Readme file:
Ancestral Characteristics Database, language level data

The language-level Ancestral Characteristics Database contains a shapefile called "langa_no_overlap_biggest_clean". This is a cleaned version of the Ethnologue version 16 shapefile. We have removed any sliver polygons and have removed overlapping polygon so that each point on earth is assigned to only one language or dialect. When doing this, we chose to keep the largest polygon.

The database also contains three stata dta files, which contain ethnographic data for each language group:

EthnoAtlas_Ethnologue16_baseline_by_language.dta
EthnoAtlas_Ethnologue16_extended_EE_Siberia_by_language.dta
EthnoAtlas_Ethnologue16_extended_EE_Siberia_WES_by_language.dta

The differences between and sources of the underlying ethnographic data are described in Guiliano and Nunn (2017). The files contain identifiers that can be used to match / join the ethnographic data to the shapefile. The variable id is the Ethnologue language identifier. V107 is the identifier/name of the ethnic group from the original ethnographic data (e.g., *Ethnographic Atlas*).

The dataset also contains a variable called "poor_match," which indicates the quality of the match between the Ethnologue language/dialect variable and the ethnic group in the ethnographic data. A value of 0 indicates that a direct match was possible. A value of 1 indicates that a direct match wasn't possible because the language was not represented in the ethnographic data. A value of 2 indicates that a direct match wasn't possible and this is because the language is a mixed language e.g., pidgin or creole. For both 1 and 2 matches, geographic proximity was used to match language to an ethnic group.