Matthijs Vákár Internship Report

Name: Matthijs Vákár

Current status: currently spending a gap year to find a good PhD-project

Where I worked: Worked in Department of Linguistics, University of Cambridge

Supervisor: Worked with Jeffrey Watumull, a student of Prof. Ian Roberts and Prof. Noam Chomsky

Summary of work: There are sciences of which the mathematical nature is well-known. Take physics or computer science for instance or, by extension, even any natural science. Linguistics is not generally the first to spring to mind. In fact, it is usually even grouped with the humanities. This is understandable considering the subfields of pragmatics, stylistics, psycholinguistics and historical linguistics. Much of the work done in linguistics however has a very mathematical flavour: phonetics, morphology, syntax, semantics, discourse analysis and computational linguistics. In fact, many of the best known linguists in these fields, including Chomsky, actually started out as mathematicians. It might therefore be less surprising that I was received with such a warm welcome by all linguists I talked to, despite my initial lack of knowledge of linguistics. This summer, I submerged myself in natural language syntax, the field of linguistics that studies grammars, reading books and papers and discussing what I was reading with my supervisor.

Given an alphabet (which we take to be finite), one can define the set of sentences (finite strings) over that alphabet and its powerset, the set of all languages over that alphabet (cardinality of continuum). Arguing that natural languages are learnable by children based on finite positive evidence, one sees that they must be in some sense finitely generated (by a grammar), more precisely that they are computable. Since the set of computable languages has the cardinality of the natural numbers, we conclude that most languages cannot be natural languages.

Based on the languages we observe `in the wild' however, it seems that this classification of natural languages as computable is still too coarse. There are many different very successful frameworks for grammars in which linguists have fitted a couple of hundred of the best studied of the circa 6000 languages known to exist to this date. These frameworks appear very different and have often been developed independently. However, they turn out to define roughly the same subclass of the computable languages. To be more precise, they all define one of 4 or 5 subclasses of the computable languages. (c.f. Chomsky Hierarchy) This is a welcome unification in the disparate landscape of theories of syntax. It is now the job of empirical research to point out which of these classes and therefore which frameworks for grammar fit(s) natural language best. Slowly but surely, one hopes to converge to an appropriate formal notion of natural language this way. Ideally, this would not only give us a framework in which we can fit all the world's languages, but more importantly, it would teach us something about one of the most fundamental aspects of human existence, about the very nature of language.

The project has been very worthwhile. The discussions I had with my supervisor never failed to be

intellectually stimulating. I do not doubt that we will stay in touch. Moreover, the subject matter turned out to be a very nice mix between serious formal mathematics and real world applications. This leaves me so enthusiastic that I am currently seriously considering to continue in the direction of mathematical linguistics.

Recommended reading:

What is linguistics?

- - http://grammar.about.com/od/grammarfaq/a/What-Is-Linguistics.htm

Relevant papers for my project:

- - Chomsky – Three Models for the Description of Language (1956)
- - Watumull – A Turing Program for Linguistic Theory (2012)
- - Nowak, Kamarova, Niyogi – Computational and Evolutionary Aspects of Language (2002)
- - Stabler – Recursion in Grammar and Performance (2011)