of allowable structures for the learning problem. It is a reasonable to expect that the formal relationships among the allowable grammars that make language learnable are intimately related to the formal properties of the linguistic theory defining the class of allowable grammars (Dresher, 1999; Heinz, 2010*a*).

However, learnability is about classes: it is a property of classes of grammars, and by extension, languages. A class is learnable if an algorithm exists which can select, based on input data, the target grammar (the one generating that data), for data from every grammar in the class. What makes learning easier or harder, as noted above, is the amount of data and time needed to robustly identify the target grammar out of the class of possible grammars. The challenge arises from the need to distinguish between different possible grammars, rather than from the individual grammars themselves.

Still, learnability oftentimes finds itself at odds with linguistic theories. Tesar (1995, pp. 108-109), explaining the Principles & Parameters triggering approach to learning as a motivation for a learnability account in OT, captures the tension elegantly:

> It is significant that a trigger provides information about the value of a single parameter, rather than relationships between the values of several parameters

> . . . The result is that learnability concerns in this framework favor parameters which are independent: they interact with each other as little as possible, so that the effects of each parameter setting can be distinguished from the effects of the other parameters. In fact, this property of independence has been proposed as a principle for grammars (Wexler & Manzini 1987). Unfortunately, this results in a conflict between the goals of learnability, which favor independent parameters with restricted effects, and the goals of linguistic theory, which favor parameters with wide-ranging effects and greater explanatory power.
>
> Optimality Theory may provide the opportunity for this conflict to be avoided . . . The Constraint Demotion learning algorithm not only tolerates constraint interaction, but is based upon it. Informative data provide information not about one constraint in isolation, but about the results of interaction between constraints. Constraints which have wide-ranging effects benefit learnability. If these properties are successfully preserved in a more general account of language learning within Optimality Theory, explanation and learnability will work together; they will both favor interacting constraints with wide-ranging effects and explanatory power.

While Tesar is attempting to eradicate the tension by presenting the theory of constraint interaction as a remedy, this is less a dismissal of learnability than an appeal to situate it in a particular theory. The results and problems brought forth to phonology do not disappear within constraint interaction. In fact, many of the issues Tesar mentions, in addition to those outlined earlier, are still the subject of ongoing debate. Alexander's Clark research, for example, insists that learnability ought to be a primary factor in the development of linguistic theories (Clark, 2010; Clark & Lappin, 2011).

All this serves to highlight, once again, the independent nature of learnability results. However, the AI perspective offered earlier rears its head, as learnability must bend the knee to typology and vice versa. Whether learnability is taken to be a feature of linguistic theory or an additional constraint atop it remains to be seen.

5.2 Methodological Choices and Research Goals

In addition to concerns over what is being modeled, another tension that pervades the literature in phonological learning is more methodological. Researchers necessarily make many choices when studying learning, from restrictions in the scope and naturalness of the phonological data, to the scope of generalizations, and the nature of how they are inferred. In principle, these choices can complement each other. Niyogi (2006) distinguishes between mathematical descriptions of learning and computational learning, and points out that

> Mathematical models with their equations and proofs and computational models with their programs and simulations provide different and important windows of insight into the phenomena at hand. In the first, one constructs idealized and simplified models but one can now reason precisely about the behavior of such models and therefore be sure of one's conclusions. In the second, one constructs more realistic models but because of the complexity, one will need

> to resort to heuristic arguments and simulations. In summary, for mathematical models the asumptions are more questionable but the conclusions are more reliable — for computational models, the assumptions are more believable but the conclusions more suspect.

Let's take each of these in turn. On the mathematical side, one can readily study the abstract characteristics of formal languages in terms of their learnability and apply the results to natural language learning. This often involves two parts. One can show that a particular subclass of grammars is learnable in some conditions. Then one can show that a particular set of patterns uniquely characteristic of human phonology sits in that class under certain representational conditions. In phonology, the patterns and transformations from underlying to surface forms inhabit subclasses of the regular languages and relations. These subclasses are both computationally simple, feasibly learnable in the limit from positive data, and more easily learnable by humans in learning experiments (see Heinz (2018) for an overview).

At the same time, increasing accessability, ease, and power of computation enabled the use of extensive simulation studies in learning, both in phonology and in linguistics more generally. Even Bradfield (2017), who strongly echoes Niyogi's cautions about the robustness of conclusions drawn from simulations, notes that simulation and modeling is a necessary part of science and that, within phonology, it has encouraged the incorporation of continuous parameters in vowel phonology, the use of probabilistic and stochastic models, agent-based models of phonological change, and iterative models of phonological learning in the individual, among others.

Much work in phonological learning has shown that using simulations can be useful. If the possible model proposed is sufficiently complex, it may be surprising and not obvious that it can describe or account for some phenomenon. Additionally, simulations may show which theories or models cannot account for a particular phenomena. Bradfield (2017) notes that in general, more comprehensive testing of analyses and theories in phonology is necessary to proceed with simulations in an effective and clarifying way. He states,

> . . . if simulations are designed with careful analysis of the underlying theories, analyses of the sources of error, and the rest of the apparatus usual in physical and engineering science simulation studies, and then simulations are conducted over a wide range of possible configurations and parameter settings, one might establish with some confidence what is not possible, as well as what is possible; and one may even produce numerical results, testable for agreement with real empirical data.

Many of the core results, problems, and theories characteristic to learnability showcased in this chapter have been justified on mathematical grounds. While studies of the learnability of formal languages are relevant to phonological theory, they are also in a sense independent of it. The use of formal languages is an abstraction intended not to eliminate detail, but instead to clarify the contributions that particular assumptions make. Phonologists who are uncomfortable with such abstractions may prefer to work with data collected from corpora or experimentation. For example, a successful research program developed within phonology trains learning models on corpora intended to simulate an adult's lexicon and then tests these models by comparing their performance on test data with humans' performance on the

same test data in behavioral experiments (Albright & Hayes, 2003; Hayes & Londe, 2006). The underlying idea is of course attractive: if the learning model behaves like the human subjects on the test then perhaps the human subjects have an internal learning mechanism similar to the one represented by the learning model.

Heinz & Riggle (2011, pp.67–68) explain the tension:

> A shift in focus from the analysis of properties that define various learnable classes of languages to the behavior of humans is undoubtedly appealing to any who feel that the results of learnability theory are too abstract and remote from real-world learning problems. On the other hand, having observed that an algorithm A and human subject H give similar responses for a particular set of test items T after being exposed to a set of training data D, it is not clear what we can conclude about H or the relationship between A and H, because they might wildly diverge for some other data T' and D'. The goal of determining which properties of the data critically underlie learnability — or in this case the correlation between A and H — is precisely why learning theory focuses mainly on the *properties of classes of languages* or the *general behavior of specific algorithms*, as opposed to the specific behavior of specific algorithms. [emphasis in original]

The distinction between natural phonological data and artificial data designed around representational or computational assumptions has caused much methodological angst. On the one hand, the search for the principles underlying phonological learning involves fixing representational and computational assumptions them in order to determine their effects on a well-understood class of learning problem. The reason for this is simple: it is desirable to know with certainty that a learning algorithm will succeed, and under what conditions, before testing it on data. Computational science generally is about problems and the algorithms that solve them reliably, correctly, and with so much resources. For example, no one would use a sorting algorithm in a software library if there was no proof of its correctness. The proof of correctness may be derived from a specification of the program, but neither the program itself nor simulations of it on data are sufficient to derive it.

However, many take the study of phonological learning to only encompass testing models on naturalistic, or experimentally obtained, data. The reason for this is also simple: the explananda of phonological learning is phonology itself, and all of the interactions between phonological phenomena a learner faces. The attitude seems to be that if a model performs well according to some evaluation, then the learning problem has been adequately addressed. To us, this view seems shortsighted. Of course it is important to obtain such results, but computational science, as explained above, demands more. Furthermore, there is the important point that all data is "cooked" rather than "raw" (Hammarberg, 1981), meaning it involves some abstraction on the part of the researcher. This means the debate over abstractness and idealization has only to do with *how much* and *where* researchers prefer abstraction in phonological learning, since all accept abstraction at some level. The disagreement is not about abstraction per se, but the right level of it.

5.3 The psychological reality of grammar

Since this beginning, the nature of phonological discovery and learning has continually run into the question of whether results in this domain have discernible mental content. Since the generative program in phonology explicitly initiated by Chomsky & Halle (1968), the conditions of psychological reality as a measure of an explanatorily adequate theory of phonological knowledge have created another tension in phonological learnability research. How should the results of learnability be interpreted with respect to the cognitive content of phonological grammar?

The first discussion of "psychological reality" in linguistics comes in fact from phonology, in Edward Sapir's 1933 "La Réalité Psychologique des Phonèmes" . Considering first what is often called "linguistic evidence," Sapir constructed an abstract system of rules and underlying representations that offered a plausible account of the linguistic data. Sapir then argued that if the phonemes revealed by linguistic analysis are psychologically real (as he assumes they are), then we ought to see them reflected, not just in grammatical patterns, but also in various other types of behavior. He turned to perceptual tests, whose outcome convinced him that his theoretical constructions had "psychological reality." As Dresher (2010) notes, Sapir is demonstrating the utility of corpus-internal evidence (the synchronic patterning of sounds) that the phoneme is real, and additional evidence (external to the synchronic pattern). The novelty of his approach is underscored by the fact that Sapir wrote it at a time when the term "psychological" was "so offensive in North America that he published the paper in France" (Dresher, 2010, p. 2).

Thus another tension is the one between the distinctly mentalist approach, which favor mental structures first which describe a range of typological or language data from linguistic evidence, and a behaviorist approach which favors models which capture results and generalizations from task-based experiments. Sapir's approach, as Dresher (2010) notes, is often misinterpreted, and has led over the years to "the subsequent practice of dividing evidence into two types: evidence which does not bear on psychological reality—synchronic and diachronic patterns, i.e. conventional linguistics—and evidence which does—evidence from acquisition, aphasia, speech errors, or any experiment which requires special apparatus or clothes, or consent forms." Dresher concludes from this that discussion of psychological reality as distinct from linguistic reality is a red herring.

The importance of experimental tasks in defining psychological reality is a crucial part of this tension. This is especially true as the majority of successes claimed within machine learning center around benchmark performance on various tasks that are said to model real-world tasks emblematic of the sort intelligent creatures perform. It is notable that this tension pervades all facets of cognition. In setting out a program to unite cognitive science, neuroscience, and AI, Kriegeskorte & Douglas (2018) assert that "understanding brain information processing requires that we build computational models that are capable of performing cognitive tasks." In this view, one's theory of learning, should model say, the quantitative nature of judgments given by learners on a forced-choice discrimination task on various phonotactic patterns. The patterns themselves can only serve to aid the experiment design.

In contrast, early generative work considered linguistic data by itself to be sufficient evidence of psychological reality—that is, the kinds of language patterns seen in data reflect a

grammar capable of generating them, a grammar which has been internalized and learned. In this view, it is sufficient for a theory of learning to consider various abstractions over linguistic data which may be structured in a mathematically rigorous way, leading to learnability characterizations. This is the path pursued by almost all work in formal learning theory and its applications to linguistic patterns.

The relationship between grammars and psychology seems here to stay in generative phonology, and so results in learnability will continue to confront this issue. We close this section with a remark from Zenon Pylyshyn's 1984 book discussing the nature of cognition and of cognitive theories. He cautions that "the kinds of theories cognitive scientists entertain are intimately related to the set of tacit assumptions they make about the very foundations of the field of cognitive science. In cognitive science the gap between metatheory and practice is extremely narrow."

6 Conclusion

The brief history of learnability in phonology presented here showcases a field undergoing rapid growth and change. Such a field traces its issues and inspirations as far back as the development of the foundations of mathematics and through subsequent computational and cognitive turns. The problems it has presented permeate every theory of phonological learning, and will continue to do so. While much of the field has borrowed techniques from the outside, ther are notable contributions developed within the phonological tradition proper. The strong tensions that inward glances into such a tumultuous field reveal reflect strong divides on linguistic typology, methodology, and the views of cognitive reality. These tensions are perhaps inevitable given the range of tacit assumptions entertained by researchers of language and the mind.

We want to encourage phonologists to be more directly involved in the science of learning. The fundamental problems in learning raised in this chapter are ones that researchers face in many fields apart from phonology, but we believe phonology provides a concrete, knowledge-rich domain in which solutions to those problems can be developed and studied. In other words we urge linguists interested in the learning problem to not just borrow ideas from the machine learning, artificial intelligence and statistical inference literatures, but to actively contribute to them.

Lastly, we hope that phonologists take the long view. The multitude of approaches and insights resulting from and required of work in learnability should be embraced and not a source of division. In our view, a complete, computationally explicit theory of the acquisition of phonological grammars will not be the product of singular efforts but will instead be the product of many results from many methodologies. And for the phonologists perplexed by all the fuss about learning, we hope they look at this brief history and let the flowers bloom.

7 Acknowledgments

We would like to thank Jane Chandlee, Elan Dresher, Gaja Jarosz, and Harry van der Hulst for helpful comments on this chapter.

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26(1): 9–41.

Albright, Adam and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the sixth meeting of the acl special interest group in computational phonology*, 58–69. Somerset, NJ: Association for Computational Linguistics.

Albright, Adam and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90: 119–161.

Albright, Adam and Bruce Hayes. 2011. Learning and learnability in phonology. In Jason Riggle John Goldsmith and Alan Yu (eds.), *Handbook of phonological theory*, 661–690.

Angluin, Dana. 1980. Finding patterns common to a set of strings. *Journal of Computer and System Sciences* 21: 46–62.

Angluin, Dana. 1982. Inference of reversible languages. *Journal for the Association of Computing Machinery* 29(3): 741–765.

Applegate, R.B. 1972. *Ineseño Chumash grammar*. Ph.D. thesis, University of California, Berkeley.

Bach, Emmon. 1975. Long vowels and stress in Kwakiutl. *Texas Linguistic Forum* 2: 9–19.

Beard, Alex. 2018. How babies learn — and why robots can't compete. *The Guardian* .

Berwick, Robert. 1985. *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

Berwick, Robert and Amy Weinberg. 1984. *The grammatical basis of linguistic performance*. Cambridge, MA: MIT Press.

Bloomfield, Leonard. 1926. A set of postulates for the science of language. *Language* 2(3): 153–164.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* 21. University of Amsterdam.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32: 45–86.

Bradfield, Julian. 2017. The sound of a spherical cow. *Phonology* 34(2): 347–362.

Calamaro, Shira and Gaja Jarosz. 2015. Learning general phonological rules from distributional information: A computational model. *Cognitive Science* 39(3): 647–666.

Carnap, Rudolf. 1928. Der logische aufbau der welt: Versuch einer konstitutionstheorie der begriffe. *Welt-Kreis, Berlin* .

Cauchy, Augistin-Louis. 1821. *Analyse algebrique*. Paris, France: Debure Frères.

Chandlee, Jane, Rémi Eyraud, and Jeffrey Heinz. 2014. Learning strictly local subsequential functions. *Transactions of the Association for Computational Linguistics* 2: 491–503.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 113–124. IT-2.

Chomsky, Noam. 1959. On certain formal properties of grammars. *Information and Control* 2: 137–167.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Church, Alonzo. 1932. A set of postulates for the foundation of logic. *Annals of Mathematics* 33(2): 346–366. Series 2.

Clark, Alexander. 2010. Three learnable models for the description of language. In Carlos Martín-Vide Adrian-Horia Dediu, Henning Fernau (ed.), *Language and automata theory and applications, fourth international conference, lata 2010*, LNCS, 16–31. Springer.

Clark, Alexander and Shalom Lappin. 2011. *Linguistic nativism and the poverty of the stimulus*. Wiley-Blackwell.

Clark, Robin. 1989. On the relationship between the input data and parameter setting. In Juli Carter Rose-Marie Dèchaine (ed.), *North eastern linguistic society (nels) 19*, 48–62. Amherst, MA: University of Massachusetts, Graduate Linguistic Student Association.

Clements, G. N. and Samuel Jay Keyser. 1983. *CV Phonology: A Generative Theory of the Syllable*. Cambridge, Mass.: MIT Press.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. Conll-sigmorphon 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the conll sigmorphon 2017 shared task: Universal morphological reinflection*, 1–30. Vancouver: Association for Computational Linguistics.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The sigmorphon 2016 shared task: Morphological reinflection. In *Proceedings of the 14th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, 10–22. Berlin, Germany: Association for Computational Linguistics.

Cotterell, Ryan, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics* 3: 433–447.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12(1): 31–37.

Dresher, B Elan. 2010. There's no reality like psychological reality. *Toronto Working Papers in Linguistics* 32.

Dresher, Elan. 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry* 30: 27–67.

Dresher, Elan and Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 34: 137–195.

Garcia, Pedro, Enrique Vidal, and José Oncina. 1990. Learning locally testable languages in the strict sense. In *Proceedings of the workshop on algorithmic learning theory*, 325–338.

Gildea, Daniel and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics* 24(4): 497–530.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1(1): 3–55.

Gold, E.M. 1967. Language identification in the limit. *Information and Control* 10: 447–474.

Goldberg, Yoav. 2017. *Neural network methods for natural language processing*. Morgan and Claypool Publishers.

Goldsmith, John. 1994. A dynamic computational theory of accent systems. In Jennifer Cole and Charles Kisseberth (eds.), *Perspectives in phonology*, 1–28. Stanford: Center for the Study of Language and Information.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computatonal Linguistics* 27(2): 153–198.

Goldsmith, John and Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language and Linguistic Theory* 30(3): 859–896.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl (eds.), *Proceedings of the stockholm workshop on variation within optimality theory*, 111–120.

Goldwater, Sharon and Mark Johnson. 2004. Priors in bayesian learning of phonological rules. In *Proceedings of the 7th meeting of the acl special interest group in computational phonology: Current themes in computational phonology and morphology*, 35–42.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. The MIT Press.

Greene, Lane. 2017. Machine translation: Beyond babel. *The Economist* .

Hale, Mark and Charles Reiss. 2000. Substance abuse and dysfunctionalism: Current trends in phonology. *Linguistic Inquiry* 31: 157–169.

Hammarberg, Robert. 1981. The cooked and the raw. *Journal of Information Science* 3(6): 261–267.

Hayes, Bruce. 2004. Phonological acquisition in optimality theory: the early stages. In Rene Kager, Joe Pater, and Wim Zonneveld (eds.), *Fixing priorities: Constraints in phonological acquisition*. Cambridge University Press.

Hayes, Bruce and ZsuZsa Londe. 2006. Stochastic phonological knowledge: the case of hungarian vowel harmony. *Phonology* 23(1): 59–104.

Hayes, Bruce and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379–440.

Heinz, Jeffrey. 2008. Left-to-right and right-to-left iterative languages. In Alexander Clark, François Coste, and Lauren Miclet (eds.), *Grammatical inference: Algorithms and applications, 9th international colloquium*, vol. 5278 of *Lecture Notes in Computer Science*, 84–97. Springer.

Heinz, Jeffrey. 2010*a*. Learning long-distance phonotactics. *Linguistic Inquiry* 41(4): 623–661.

Heinz, Jeffrey. 2010*b*. String extension learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 897–906. Uppsala, Sweden: Association for Computational Linguistics.

Heinz, Jeffrey. 2018. The computational nature of phonological generalizations. In Larry Hyman and Frans Plank (eds.), *Phonological typology*, Phonetics and Phonology, chap. 5, 126–195. De Gruyter Mouton.

Heinz, Jeffrey, Colin de la Higuera, and Menno van Zaanen. 2015. *Grammatical inference for computational linguistics*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Heinz, Jeffrey, Anna Kasprzik, and Timo Kötzing. 2012. Learning with lattice-structured hypothesis spaces. *Theoretical Computer Science* 457: 111–127.

Heinz, Jeffrey and Jason Riggle. 2011. Learnability. In Marc van Oostendorp, Colin Ewen, Beth Hume, and Keren Rice (eds.), *Blackwell companion to phonology*. Wiley-Blackwell.

de la Higuera, Colin. 2010. *Grammatical inference: Learning automata and grammars*. Cambridge University Press.

Hilbert, David. 1928. Die grundlagen der mathematik. In *Die grundlagen der mathematik*, 1–21. Springer.

Hofstadter, Douglas. 2018. The shallowness of Google Translate. *The Atlantic* .

Hulden, Mans, Markus Forsberg, and Malin Ahlberg. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics*, 569–578. Gothenburg, Sweden: Association for Computational Linguistics.

Hutson, Matthew. 2018. Hackers easily fool artificial intelligences. *Science* 361(6399): 215–215. doi:10.1126/science.361.6399.215. URL `http://science.sciencemag.org/content/361/6399/215`.

Jäger, Gerhard and James Rogers. 2012. Formal language theory: Refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B* 367(1598): 1956–1970.

Jarosz, Gaja. 2006. *Rich lexicons and restrictive grammars – maximum likelihood learning in Optimality Theory*. Ph.D. thesis, Johns Hopkins University.

Jarosz, Gaja. 2013. Learning with hidden structure in optimality theory and harmonic grammar: Beyond robust interpretive parsing. *Phonology* 30(1): 27–71.

Jarosz, Gaja. 2019. Computational modeling of phonological learning. *Annual Review in Linguistics* .

Jesney, Karen and Anne-Michelle Tessier. 2011. Biases in harmonic grammar: the road to restrictive learning. *Natural Language & Linguistic Theory* 29(1): 251–290.

Johnson, Mark. 1984. A discovery procedure for certain phonological rules. In *Proceedings of 10th international conference on computational linguistics and 22nd annual meeting of the association for computational linguistics*, 344–347.

Kleene, Stephen Cole. 1952. *Introduction to metamathematics*, vol. 483. van Nostrand New York.

Kobele, Gregory. 2006. *Generating copies: An investigation into structural identity in language and grammar*. Ph.D. thesis, University of California, Los Angeles.

Kriegeskorte, Nikolaus and Pamela K Douglas. 2018. Cognitive computational neuroscience. *Nature Neuroscience* 1148–1160.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553): 436–444.

Lenneberg, Eric. 1967. *Biological foundations of language*. Oxford, England: Wiley.

Levesque, Hector. 2017. *Common sense, the turing test, and the quest for real ai*. MIT Press.

Lewis-Kraus, Gideon. 2016. The great A.I. awakening. *The New York Times Magazine* .

Ling, Marinov M., C. X. 1993. Answering the connectionist challenge: a symbolic model of learning the past tenses of english verbs. *Cognition* 49: 235–290.

Magri, Giorgio. 2009. *A theory of individual-level predicates based on blind mandatory implicatures. Constraint promotion for Optimality Theory.* Ph.D. thesis, MIT.

Magri, Giorgio. 2013. The complexity of learning in OT and its implications for the acquisition of phonotactics. *Linguistic Inquiry* 44(3): 433–468.

Marcus, Gary and Ernest Davis. 2018. A.I. is harder than you think. *New York Times* A21.

Nazarov, Aleksei and Gaja Jarosz. 2017. Learning parametric stress without domain-specific mechanisms. In *Proceedings of the annual meetings on phonology*, vol. 4.

Niyogi, Partha. 2006. *The computational nature of language learning and evolution.* Cambridge, MA: MIT Press.

Oncina, José, Pedro García, and Enrique Vidal. 1993. Learning subsequential transducers for pattern recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15: 448–458.

Osherson, Daniel and Scott Weinstein. 1983. Formal learning theory. In M. Gazzaniga and G. Miller (eds.), *Handbook of cognitive neurology.* Plenum, New York.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* 101(3): B31 – B41.

Piattelli-Palmarini, Massimo (ed.). 1980. *Language and learning : the debate between jean piaget and noam chomsky.* Harvard University Press Cambridge, Mass.

Pinker, Steven. 2004. Clarifying the logical problem of language acquisition. *Journal of Child Language* 31(4): 949–953.

Post, Emil L. 1936. Finite combinatory processes – formulation 1. *Journal of Symbolic Logic* 1(3): 103–105.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt, and Michael Becker. 2010. Harmonic grammar with linear programming: from linear systems to linguistic typology. *Phonology* 27: 77–117.

Prince, Alan and Bruce Tesar. 2004. Learning phonotactic distributions. *Constraints in phonological acquisition* 245–291.

Pylyshyn, Zenon Walter. 1984. *Computation and cognition.* MIT press Cambridge, MA.

Rahimi, Ali and Ben Recht. 2017. Reflections on random kitchen sinks. Accessed from `http://www.argmin.net/2017/12/05/kitchen-sinks`.

Rasin, Ezer and Roni Katzir. 2016. On evaluation metrics in optimality theory. *Linguistic Inquiry* 47(2): 235–282.

Rumelhart, D. E. and J. L. McClelland. 1986. On learning the past tenses of English verbs. In J.L. McClelland and D. E. Rumelhart (eds.), *Parallel distributed processing, volume 2*, 216–271. Cambridge MA: MIT Press.

Sapir, Edward. 1933. La rèalitè psychologique des phonémes. *Journal de psychologie normale et pathologique* 247-65. Reprinted in translation as 'The psychological reality of phonemes' in David G. Mandelbaum (ed.), *Edward Sapir: Selected writings in language, culture, and personality, 46–60. Berkeley: University of California Press, 1949.*

Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8: 333–343.

Simpson, Marc. 2010. *From alternations to ordered rules: A system for learning derivational phonology.* Master's thesis, Concordia University.

Somers, James. 2013. The man who would teach machines to think. *The Atlantic* 11.

Steedman, Mark. 2008. On becoming a discipline. *Computational Linguistics* 34(1): 137–144.

Tesar, B. 1998*a*. An iterative strategy for language learning. *Lingua* 104: 131–145.

Tesar, Bruce. 1995. *Computational Optimality Theory.* Ph.D. thesis, University of Colorado at Boulder.

Tesar, Bruce. 1998*b*. Error-driven learning in optimality theory via the efficient computation of optimal forms. Ms.

Tesar, Bruce. 2007. Learnability. In Paul de Lacy (ed.), *The cambridge handbook of phonology*, 555–574. Cambridge University Press.

Tesar, Bruce. 2014. *Output-driven phonology.* Cambridge University Press.

Tesar, Bruce and Paul Smolensky. 1998. Learnability in optimality theory. *Linguistic Inquiry* (29): 229–268.

Tesar, Bruce and Paul Smolensky. 2000. *Learnability in Optimality Theory.* MIT Press.

Tessier, Anne-Michelle. 2009. Frequency of violation and constraint-based phonological learning. *Lingua* 119(1): 6–38.

Thue, Axel. 1914. Probleme über veränderungen von zeichenreihen nach gegebenen regeln. *Christiana Videnskabs-Selskabs Skrifter* I. Math.-naturv. Klasse 10.

Tomalin, Marcus. 2006. *Linguistics and the formal sciences: the origins of generative grammar*, vol. 110. Cambridge University Press.

Turing, Alan. 1937. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society* s2(42): 230–265.

Turing, Alan. 1950. Computing machinery and intelligence. *Mind* 59(236): 433–460.

Valiant, L.G. 1984. A theory of the learnable. *Communications of the ACM* 27: 1134–1142.

Vu, Mai Ha, Ashkan Zehfroosh, Kristina Strother-Garcia, Michael Sebok, Herbert G. Tanner, and Jeffrey Heinz. 2018. Statistical relational learning with unconventional string models. *Frontiers in Robotics and AI* In press.

Whitehead, Alfred North and Bertrand Russell. 1912. *Principia mathematica*, vol. 2. University Press.

Zinn, Howard. 2010. *You can't be neutral on a moving train: A personal history of our times.* Beacon Press.