## [LibriSpeech](http://www.openslr.org/12/): This corpus contains roughly 1,000 hours of English speech, composed of audiobooks read by multiple speakers. The data is organized by chapters of each book.

* [Spoken Wikipedia Corpora](https://nats.gitlab.io/swc/): Containing hundreds of hours of audio, this corpus is composed of spoken articles from Wikipedia in English, German, and Dutch. Due to the nature of the project, it also contains a diverse set of readers and topics.
* [Amazon Reviews](https://snap.stanford.edu/data/web-Amazon.html): This dataset contains around 35 million reviews from Amazon spanning a period of 18 years. It includes product and user information, ratings, and the plaintext review.
* [Google Books Ngrams](https://aws.amazon.com/datasets/google-books-ngrams/): A Google Books corpora of n-grams, or ‘fixed size tuples of items’, can be found at this link. The ‘n’ in ‘n-grams’ specifies the number of words or characters in that specific tuple.
* [UCI’s Spambase](https://archive.ics.uci.edu/ml/datasets/Spambase): Originally created by a team at Hewlett-Packard, this large spam email dataset is useful for developing personalized spam filters.
* [Stanford Sentiment Treebank](http://nlp.stanford.edu/sentiment/code.html): Also built from movie reviews, Stanford’s dataset was designed to train a model to identify sentiment in longer phrases. It contains over 10,000 snippets taken from Rotten Tomatoes.
* [Recommender Systems Datasets](http://cseweb.ucsd.edu/~jmcauley/datasets.html): Datasets from a variety of sources, including fitness tracking, video games, song data, and social media. Labels include star ratings, time stamps, social networks, and images.
* [Project Gutenberg](http://www.gutenberg.org/wiki/Gutenberg:Offline_Catalogs): Extensive collection of book texts. These are public domain and available in a variety of languages, spanning a long period of time.\
* [M-AI Labs Speech Dataset](https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/): Nearly 1000 hours of audio plus transcriptions. It includes multiple languages arranged by male voices, female voices, and a mix of the two.
* [European Parliament Proceedings Parallel Corpus](http://www.statmt.org/europarl/): Sentence pairs from Parliament proceedings. There are entries from 21 European languages including some less common entries for ML corpus.