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

提供者：渡辺美知子（国立国語研究所），株式会社タイムヒル

プロジェクト

本コーパスは，国立国語研究所コーパス開発センタープロジェクトの一環として，株式会社ストレートワード協力のもと，以下の科学研究費の助成を受けて構築された。
1) 日本語話し言葉コーパスにおける言い淀み分類の精緻化と機能の対照分析
（2012-2014 年度，課題番号：24520494，研究代表者:：渡辺美知子）
2) 後続要素の複雑さが言い淀みの発生に及ぼす影響についての日英語対照研究
（2015-2018 年度，課題番号: 15K02553，研究代表者：渡辺美知子）
3) 日本語と英語のパラレルコーパスを用いた言い淀みの対照言語学的研究
（2018-2021 年度, 課題番号: 18K00559, 研究代表者:渡辺美知子, 研究分担者:Ralph Rose）
本コーパス公開に際し，国立国語研究所研究倫理委員会（2021)の承認を受けている。

概要

本コーパスは『日本語話し言葉コーパス（CSJ)』模擬講演の一部との対照研究ができるようにデザインされている（https://ccd.ninjal.ac.jp/csj/)。具体的には，講演のトピック，講演者の年齢・性別・学歴，録音環境が，CSJ コア内の模擬講演の一部に合わせてある。講演トピックは，“the most memorable experience in my life”である。録音時（2012-13）アメリカ，ロサンジェルスまたはアナハイムに住む 20 代～30 代前半のアメリカ英語話者（学生または大学卒業者）男女各 10 名が少人数の知人，友人を前に，約 10 分間のインフォーマルなスピーチを行った。本コーパスはその録音音声と書き起こしテキストを含んでいる。総計 41,062 語，3.81 時間から成る。書き起こしテキストには，非流暢性ラベル，節境界ラベル，POS ラベル，語境界の時間情報が付与されている。ラベリングの詳細については，documents フォルダ内の以下のファイルを参照されたい。これら全ての情報は XML ファイルに格納されている。

documents フォルダ内のファイル

ファイル名の説明 -> file_name_scheme.txt
非流暢性ラベル -> disfluency_labeling_Japanese.docx
節境界ラベル -> clause_boundary_labels_Japanese.docx
XML files 内の節境界ラベル -> PennTreebankII_tags.txt
POS ラベル -> 形態素解析には“Enju”を用いた。品詞タグリストは，enju_POS_tags.txt を参照のこと。https://mynlp.is.s.u-tokyo.ac.jp/enju/
時間情報：各語，ポーズ，節の開始と終了の時間情報が付与されている。これは，HTK (http://htk.eng.cam.ac.uk/)を用いた強制切り出しの結果である。ポーズ境界，フィラー境界のみ，人手での確認・修正がなされている。
XML files 内のラベル -> XML_labels_ Japanese.docx
publication_with_COPE.docx -> COPE を用いた研究業績リスト

フォルダ構成

・documents：各種ラベリング仕様についての文書を格納したフォルダ
・speech：20 講演の音声ファイルフォルダ
　音声ファイルフォーマート：WAV 形式（16kHz・16bit・Mono）
・XML：XML ファイルフォルダ
・text_clean：非流暢性表記も言語情報タグもない書き起こしテキストフォルダ
・text_annotated：非流暢性の記述と言語情報タグのある書き起こしテキストフォルダ
・lab_files：語，フィラー，ポーズの開始・終了時間を waveSurfer の lab file 形式で記載したテキストファイルフォルダ。フィラーとその前後のポーズは人手修正がなされているが，それ以外の箇所は HTK を用いた強制切り出しの出力結果で，人手修正はなされていない。

利用範囲：研究・教育目的に限る
利用価格：無償