Corpus name	Words	Description
Brown Corpus	1M	The first modern, computer readable corpus. Consists of various texts in American English.
BNC	100M	A large corpus of spoken and written British English, completely annotated with part-of-speech tags.
Penn Treebank	1M	Wall Street Journal sentences annotated with parse trees.
CHILDES	various	A collection of child language data corpora.
Switchboard	3M	Transcribed telephone conversations and spoken texts. Includes recordings of the sentences used in the Penn Treebank.
HCRC Map Task	145K	Audio, video, and transcriptions of spoken dialogue between individuals participating in a cooperative task.
Canadian Hansard	20M	Bilingual sentence-aligned French and English proceedings of the Canadian parliament.