日本語日常会話コーパスの Universal Dependencies: UD_Japanese-CEJC

大村 舞 [1]　若狭 絢 [2]　松田 寛 [3]　浅原 正幸 [4]
[1] 大阪樟蔭女子大学　[2] 東北大学
[3] 株式会社リクルート Megagon Labs　[4] 人間文化研究機構国立国語研究所
omura.mai@osaka-shoin.ac.jp　aya.wakasa.c3@tohoku.ac.jp
matsudahiroshi@r.recruit.co.jp　masayu-a@ninjal.ac.jp

掲載号の情報

32 巻 1 号 pp. 55-90.
doi: https://doi.org/10.5715/jnlp.32.55

概要

本研究では，日本語日常会話コーパス (CEJC) を Universal Dependencies 形式に変換した日本語話し言葉のツリーバンク UD Japanese-CEJC を開発・構築したので，そのデータについて報告する．日本語日常会話コーパスは，日本語の様々な日常会話を収録した大規模な音声言語コーパスであり，単語区切りや品詞のアノテーションが含まれている．我々は，UD Japanese-CEJC のために，CEJC の長単位形態論情報と文節係り受け情報を新たにアノテーションした．UD Japanese-CEJC は日本語形態論情報と文節ベースの依存構造情報および CEJC から手作業で整備された変換ルールに従って構築した．構築した UD Japanese-CEJC に対して，日本語書き言葉コーパスとの比較や UD 依存構造解析精度の評価をおこない，CEJC における UD 構築に関する様々な問題点を検討した．