Overview of “The Corpus of Oral Presentations in English (COPE)”

Provider: Michiko Watanabe (NINJAL) and TimeHill Inc.

Project

COPE was built as a project of Center for Corpus Development, NINJAL, in cooperation with Straightword Inc., supported by the following JSPS KAKENHI.

1) Contrastive analysis of English and Japanese disfluencies using parallel spontaneous speech corpora of the two languages

(2012-2014, project number: 24520494, principal researcher: Michiko Watanabe)

2) A contrastive study of speech disfluencies: How does the complexity of the following constituents affect the occurrence of filled pauses?

(2015-2018, project number: 15K02553, principal researcher: Michiko Watanabe, co-investigator: Ralph Rose)

3) A contrastive study of speech disfluencies using parallel spontaneous speech corpora of English and Japanese

(2018-2021, project number: 18K00559, principal researcher: Michiko Watanabe, co-investigator: Ralph Rose)

The corpus publication was approved by Ethics Review Committee of NINJAL (2021).

Corpus overview

COPE was designed for contrastive studies of informal presentation speeches in English and Japanese. The speech topic, speakers’ age, sex, academic backgrounds, and the recording settings are matched to those of a part of “Simulated Public Speaking (SPS)” in the “Core” of “The Corpus of Spontaneous Japanese (CSJ)”. https://ccd.ninjal.ac.jp/csj/en/index.html

The given speech topic was “the most memorable experience in my life.” Twenty speakers were college students or graduates in their 20s and early 30s living in Los Angeles or Anaheim at the time of recording (2012-13). Half of them were male and the other half female speakers. The speakers gave their talks in front of a small audience including their friends in a relaxed atmosphere. Each speech lasts about 10 minutes. The corpus consists of 41,062 words in 3.8 hours of speech in total.

The following annotations were given to the speech transcripts and compiled in XML format. Refer to files in the documents folder for labeling details.

Files in documents folder:

Explanation about file names -> file_name_scheme.txt

Disfluency labels -> disfluency_labeling_English.docx
Clause boundary labels -> clause_boundary_labels_English.docx
Clause boundary labels in XML files -> PennTreebankII_tags.txt
POS labels -> A parser "Enju" was used for POS analysis. Refer to enju_POS_tags.txt for the list of POS tags. https://mynlp.is.s.u-tokyo.ac.jp/enju/
Time stamps are given at word and clause boundaries using forced alignment techniques with HTK (http://htk.eng.cam.ac.uk/). SP and filler boundaries are manually checked, whereas other boundaries are not.
Labels in XML files -> XML_labels_English.docx
publication_with_COPE.docx -> a list of research articles using COPE

Folders
・documents: explanations about labeling
・speech: 20 wav files of presentation speeches (16kHz・16bit・Mono).
・XML: 20 XML files.
・text_annotated: transcripts with disfluency and clause boundary labels.
・lab_files: the beginning and the end time of each word, filler and silent pause in each file in the "lab" format of waveSurfer (https://wavesurfer-js.org/).
・text_clean: plain transcripts without disfluencies and any linguistic labels

Purpose of use: for academic research and education only
License fee: free