

THE WEST POINT BRAZILIAN PORTUGUESE SPEECH CORPUS

The Center For Technology Enhanced Language Learning
United States Military Academy
Department Of Foreign Languages
745 Brewerton Road
West Point, NY 10996
Email: john.morgan@usma.edu
Phone: 845-938-6077
Fax: 845-938-3585

March 23, 2004

The West Point Brazilian Portuguese Speech Corpus is a data base of digital recordings of spoken Brazilian Portuguese . It was collected by staff and faculty of the Department of Foreign Languages (DFL) and Center for Technology Enhanced Language Learning (CTELL) to develop acoustic models for speech recognition systems. The US government uses These systems to provide speech-recognition enhanced language learning courseware to government linguists and students enrolled in various government language programs.

The corpus was collected in March of 1999 in Brasilia Brazil. Informants were recruited from the Brazilian Military Academy. The corpus includes read speech from native and non-native speakers.

The speech was elicited from a prompt script. Informants read short sentences and phrases typically used in language learning situations. The scripts contain a total of 296 sentences. The prompts are listed in the file prompts.txt. Each line of this file has two fields separated by a tab, the first denoting the base name of the waveform file, and the second the prompt used in recording the utterance.

A pronouncing dictionary is provided in the file SANTIAGO.txt. The dictionary was developed by Dr. Sheila Ackerlind with help from cadet Sterling Packer.

Speech data was collected using 4 laptop computers running MS Windows. Three of the computers recorded with a 16 bit data size and sampling rate of 22050 Hz, the other laptop recorded with an 8 bit data size at 11025 Hz. The recording script presented a visual display of the sentence to be recorded. The informant pressed a key and spoke the sentence. The recording was played back for review allowing the utterance to be re-recorded. A member of the data collection team was on hand during the recording session to verify recordings and provide technical assistance in case of malfunctioning equipment.