Long-distance phonotactics:
Integrating experimental and computational approaches to language learnability

Research questions

How do humans learn long-distance dependencies in the sound system of a language? Can we replicate this process by enhancing existing computational models with a richer linguistic structure?

Context

An important property of any language's sound system is its PHONOTACTICS – the unique way in which it allows its inventory of speech sounds to combine into words. In English, "plick" and "snick" would be suitable words, but "tlick" and "bnick" would not, since the English grammar prohibits word-initial sequences of [tl] and [bn]. Likewise, in the Samala language of Southern California, [s] and [ʃ] (the "sh" sound of English) cannot co-occur anywhere within the same word. Evidence of this can be seen when the prefix *s-* of ha-s̲-xintila 'his gentile name' changes to *ʃ-* when the suffix *–waʃ* is added, becoming ha-ʃ-xintila-waʃ 'his former gentile name' (Applegate 1972; Hansson 2010a). This pattern, known as sibilant harmony, is relatively common across the world's languages and illustrates that restrictions on the co-occurrence of speech sounds within a word can apply at varying levels of LOCALITY, sometimes at an unbounded distance. These LONG-DISTANCE DEPENDENCIES have long been a prominent topic in theoretical phonology, and are the focus of my research program.

Long-distance phonotactics are known to cause serious problems for theories of LEARNABILITY (Heinz 2010; Heinz *et al.* 2011). A learner needs an enormous amount of computational power to discover a dependency over an arbitrary distance (within an unbounded search space, moreover), making them unlearnable in practice. However, the existence of these patterns in natural languages such as Samala is evidence that they are indeed learnable. To resolve this paradox, researchers posit a set of cognitive LEARNING BIASES that facilitate the learning of patterns with certain properties but not others. Recent experimental studies have shown, for example, that non-adjacent dependencies are easier to learn if the interacting elements are relatively similar (Moreton 2008, 2012; Newport & Aslin 2004) and if they are closer together (Finley 2012).

The proposed project uses a multi-disciplinary approach to study the learning biases associated with long-distance dependencies. The outcomes will further our understanding of how the patterns of natural languages are restricted, highlighting a fundamental aspect of human cognition shared by everyone.

Objectives

During the award tenure, I will answer the above research questions by achieving two main objectives:

1. Expand our knowledge of the biases that influence how humans learn long-distance dependencies.
2. Develop a computational learning algorithm that can acquire the patterns found in natural languages, imitating the performance of humans from experimental studies.

Methodology

The *experimental* portion of this project will use an ARTIFICIAL LANGUAGE LEARNING paradigm to test for possible learning biases that cannot otherwise be observed. This has become a popular methodology for linguists and cognitive psychologists interested in language learning (Finley & Badecker 2009; Nevins 2010; for a review see Moreton & Pater 2012a,b), as the relative rarity (or non-existence) of certain patterns makes it unfeasible to use the more traditional method of studying children throughout the process of language acquisition.

An example of a typical study comes from McMullin & Hansson (2013). Adult participants first complete a training phase in which they hear and repeat words from a language, constructed for the study, that prohibits l...r and r...l sequences, but allows l...l and r...r (e.g. pel̲okir̲u would never appear in the language, but per̲okir̲u would be an acceptable form) – an attested pattern known as liquid harmony. After being exposed to about 600 training items (~25 minutes) that conform to the pattern, all participants are tested to determine whether they prefer liquid harmony in novel words. The results reveal whether they have learned the pattern, and more importantly if they generalize the pattern to similar contexts that were not encountered in their training.

I will run three experiments focusing especially on the learning biases related to locality, or the distance between the two elements in a dependency. Experiment 1 will restrict the participants' input to one level of locality (separated by one, two, or three syllables), but will test whether they apply the pattern to the previously unseen distances. Experiment 2 will give training that exhibits a dependency at multiple distances, testing whether a varied input can help the learner discover the pattern. Experiment 3 will give evidence for the pattern at one distance, but evidence against the pattern at another distance. The results of this study will be particularly interesting, since we do not know whether some of these patterns are learnable, as they are unattested in natural languages (e.g. a dependency that holds across two syllables, but not in adjacent syllables; Hansson 2010a).

The goal of the *computational* component of this project is to develop a FORMAL LEARNING MODEL that learns the same set of long-distance dependencies as humans. The field of linguistics has a long-established set of criteria to define a pattern as learnable by a particular algorithm (i.e. identifiable in the limit from positive data; Chomsky 1956; Gold 1967). As a result, the literature does provide a few basic models that show potential for application to long-distance interactions (Hayes & Wilson 2008; Heinz 2010), but that have some unaddressed problems.

I will build off of an existing model known as a precedence learner (Heinz 2007; 2010). In principle, the model is quite simple. For the word 'dogs', the learner would record the adjacent bigrams {do}{og} and {gs} as well as the precedence relations {d-o}{d-g}{d-s}{o-g}{o-s} and {g-s}, with no reference to distance or intervening material, just that one sound precedes the other. These two properties provably allow it to learn unbounded long-distance dependencies (Heinz 2010). However, this model is too general, being able to learn a dependency between any two elements, not just between two that are relatively similar – a known human learning bias. Furthermore, it cannot learn an attested, though more restrictive type of liquid harmony in which words can contain sequences of rVCVl and lVCVr, but not rVl and lVr (where *V* is any vowel and *C* is any consonant).

I propose that a single modification of the model resolves these issues: allowing it to operate over strings of sounds that can be defined as a phonological class (e.g. vowels, consonants, sibilants, or liquids), rather than across all segments in the word. Phonological theory relies heavily on the categorization of linguistic sounds into natural classes based on a shared feature, but this structure has never been integrated into a precedence model of learning.

Contribution to the advancement of knowledge

Though long-distance dependencies are quite simple descriptively, they are puzzling from many perspectives. Computationally they are quite complex, classifying a set of patterns that are likely at the boundaries of human learning capabilities. They also cause problems for theories of language change. The existence of a sound pattern is often explained as an error in transmission from one generation to the next, arising from systematic misperceptions or misproductions, but this study tests the hypothesis that learning biases can also help to shape the patterns found in the world's languages. Though my undertaking of this project is the result of an interest in theoretical phonology, it connects such diverse disciplines as computational learning theory, cognitive psychology, and historical linguistics.

Additional applicant and project information

I am currently a Ph.D. student in UBC's Department of Linguistics. Having transferred from the M.A. level in 2011, I am expected to graduate in 2016. My committee includes my supervisor Dr. Gunnar Hansson, an expert on long-distance phenomena in phonology (Hansson 2001; 2010a,b), as well as Dr. Carla Hudson Kam, the Canada Research Chair in Language Acquisition, who has conducted several artificial language learning studies in the past (Hudson Kam & Newport 2005; 2009). She is also the director of UBC's Language and Learning Lab, which provides me with the necessary facilities for completing this project. I have already run several artificial language studies that will shape the initial chapters of my thesis, establishing the relationship between learning biases and linguistic typology. By accomplishing the above objectives, I will round out my dissertation by using further experimental results to inform computational models of learning – a much-needed addition to the literature.