Analyzing Pre-Indo-European Theory of the Etruscan Language Origins Using Topological Data Analysis

Welch, Helena (School: Los Alamos High School)

High-dimensional data is often difficult to analyze because of the exponential growth of the size of the space in which the data lives as the dimension increases. One example of high-dimensional data comes from language, which contains many different characteristics (dimensions) with which it can be quantified, but not always sufficient data to detect patterns in it. This is especially true for ancient languages, as there is a sparsity of texts from which to draw. The ancient Etruscan language is currently classified as a non-Indo-European isolate. However, the Etruscans lived in an Indo-European-speaking region and appear to be genetically related to Indo-Europeans. This study aims to bring a quantitative measure, topological data analysis (TDA), to ongoing investigations of Etruscan to more concretely determine Etruscan's similarity to different Indo-European languages. Phonetic patterns in a specific word list translated into different languages by large-language models are encoded, and the distance between two given phonemes based on this encoding is calculated. Results indicate that Sanskrit has the highest correlation to Etruscan. Etruscan appears similar to older Indo-European languages and thus may be older than neighboring languages, explaining its uniqueness compared to Indo-European languages that developed later in time.

Awards Won: