Constructional Diversity Analyzer (CDA)

https://haerimhwang.github.io/tools/English-constructional-diversity-analyzer

■出典

Hwang, H., & Kim, H. (2022). Automatic analysis of constructional diversity as a predictor of EFL students’ writing proficiency. Applied linguistics. (online first)

Hwang, H., Choe, A., & Zenker, F. (2020). Construction Counter: A tool to measure (nonnative) language development. Brown Bag Series, University of Hawai‘i at Mānoa, Honolulu, USA. (April 16)

■ spaCy と Kivy を使用

1.Natural Language Toolkit (Bird et al. 2009) の sent_tokenize で文ごとに区切る
2.spaCy (Honnibal and Montani 2019) で、節に分割
    1.token.lemma_ でレマ化
    2.token.pos_ で品詞タグ付け
    3.token_dep_ で統語依存関係タグ付け

■分析手順

1.11 種類の構文に分類
2. 構文の延べ頻度（節の数と同じ）を計算
3. 構文の異なり頻度を計算（構文多様性）
4. 各構文の頻度と割合を計算
5. 各構文内の動詞の異なり頻度を計算（動詞多様性）

■ 11 の構文

・6 は構文文法より（Goldberg 1995）
    1. 使役移動 : She faxed a letter to him.
    2. 二重目的語 : She faxed him a letter.
    3. 自動詞移動 : The fly buzzed into the room.
    4. 自動詞結果 : The pond froze solid.
    5. 句動詞　: The girl looked the name up.
    6. 他動詞結果 : The girl made the can flat.
・5 は基礎的
    7. 叙述　: She is a student.
    8. 受け身　: It was folded.
    9. 単純自動詞 : I worked.
    10. 単純他動詞 : The man kicked the ball.
    11. 存在の there: There is a house.

■精度検証

・ American National Corpus (Reppen et al 2005)
・ 1000 の節の分類
    ・ 応用言語学者二人　Cohen's kappa 1.00
    ・ CDA
        ・ recall 0.82

・ precision 0.86
・ F1 0.82

■構文多様性スコア計算方法 Constructional Diversity Score

・log10(頻度 +1)
　・タイプ頻度に 1 をたしているのはラプラシアン平滑化（Manning et al. 2008）
　・対数を取っているのは標準化するため (Comptone et al. 2020)
　・各構文の比率を出して、
　・それを逆正弦変換する (Studebaker 1985)
　　・比率・割合データは、逆正弦変換することで正規分布に変換できやすい

■熟達度の予測

・A1 から C1 の 9 段階は 1 から 9 の連続変数として
　・ 4 レベル以上の順序変数は連続変数として扱う慣例 (Labovitz 1970, Robitzsch 2020)
・構文多様性指標は正規分布に従っていた（Q-Q plot）
・VIF を考慮して残ったのは以下の 7 つ
　1.there 構文
　2. 受け身
　3. 動詞句
　4. 使役移動
　5. 単純自動詞
　6. 二重目的語
　7. 叙述（ただし係数はマイナス）
・説明率は 11.5%

使い方

■ダウンロード (430MB)

・圧縮されているが、単一のアプリ
・起動に時間がかかる（3 分くらい）

■何を出力するか選択（チェック入れる）

・構文多様性 constructional diversity
・動詞多様性 verbal diversity

■分析対象ファイルの入っているフォルダーを選択

・ [Location] で選ぶ
　・ プログラムの入っているドライブ内からしか選択できない
　・ ルートから順に下にたどっていく
　・ 日本語ファイル名は文字化け
・ プログラムのフォルダー内にデータをコピーしたほうが便利
・ フォルダーを選んで、[Select]
・[Process] を押して実行
　・ 少し時間がかかる（量にもよるが、2 分くらい）

■結果は、分析対象にしたフォルダー内に csv で保存される

■ constructional_diversity.csv

■ verbal_diversity.csv