Motivation
There was no multilingual corpus containing many languages with enough data for various NLP applications. We decided to build a web corpus containing at least 100 languages with a minimum of 10MB of text for each language. This language data resource can be of use particularly to researchers specialized in development of multilingual technologies.

Building a Web Corpus
Can be divided into several steps as follows:
- retrieving metadata
- building an initial corpus
- generating word tuples
- using them as search queries
- downloading found pages
- removing boilerplate code
- identifying language
- removing duplicate content
- evaluating text quality

Metadata
Combination of multiple sources: SIL International, Wikipedia, and Ethnologue. These metadata are now accessible through RESTful API.

Language Identification
YALL, our language identifier, uses a scoring function based on byte 4-grams.

Duplicity Reduction
Duplicity sources: spam, common passages and incorrectly detected boilerplate code. Only unique paragraphs are preserved.

W2C Web Corpus

World Coverage
States covered by languages included in W2C Corpus according to Ethnologue