Researchers reconstruct major branches in the tree of language

September 10 2021

Representative image of a branching tree. Credit: Kevin Wenning/Unsplash.com

The diversity of human languages can be likened to branches on a tree. If you're reading this in English, you're on a branch that traces back to a common ancestor with Scots, which traces back to a more distant

ancestor that split off into German and Dutch. Moving further in, there's the European branch that gave rise to Germanic; Celtic; Albanian; the Slavic languages; the Romance languages like Italian and Spanish; Armenian; Baltic; and Hellenic Greek. Before this branch, and some 5,000 years into human history, there's Indo-European—a major proto-language that split into the European branch on one side, and on the other, the Indo-Iranian ancestor of modern Persian, Nepali, Bengali, Hindi, and many more.

One of the defining goals of historical linguistics is to map the ancestry of modern languages as far back as it will go—perhaps, some linguists hope, to a single common ancestor that would constitute the trunk of the metaphorical tree. But while many thrilling connections have been suggested based on systemic comparisons of data from most of the world's languages, much of the work, which goes back as early as the 1800s, has been prone to error. Linguists are still debating over the internal structure of such well-established families as Indo-European, and over the very existence of chronologically deeper and larger families.

To test which branches hold up under the weight of scrutiny, a team of researchers associated with the Evolution of Human Languages program is using a novel technique to comb through the data and to reconstruct major branches in the linguistic tree. In two recent papers, they examine the ~5,000-year-old Indo-European family, which has been well studied, and a more tenuous, older branch known as the Altaic macrofamily, which is thought to connect the linguistic ancestors of such distant languages as Turkish, Mongolian, Korean, and Japanese.

"The deeper you want to go back in time, the less you can rely on classic methods of language comparison to find meaningful correlates," says co-author George Starostin, an Santa Fe Institute external professor based at the Higher School of Economics in Moscow. He explains that one of the

major challenges when comparing across languages is distinguishing between words that have similar sounds and meanings because they might descend from a common ancestor, from those that are similar because their cultures borrowed terms from each other in the more recent past.

"We have to get to the deepest layer of language to identify its ancestry because the outer layers, they are contaminated. They get easily corrupted by replacements and borrowings," he says.

To tap into the core layers of language, Starostin's team starts with an established list of core, universal concepts from the human experience. It includes meanings like "rock," "fire," "cloud," "two," "hand," and "human," amongst 110 total concepts. Working from this list, the researchers then use classic methods of linguistic reconstruction to come up with a number of word shapes which they then match with specific meanings from the list. The approach, dubbed "onomasiological reconstruction," notably differs from traditional approaches to comparative linguistics because it focuses on finding which words were used to express a given meaning in the proto-language, rather than on reconstructing phonetic shapes of those words and associating them with a vague cloud of meanings.

Their latest re-classification of the Indo-European family, which applies the onomasiological principle and was published in the journal *Linguistics*, confirmed well-documented genealogies in the literature. Similar research on the Eurasian Altaic language group, whose proto-language dates back an estimated 8,000 years, confirmed a positive signal of a relationship between most major branches of Altaic—Turkic, Mongolic, Tungusic, and Japanese. However, it failed to reproduce a previously published relationship between Korean and the other languages in the Altaic grouping. This could either mean that the new criteria were too strict or (less likely) that previous groupings were

incorrect.

As the researchers test and reconstruct the branches of human language, one of the ultimate goals is to understand the evolutionary paths languages follow over generations, much like evolutionary biologists do for living organisms.

"One great thing about historical reconstruction of languages is that it's able to bring out a lot of cultural information," Starostin says. "Reconstructing its internal phylogeny, like we're doing in these studies, is the initial step to a much larger procedure of trying to reconstruct a large part of the lexical stock of that language, including its cultural lexicon."