Syntactic locality: an interface of typology, theory and computation

Thomas McFadden* (PI)
John Bailyn*, Thomas Graf*,†, Sandhya Sundaresan* (Co-PIs)
*Department of Linguistics † Institute for Advanced Computational Science

1 Overview/abstract

Dependencies between elements in natural language syntax — i.e. the structures of sentences — are subject to locality constraints. That is, two words or other syntactic items are able interact in some grammatical way only if they are close enough to each other in some relevant sense. For example, an English reflexive pronoun like *myself* has to appear (roughly) in the same clause as a pronoun or noun phrase that it co-refers with, so it's possible to say *I considered myself*, but not *I think that she considered myself*. Locality has played a central role in theoretical work on syntax for decades, and empirical work over that time has uncovered voluminous evidence for locality being at play in diverse phenomena across the languages of the world. The result, however, is a wide array of competing characterizations and theoretical approaches to locality — including quite disparate conceptions of what locality actually is — with distinct but overlapping patterns of empirical coverage that make principled comparison extremely difficult.

The project proposed here aims to address this situation by building three resources: 1) A representative typology of locality phenomena 2) A curated survey of theoretical approaches to locality 3) Targeted insights from mathematical and computational linguistics, which can provide ways to adjudicate among the competing theoreties. The resources will then be brought together to build new theoretical and computational tools for work on the topic. In so doing, the project would take advantage of crucial strengths of the Department of Linguistics at Stony Brook that distinguish it from most other centers for linguistics: the presence of world-class researchers both in theoretical syntax and in computatinal linguistics, who are uniquely well-suited to a collaboration of this type.

The current proposal seeks OVPR seed grant funding for the preliminary steps in this process, centered around building initial versions of the three components described above. These will provide the basis for the project to develop proposals for two NSF grants. The first, to be submitted to the Linguistics Program (https://new.nsf.gov/funding/opportunities/linguistics), will be to fund a larger project to construct a novel theoretical approach to syntactic locality that synthesizes the insights from typology, theory and computation that come out of the three components. The second, also to be submitted to the Linguistics Program, but also targetting the Computational Cognition (CompCog) initiative (https://new.nsf.gov/funding/opportunities/stimulating-integrative-research-computational-1), will build specifically on the typological component to create a corpus of linguistic material exemplifying the range of locality effects, which can be used as a test battery to evaluate Large Language Models (LLMs) and other natural language systems on their ability to learn the relevant effects. The seed grant will primarily go to funding a graduate RA who will play central roles in carrying out the research and in coordinating among the team members and the different portions of the project. It will also fund an intensive workshop with invited external experts who will provide input from their own areas of specific expertise relating to locality, as well as critical feedback on intermediate results of the project, in preparation for writing the NSF grant proposals.

A note on terminology: we use the term 'theoretical syntax' throughout as a shorthand for referring to theoretical approaches that are not explicitly computational or mathematical in their goals and methods. This is not meant to imply that computational and mathematical approaches to natural language are any less theoretical than the non-computational approaches — they certainly are not — but is used as a matter of convenience reflecting common usage in the field and the lack of a satisfactory alternative.