How many languages are on the web? The Crúbadán project ten years on

Kevin Scannell

Around 2003 we began collecting text corpora from the web for about 140 indigenous and minority languages to assist communities in the development of basic language technologies like spelling and grammar checkers. Since then, we've expanded the number of languages to nearly 1500, and the corpora we've created have been used by hundreds of linguists, software developers, and community members in a wide range of projects. We will discuss our progress in trying to answer the (problematic) question "How many languages are on the web?", focusing on the many challenges that arise when scaling a project to thousands of languages.