『昭和話し言葉コーパス』短単位語彙表・語数表 ver. 2025.03 解説

１．データの概要

本データは、『昭和話し言葉コーパス』(Showa Speech Corpus, SSC) の『中納言』(データバージョン 2022.02) に収録されているデータに基づいて、頻度 1 までの見出し語を対象にして作成した語彙表と語数表である。SSC には様々な属性情報が付与されているが、語彙表・語数表で取り上げる主な属性は次のとおり。

会話の属性：「形式」(講演、挨拶・祝辞、司会、雑談、用談相談、会議会合)
話者の属性：「性別」と「年齢」(5歳きざみ)

２．語彙表

ファイル名　1_ssc_frequencylist_suw_token.tsv (UTF8 タブ区切り)

- 語彙素読み、語彙素、品詞、語彙素細分類、語種で見出し語を特定したもの

３．語数表

ファイル名　2_ssc_wc.tsv (UTF8 タブ区切り)

- 資料ID、話者ID別の延べ語数表

ファイル名 2b_ssc_wc.xlsx

- 2_ssc_wc.tsvを対象に形式・年齢・性別で延べ語数を集計して作成。
- 粗頻度から調整頻度を求める場合の参照用に作成。
- その他の語数表が必要な場合は、各語彙表・語数表をもとに集計されたい。

４．品詞構成表

ファイル名 3_ssc_frequencylist_pos.tsv　(UTF8 タブ区切り)

- 以下の4つの表を収めた。
  (1) 短単位における品詞の語数（延べ語数）
  (2) 短単位における品詞の語数（異なり語数）
  (3) 短単位における品詞の割合（延べ語数）
  (4) 短単位における品詞の割合（異なり語数）
- 列は、形式、性別、年齢、性別と年齢

５．語種構成表

ファイル名 4_ssc_frequencylist_wtype.tsv （UTF8 タブ区切り）

・品詞構成表と同様に4つの表を収めた。表の種類は品詞構成表と同じ。

６．利用上の注意

- 研究、教育目的であれば無償で自由に利用できる。申し込みの必要はない。
- 再配布は不可。商業使用（営利目的での利用）は要相談。
- 論文等に引用する際は出典とバージョンを明記すること。以下に例を示す。

  『昭和話し言葉コーパス』短単位語彙表 ver. 2025. 03

  『昭和話し言葉コーパス』短単位語数表 ver. 2025. 03

  『昭和話し言葉コーパス』短単位品詞構成表 ver. 2025. 03

  『昭和話し言葉コーパス』短単位語種構成表 ver. 2025. 03
- 本データの著作権（編集著作権）は国立国語研究所が有する。
- データの瑕疵による損害についてはいかなる場合でも補償しない。
- 内容の改善のため予告なく更新することがある。
- 本データに関する問い合わせ先：kotonoha[at]ninjal. ac. jp