1 Abstract

Much of the recent progress in natural language processing (NLP) has been driven by ever larger data sets and ever more complex algorithms that take advantage of them. This approach has proven extremely successful on languages like English, French, or Modern Standard Arabic (MSA) for which such large data sets are available, but it has left other languages, and other communities, behind. Most of the world's roughly 7,000 languages are "low-resource" from the perspective of NLP: they have relatively few written resources, lack standardized orthographies, and coexist in multilingual situations where better-resourced languages are generally used in more formal settings.

Though we generally refer to it as a single language, Arabic is actually a family of related albeit distinct varieties which are often not mutually intelligible. Although these varieties collectively have roughly as many native speakers as English, only MSA has the standardized orthography and sufficient corpus data suitable for large-scale NLP. The Arabic varieties are a particularly valuable area for developing computational learning systems since they are characterized by linguistically complex systems of *templatic* morphology. In Arabic, many grammatical functions are expressed by means of dramatic internal changes rather than just by the addition of suffixes; thus where English forms the plural of *book* with the addition of the *-s* suffix (*books*), in MSA, the plural of *kitaab* 'book' is *kutub*. How and why Arabic has this system poses an interesting theoretical question and a challenging problem for systems designed to automatically learn the language.

The goals of this project are two-fold: to build NLP tools for Arabic morphology by incorporating methods from linguistics while furthering our understanding of the cognitive science of Arabic word structure using the low-resource learning tasks as an empirical test bed. We propose to lay the groundwork for this research program over the next year by developing new tools and data sets for Arabic varieties and by incorporating new work in Arabic theoretical phonology and morphology into acquisition models.

We propose to synthesize three disparate approaches to language science and engineering: **1)** theoretical linguistics, which models the internalized knowledge that allows speakers to produce and understand utterances, including utterances they have never heard before; **2)** the study of language acquisition, which explores how young children acquire this complex system with exposure to limited and incomplete data; **3)** and NLP, which builds computational systems that learn to approximate the capabilities of humans in recovering meaning from language. Children are the gold-standard for low-resource language learning. They acquire their native languages on the basis of surprisingly little evidence. Most of what is possible to say in a given language is never actually said: children receive only a tiny fraction of the language examples that would facilitate learning for an adult or a conventional NLP system. How children manage to acquire their native languages so well despite the sparsity of their input is a central question for theoretical linguistics – children simply could not succeed without some well-defined representational constraints on their learning. Working out what those constraints are is a major goal of theoretical linguistics, and building computational acquisition models that take advantage of these constraints will shed further light on them. At the same time, acquisition can be thought of as a particularly severe low-resource learning problem. In building an NLP model for a given low-resource Arabic variety, we can pool resources from other varieties and learn to map between them, something the child cannot do, but also extend what we uncover about acquisition and theory to the problem. The same theoretical insights and computational methods that facilitate better acquisition models and low-resource learning models in NLP.

In the effort funded by the seed grant, we propose to validate the intuition shared by the co-PIs: that syllable structure plays a fundamental, but to date under-explored, role in modeling Arabic morphology in theory and in acquisition, aand that understanding of cross-dialectal variation in syllable structure can greatly increase the quality of learned NLP models for dialectal Arabic processing. We will develop baseline models integrating underlying syllabic representations and determine what role these models can play in human acquisition and in machine learning for NLP. These will become a central and entirely novel element in the NSF proposal we will prepare.