『日本語歴史コーパス（CHJ)』長単位語彙表解説

1.データの概要

・本データは『日本語歴史コーパス』（以下、CHJ と略す。）の中納言データ ver.2018.9 に基づく語彙表である。

・CHJ を構成するサブコーパスについて頻度 1 までの見出し語を収録した。

・下表のサブコーパスの長単位について語彙表、品詞別度数表、語種構成表を作成した。

時代編	シリーズ	サブコーパス名
奈良	Ⅰ	万葉集
平安	Ⅰ	古今和歌集
平安	Ⅰ	土佐日記
平安	Ⅰ	竹取物語
平安	Ⅰ	伊勢物語
平安	Ⅰ	落窪物語
平安	Ⅰ	大和物語
平安	Ⅰ	枕草子
平安	Ⅰ	源氏物語
平安	Ⅰ	紫式部日記
平安	Ⅰ	和泉式部日記
平安	Ⅰ	平中物語
平安	Ⅰ	堤中納言物語
平安	Ⅰ	更科日記
平安	Ⅰ	讃岐典侍日記
平安	Ⅰ	蜻蛉物語
平安	Ⅰ	大鏡
鎌倉	Ⅰ説話・随筆	今昔物語集(本朝部)
鎌倉	Ⅰ説話・随筆	宇治拾遺物語
鎌倉	Ⅰ説話・随筆	十訓抄
鎌倉	Ⅰ説話・随筆	方丈記
鎌倉	Ⅰ説話・随筆	徒然草
鎌倉	Ⅱ日記・紀行	海道記
鎌倉	Ⅱ日記・紀行	建礼門院右京大夫集
鎌倉	Ⅱ日記・紀行	東関紀行
鎌倉	Ⅱ日記・紀行	十六夜日記
鎌倉	Ⅱ日記・紀行	とはずがたり
室町	Ⅰ狂言	虎明本狂言集
室町	Ⅱキリシタン資料	天草版平家物語
室町	Ⅱキリシタン資料	天草版伊曽保物語

2.集計方法

(1)語彙素読み、語彙素、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定した。

(2)(1)で得られた見出し語の集合から以下の条件に該当するものを除外した。

1)品詞に「空白」「補助記号」「記号」の文字列を含むもの。

2)語彙素が空(null)のもの（この場合、語彙素読みも同時に空になっている）

3.CHJ 長単位語彙表

(1)語彙別

①ファイル名：CHJ_サブコーパス名_frequencylist_luw.ver2018_9.tsv

②UTF8、タブ区切り。

③第1行目は見出し。2行目以降がデータ。

④見出しは、語彙素読み、語彙素、品詞、語彙素細分類、語種、頻度の6項目。

⑤Excel等のソフトに読み込んで、目的の列で並べ替えた語彙表を作ることができる。

(2)語種別

①ファイル名：CHJ_サブコーパス名_frequencylist_luw_gohsu.ver2018_9.tsv

②UTF8、タブ区切り。

③第1行目は見出し。2行目以降がデータ。

④語種としては和語、固有語、外来語、混種語、漢語を立てている。

(3)品詞別

①ファイル名：CHJ_サブコーパス名_frequencylist_luw_pos.ver2018_9.tsv

②UTF8、タブ区切り。

③第1行目は見出し。2行目以降がデータ。

④サブコーパスにおける品詞ごとの述べ語数を示す。

4.利用上の注意

(1)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。

(2)再配布は不可。商業使用（営利目的での利用）は要相談。

(3)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、出典とバージョンの例を示す。

『日本語歴史コーパス』長単位語彙表 ver.2018.9

『日本語歴史コーパス』長単位品詞構成表 ver.2018.9

『日本語歴史コーパス』長単位語種構成表 ver.2018.9

(4)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。

(5)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。

(6)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@ninjal.ac.jp（@を半角に変えること）

以上