This dataset has been collected by Dula Bankira for his research work to translate English language to *Ho* language for understanding and conveying information to Ho people. Ho is a tribe in India with a distinct language (the *Ho* language), culture, mythology, tradition, writing system, etc. Orally, the *Ho* people communicate in their language, passing it from generation to generation. In the current technological age, it is important to disseminate knowledge and exchange information in the *Ho* language to keep the language alive, and machine translation could become a part of the solution. In my research work, I have collected dataset of 12890 sentence pairs and 4418 word pairs and manually typed and translated it in English to *Ho* language using *Warang Chiti* script.

For the word-to-word translation from English to *Ho*, the bilingual English-*Ho* dictionary was developed manually by Dula Bankira. This bilingual dictionary contains 4418 English correctly labelled with their parallel *Ho* words. This parallel corpus consists of 12890 parallel sentences consisting of 28108 English and 25528 *Ho* tokens as in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **Sentences**  **(Parallel)** | **Tokens** | |
| **English** | ***Ho*** |
| An *Ho*-Hindi Dictionary | -- | 4418 | 3525 |
| English Tense Chart | 1850 | 7400 | 5558 |
| Rapidex English Speaking Courses | 2885 | 14425 | 11543 |
| *Ho* Grammar (with vocabulary) | 805 | 1678 | 1522 |
| *Ho* Dictionary | 325 | 1307 | 1225 |
| Manually translation | 7025 | 28108 | 25528 |
| Total | 12890 | 57384 | 48902 |

**Table 1**: English-*Ho* parallel corpus details after raw data collection

This dataset can be used publicly for the purpose of research work.