# 進捗報告

## 1 今週やったこと

- JSAI 予備実験，概要作成
- 内容語に関して単語出現頻度の確認

### 1.1 JSAI 予備実験，概要作成

精度がベースラインをほとんど超えなかったが一応スクリプトを完成させて実験を回した．CAE, MLP は PyTorch で実装し直した．

実験としては画像と台詞を分散表現化し，MLP を用いてそのマッチング問題を解いた．

訓練には少年漫画タッチ以外のタイトルを用い，テストには少年漫画タッチを用いた．訓練と評価画像の分割は 9:1 にした．また，3 層 MLP の中間層は 1500 とした．総データ数は 2802 件 (内テストが 80 件) である．

評価指標には平均逆順位 (MRR) を用い，分散表現同士の距離の指標にはユークリッド距離を用いた．

図 1 に CAE でデコードされた画像を示す．ダウンサンプリングには map pooling, アップサンプリングには linear upsampling を用いた．

図 2 に CAE のロスを示す．

復元された画像やロスを見ても CAE の学習はうまくいっているように見える．

表 1 に数値実験の結果を示す．

表 1: 台詞マッチング識別結果

| - | 訓練識別率 | 評価識別率 | テスト識別率 | ベースライン |
|---|---|---|---|---|
| 画像から台詞 | 0.999 | 0.345 | 0.255 | 0.293 |
| 画像 (分散表現) から台詞 (分散表現) | 0.625 | 0.340 | 0.272 | 0.293 |
| 台詞 (分散表現) から画像 (分散表現) | 0.352 | 0.304 | 0.284 | 0.293 |

また，図 3 にロスの遷移を示す．

### 1.2 内容語に関して単語出現頻度の確認

表 2 にデータセットの単語出現分布を示す．

下に各タイトルの各品詞におけるヒストグラムを示す．全然関係ないが matplotlib の日本語豆腐問題がアップデートで解決してて地味に嬉しかった．

## 2 来週の予定

DCAI を書きます．月 火曜日を目処に第一稿を仕上げようと思います．

表 2: 内容語に関する単語出現頻度の多いもの 4 件

| タイトル | 名詞 | 動詞 |
|---|---|---|
| 幼稚園ぼうえい組 | 桃 (42), 先生 (40), こと (36), 園長 (36) | して (27), ある (11), し (10), いう (10) |
| 徹さん | 徹 (44), ワイ (42), の (35), 極道 (33) | して (43), し (32), した (27), する (26) |
| OL ランチ | の (48), こと (36), 人 (33), あたし (33) | して (37), ある (24), する (20), し (20) |
| 高校の人達 | ・・・(328), 先生 (168), の (53), 部 (45) | し (59), して (56), する (44), いう (34) |
| 少年漫画タッチ | 研究 (7), 何 (6), 服 (5), 室 (5) | して (7), 見て (4), 言わ (3), ある (3) |

| タイトル | 形容詞 | 副詞 |
|---|---|---|
| 幼稚園ぼうえい組 | いい (22), ほのか (16), ない (12), 平和 (11) | もう (6), さすが (5), ちょっと (4), なんで (4) |
| 徹さん | ない (22), でっか (17), いい (14), へん (8) | やっとり (12), ちょっと (9), 全て (4), さらに (4) |
| OL ランチ | いい (55), ない (38), かわいい (10), 多い (8) | もう (17), ちょっと (14), まだ (14), やっぱ (9) |
| 高校の人達 | いい (38), ない (28), そ (17), な (12) | もう (32), みんな (29), なぜ (15), 一番 (11) |
| 少年漫画タッチ | 悪い (3), 好きな (3), いい (3), 大丈夫です (2) | いつも (4), 少々 (3), あまり (2), ほとんど (2) |

(a) オリジナル画像

(b) デコード画像 (評価)

図 1: CAE デコード結果

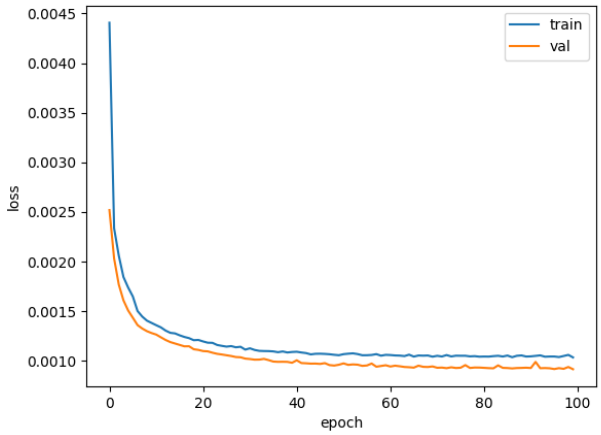
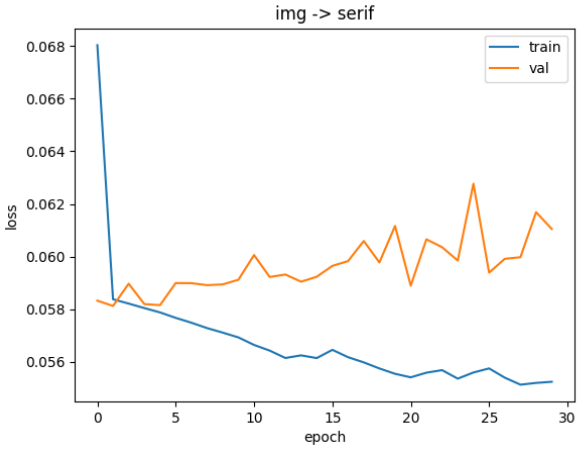
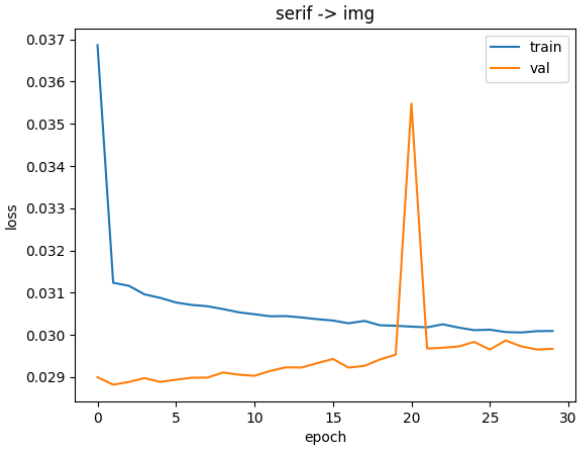



図 2: CAE ロス


(a) 画像から台詞 (分散表現)

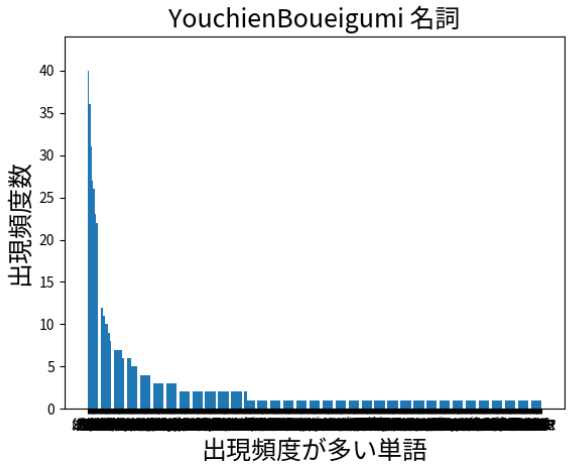


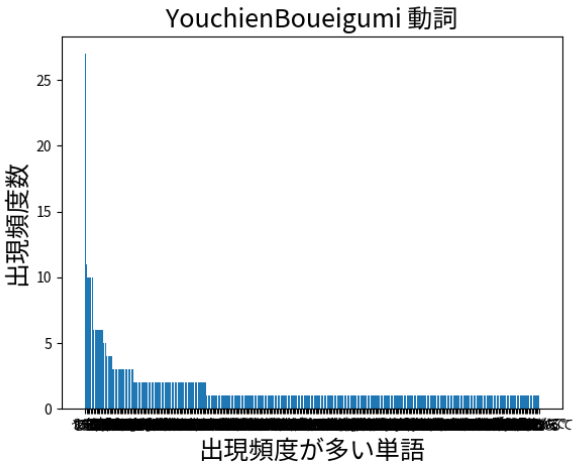

(b) 画像 (分散表現) から台詞 (分散表現)

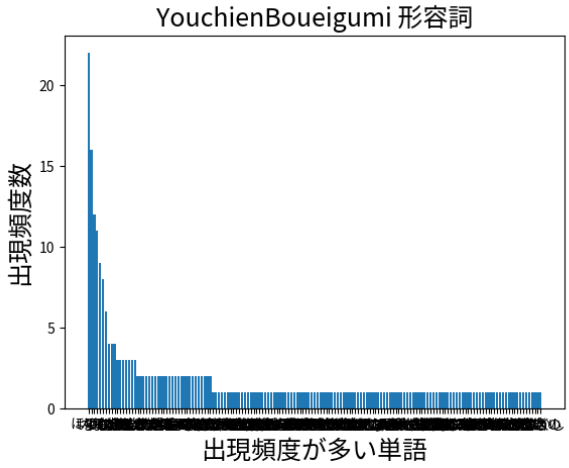


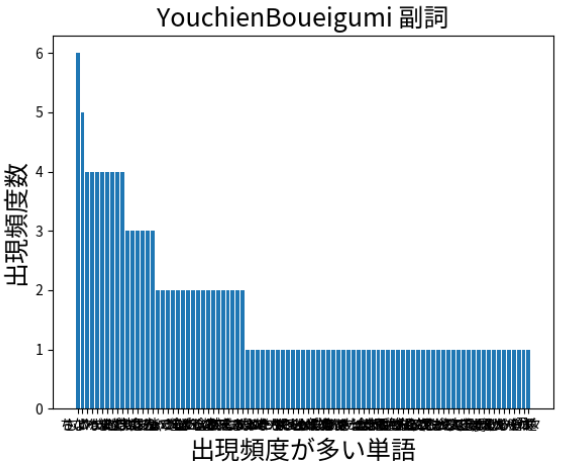

(c) 台詞 (分散表現) から画像 (分散表現)

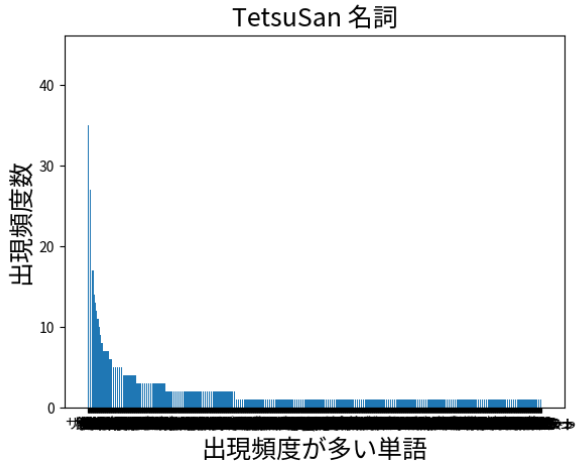
図 3: 各実験におけるロスの遷移

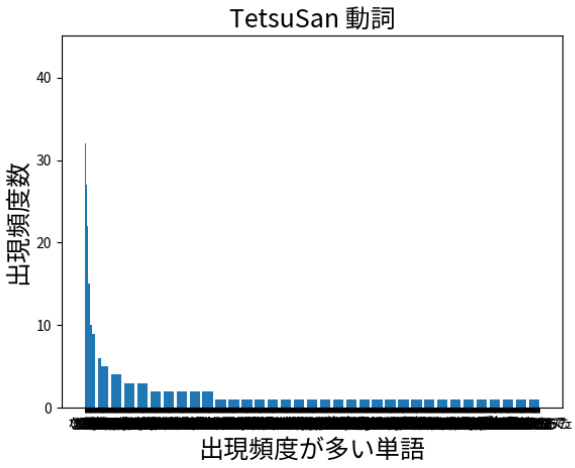

YouchienBoueigumi 名詞
出現頻度数
出現頻度が多い単語

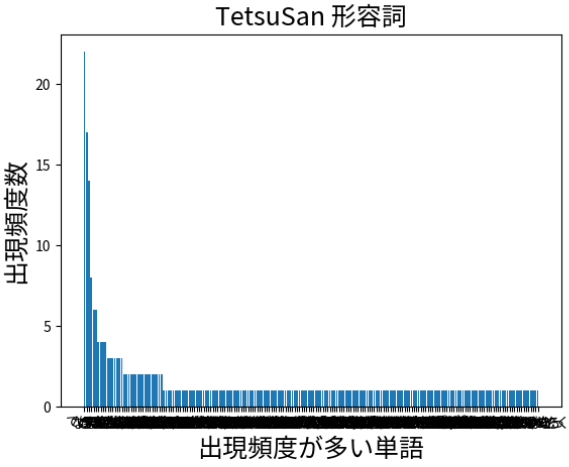

YouchienBoueigumi 動詞
出現頻度数
出現頻度が多い単語

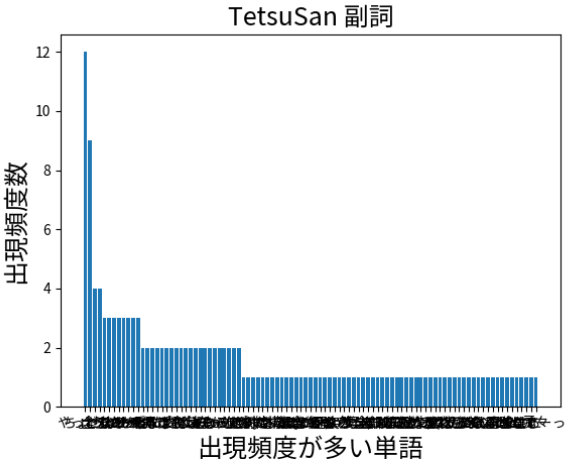

YouchienBoueigumi 形容詞
出現頻度数
出現頻度が多い単語

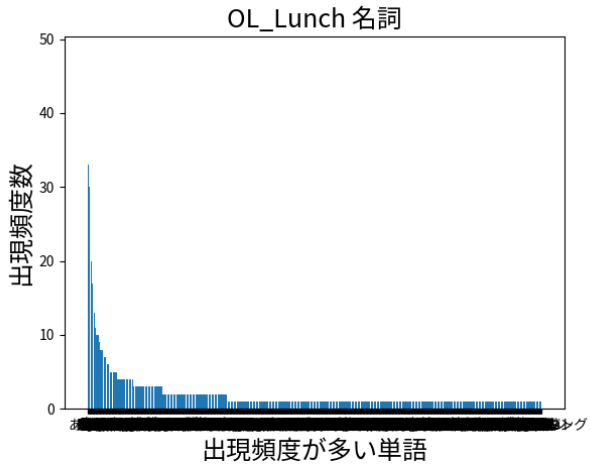

YouchienBoueigumi 副詞
出現頻度数
出現頻度が多い単語

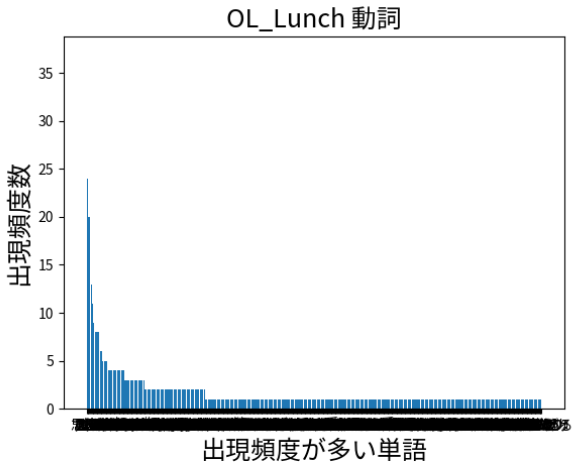

TetsuSan 名詞
出現頻度数
出現頻度が多い単語

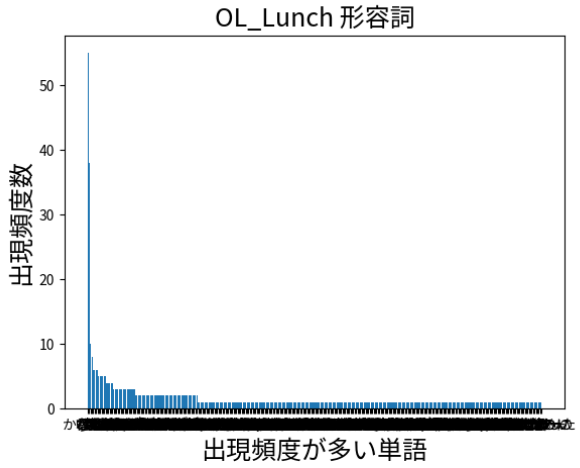

TetsuSan 動詞
出現頻度数
出現頻度が多い単語

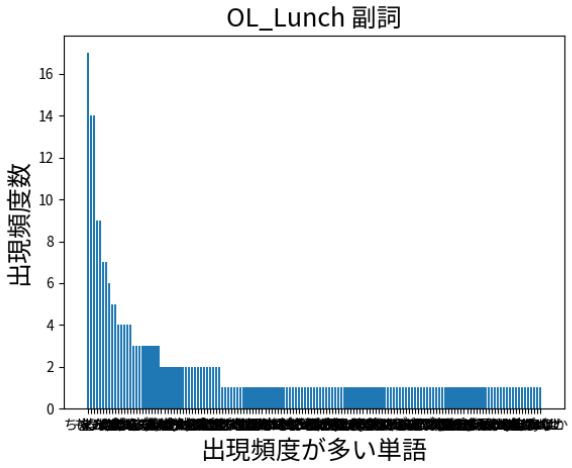

TetsuSan 形容詞
出現頻度数
出現頻度が多い単語

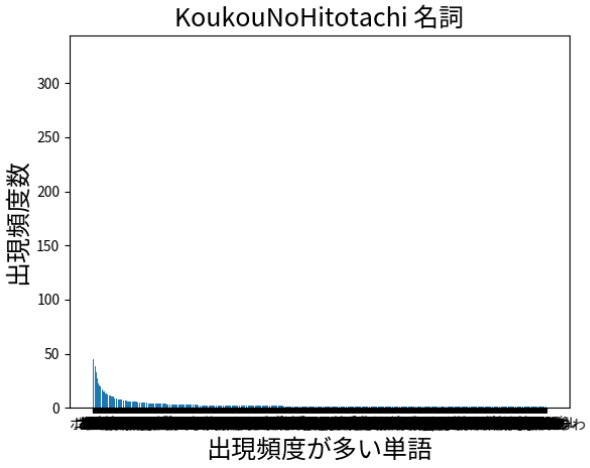

TetsuSan 副詞
出現頻度数
出現頻度が多い単語

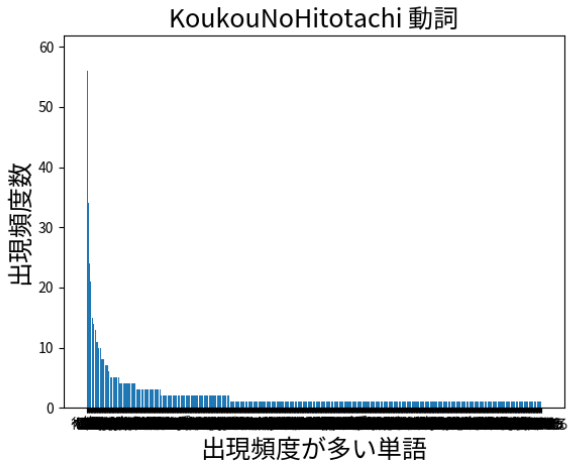

OL_Lunch 名詞
出現頻度数
出現頻度が多い単語

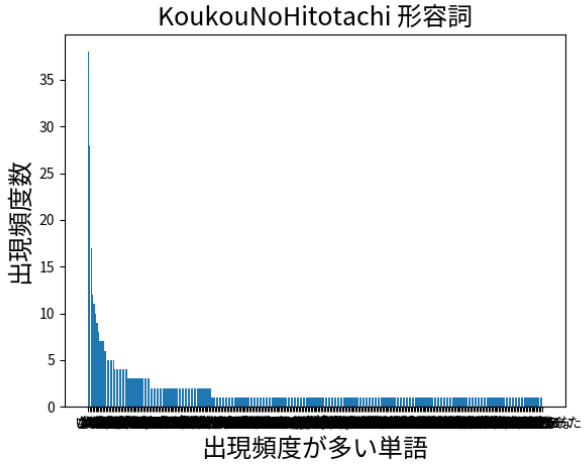

OL_Lunch 動詞
出現頻度数
出現頻度が多い単語

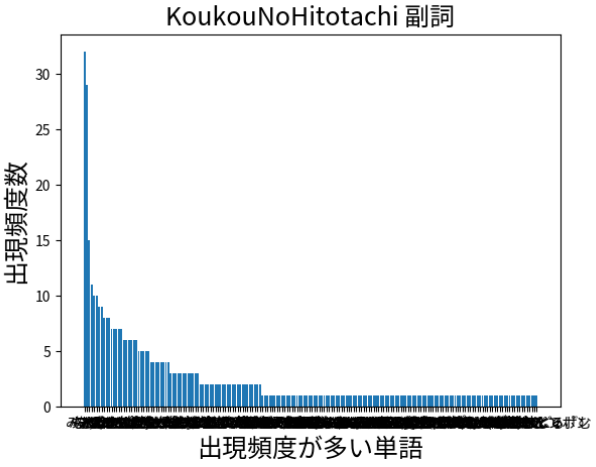

OL_Lunch 形容詞
出現頻度数
出現頻度が多い単語

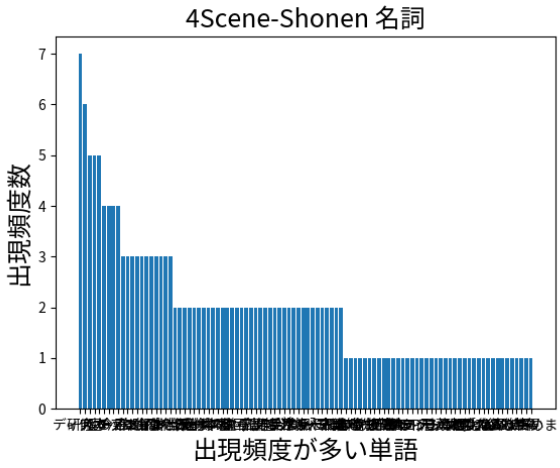

OL_Lunch 副詞
出現頻度数
出現頻度が多い単語

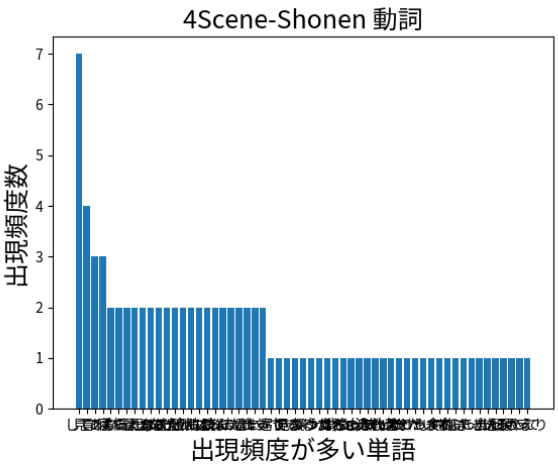

KoukouNoHitotachi 名詞
出現頻度数
出現頻度が多い単語

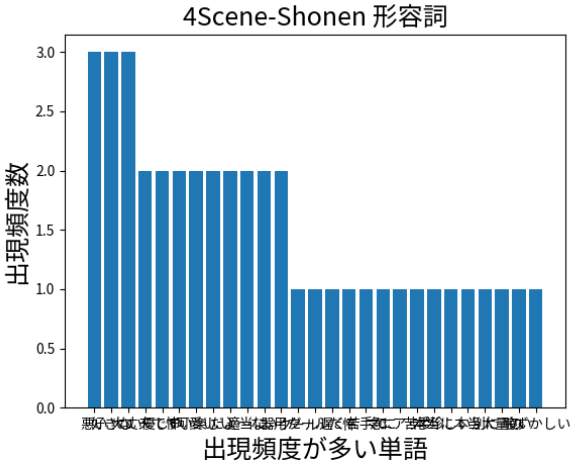

KoukouNoHitotachi 動詞
出現頻度数
出現頻度が多い単語

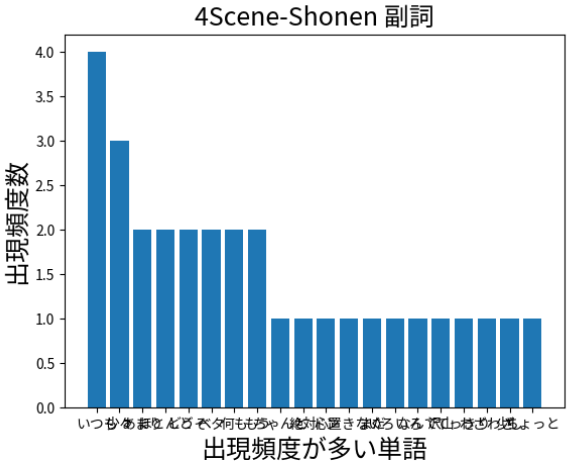

KoukouNoHitotachi 形容詞
出現頻度数
出現頻度が多い単語


KoukouNoHitotachi 副詞
出現頻度数
出現頻度が多い単語


4Scene-Shonen 名詞
出現頻度数
出現頻度が多い単語


4Scene-Shonen 動詞
出現頻度数
出現頻度が多い単語


4Scene-Shonen 形容詞
出現頻度数
出現頻度が多い単語


4Scene-Shonen 副詞
出現頻度数
出現頻度が多い単語