# Formal Language Theory in Morphology and Phonology

2:00 - 5:15pm Saturday, 6th January, 2024
Sutton Place, Lower Lobby Riverside Suite, 3rd Floor
Track LSA Organized Session

---

## Formal Language Theory in Morphology and Phonology

**Type of session**

Symposium

**Organizer**

Jeffrey Heinz
Stony Brook University
Scott Nelson
Stony Brook University
Jon Rawski
San Jose State University

**Length of session**

3 hours

**Who sponsors this session?**

n/a

**Session abstract**

This session brings together researchers interested on the relationship between formal language theory and theoretical morphology and phonology. The mathematical theory of formal languages has been invaluable for understanding natural languages for over 70 years (Chomsky, 1959). Fundamental early results in this vein showed that the class of syntactic phenomena cannot be modeled with finite-state automata, but SPE -style phonological rules can (Johnson, 1972; Kaplan and Kay, 1994). All known morphological and phonological processes, including total reduplication, are describable with one-way or two-way finite-state automata, meaning they belong to a restricted class of functions called regular (Beesley and Karttunen, 2003; Rawski et al., 2023). Moreover, phonological and morphological phenomena have been shown shown to overwhelmingly inhabit more restricted sub-regular classes (Heinz, 2011; Chandlee, 2017).

One aspect in which the modern formal language work in linguistics diverges from earlier use is a focus on logical characterizations of morpho-phonological transformations. Logical characterizations provide a bridge between computational and theoretical linguists and have been argued by (Chandlee and Jardine, 2021) and Heinz (forthcoming) to be a general formalism that can be used for phonological analysis. In the first part of this organized session, there will be a tutorial on how to do a phonological analysis using these logical formalisms. The goal of this tutorial is to provide researchers unfamiliar with these techniques an entry point so that they can add these types of analyses to their future work.

The second part of the organized session will include short talks from early career researchers studying the connections between formal language theory and linguistic theory from a variety of perspectives. Yolyan's talk examines the boundary between deterministic and non-deterministic processes with Boolean Monadic Recursive Schemes, a topic relevant to claims about the expressive power of phonological grammars and vowel harmony. Mayer's talk demonstrates the similarity between categorical and gradient grammars using semirings, arguing in favor of the latter. Lamont analyzes that OT grammars with correspondence theoretic faithfulness constraints and typical markedness constraints can be reduced to the Post Correspondence problem, which is known to be undecidable. Wang presents computational and experimental work showing how reduplicative morphology can be parsed, relating such work to questions of learnability. Markowska shows that the amount and kind of data to learn phonological processes by certain learning algorithms is significantly reduced when they operate over phonological features as opposed to segmental units like phonemes. Belth also investigates the learning problem, but from a bottom-up approach. His work develops and empirically tests learning algorithms and then relates them to the kinds of formal classes that Markowska and others begin with.

In the same way that experimental and statistical methodologies have been accepted as useful, normative practices for morphological and phonological theorizing, we hope that this session demonstrates that formal language theory and mathematical logic are similarly well suited, normative practices for these subjects.

**Involved individuals**

Jeffrey Heinz
Stony Brook University
Role Other
Time 0:00
Jane Chandlee
Haverford College
Role Other
Time 0:05
Adam Jardine
Rutgers University
Role Other
Time 0:05
Scott Nelson
Stony Brook University
Role Other
Time 0:05
Tatevik Yolyan
Rutgers University
Role Paper presenter
Time 0:50
Connor Mayer
UC Irvine
Role Paper presenter
Time 1:05
Andrew Lamont
University College London
Role Paper presenter
Time 1:35
Yang Wang
UC Los Angeles
Role Paper presenter
Time 1:50
Magdalena Markowska
Stony Brook University
Role Paper presenter
Time 2:20
Caleb Belth
University of Utah
Role Paper presenter
Time 2:35

**Linguistic field**

Computational Linguistics

**Presentation abstract 1 - Full title**

Tutorial on Morpho-phonological Analysis with Logic and Model Theory

**Presentation abstract 1**

The tutorial will be forty minutes in total, split into three sections: 1) defining phonological representations in model theory; 2) defining phonological processes with first-order translations; and 3) extending first-order logic with boolean monadic recursive schemes, which can define iterative processes and capture elsewhere condition-like effects. Participants will be given step by step instructions and upon completion will be able to provide a logical analysis on any data set of their choosing. The tutorial will also focus on bigger picture questions related to computational complexity such as choice of representation and type of logic used for the analysis.

**Presentation abstract 2 - Full title**

Weak Determinism and Simultaneous Application via Boolean Monadic Recursive Schemes

**Presentation abstract 2**

Weakly deterministic functions are a subregular class of functions that are hypothesized to describe the expressivity of natural language phonology. While there exists an informal and empirically-motivated notion of what constitutes a weakly deterministic pattern, there does not exist a consensus among phonologists on how to formalize the boundary between weakly deterministic and properly regular functions. This talk presents weakly deterministic functions through the framework of Boolean Monadic Recursive Schemes (BMRS), which provides a logical description of string functions. Within this framework, I formally define a 'simultaneous application' operator over two string functions, show that it can be used to model the computational nature of weakly deterministic functions, and explore the theoretical implications of using simultaneous application as a formal characterization of weakly deterministic maps.

**Presentation abstract 3 - Full title**

One (semi)ring to rule them all: Reconciling categorical and gradient models of phonotactics

**Presentation abstract 3**

Several frameworks have been developed over the past decades to model variability observed in phonotactic paterns (e.g. Hayes & Wilson 2008, Boersma & Pater 2016). Rather than assuming a categorical distinction between grammatical and ungrammatical forms, these frameworks assign numeric values that reflect gradient wellformedness. Recent work has argued that the gradience we observe is compatible with categorical grammars (Gorman 2013, Durvasula 2020, Kostyszyn & Heinz 2022). I present evidence from a range of domains that phonotactic grammars must be gradient to engage with realistic data and account for several basic empirical phenomena. I also argue that the apparent successes of categorical grammars in fact implicitly assume grammatical gradience. I will close by showing several examples of how the powerful concept of semirings (Goodman 1999) allows us to study the structure of the grammar independently from the values it assigns to wordforms.

**Presentation abstract 4 - Full title**

Optimality Theory is not computable

**Presentation abstract 4**

From a computational perspective, phonological mappings are not very complex. Insofar as non-cyclic rule-based models are empirically adequate, all attested mappings can be modeled with finite-state machines. Optimality Theory (OT) is a constraint-based framework most commonly used to model phonological mappings. While some variants of OT can be computed by finite-state machines, most, including those standardly employed by practicing phonologists, are not. This talk demonstrates that OT is not computable in general. In other words, it is impossible to construct an algorithm that determines the output of an arbitrary OT grammar and an arbitrary input. The result only uses mechanisms derived from the phonological literature, grounding its implications for practicing phonologists.

**Presentation abstract 5 - Full title**

What does formal language theory tell us about the nature of reduplication?

**Presentation abstract 5**

Reduplicative patterns, in a variety of subtypes, are widely attested. What they have in common is that they create identity between substrings. The fact that these identity requirements are non-regular raises two questions: what is an appropriately restrictive formal model that allows for these identity requirements, and how might learning take place in this setting? To address these questions, I will introduce a class of languages that aims to extend the regular class just far enough to accommodate attested reduplication patterns. This formal framework identifies computationally significant dimensions of variation in the typology of natural language reduplication patterns. I will then present learning experiments whose results bear on these identified dimensions of interest, and consider possible connections between these findings

and formal learnability questions. Using reduplication as a case study, I hope to demonstrate how formal language theory supplements linguistic theory with computational perspectives and motivates novel empirical research.

**Presentation abstract 6 - Full title**

Empirical and theoretical arguments for using phonological features for the learning of sequential functions

**Presentation abstract 6**

Many morphophonological processes can be modeled with sequential functions (Heinz and Lai, 2013), which in turn can be represented with deterministic finite state transducers (Sakarovitch, 2009). Jardine et al. (2014) propose an algorithm (SOSFIA) that can learn such functions in linear time given a finite characteristic sample, and an output empty transducer. The size of a transducer representing those functions depends on the size of the alphabet and the complexity of the process. As a result, the size of the characteristic sample required for SOSFIA will grow exponentially with larger alphabets and more complex processes. We show that the characteristic sample remains constant regardless of the size of the alphabet if the representation of the sample is changed from segments (phonemes) to features defining those phonemes. Consequently, this shows that generalizing over features requires significantly less data than generalizing over segments.

**Presentation abstract 7 - Full title**

The Interaction Between Learning Algorithms and Formal Language Theory

**Presentation abstract 7**

The formal-language-theoretic approach to phonology places phonological generalizations at the starting point of inquiry, by evaluating their computational properties and categorizing them in terms of these. This analysis puts immediate constraints on the properties of any algorithm that might construct such phonological generalizations. All learners have an internal structure that restricts what they can effectively learn, thus rendering some structures beyond the learner's scope, but—by the same token—making effective learning of relevant structures possible. In this talk, I will argue that a complementary way to study the nature of phonology is to directly investigate the computational procedures by which a phonological system is constructed in the mind. This places computational learning procedures at the starting point of inquiry, and works from these toward the phonological generalizations and representations that they construct. I will present experimental and computational results demonstrating how this approach complements the formal-language-theoretic approach.