

「日本語歴史コーパス 平安時代編」 語彙統計

2014 年 3 月 31 日

1. 目的

ここでは、「日本語歴史コーパス 平安時代編」の語彙に関する数値の集計結果を示す。

2. 粗頻度

ジャンル・作品名・成立年・巻名等・本文種別ごとの語数（記号等除外¹／全て）は、中納言 wiki の CHJ/短単位語数、CHJ/長単位語数各ページからダウンロードできるので、参照されたい。

短単位 CHJ_SUW_WC_v10.xlsx

長単位 CHJ_LUW_WC_v10.xlsx

3. 資料規模

表 1・2 に、中古和文 14 作品の延べ語数・異なり語数（短単位・長単位）を示す。集計にあたり、記号等は除外した。

表 1 延べ語数・異なり語数（短単位）

	延べ語数	異なり語数
古今和歌集	31257	2553
竹取物語	10317	1272
伊勢物語	13824	1646
大和物語	23090	1941
平中物語	12403	1227
土佐日記	6685	1010
落窪物語	54583	2608
堤中納言物語	15697	1852
枕草子	66034	4078
源氏物語	445674	6701
和泉式部日記	10891	1109
紫式部日記	17441	2185
更級日記	14659	1751
讃岐典侍日記	15546	1607

表 2 延べ語数・異なり語数（長単位）

	延べ語数	異なり語数
古今和歌集	29299	3010
竹取物語	9440	1548
伊勢物語	13049	1875
大和物語	21155	2478
平中物語	11773	1458
土佐日記	6292	1093
落窪物語	48430	4311
堤中納言物語	14044	2422
枕草子	60158	6046
源氏物語	387788	17089
和泉式部日記	10060	1338
紫式部日記	14981	2866
更級日記	13536	2176
讃岐典侍日記	13315	2120

¹ 「記号等」とは、空白・補助記号・漢文・中古解釈不明を指す。

4. 語種比率

図 1・2 に、中古和文 14 作品の語種比率（短単位・長単位）を示す。語数については、中納言 wiki の語彙統計ページからダウンロードできるので、参照されたい。

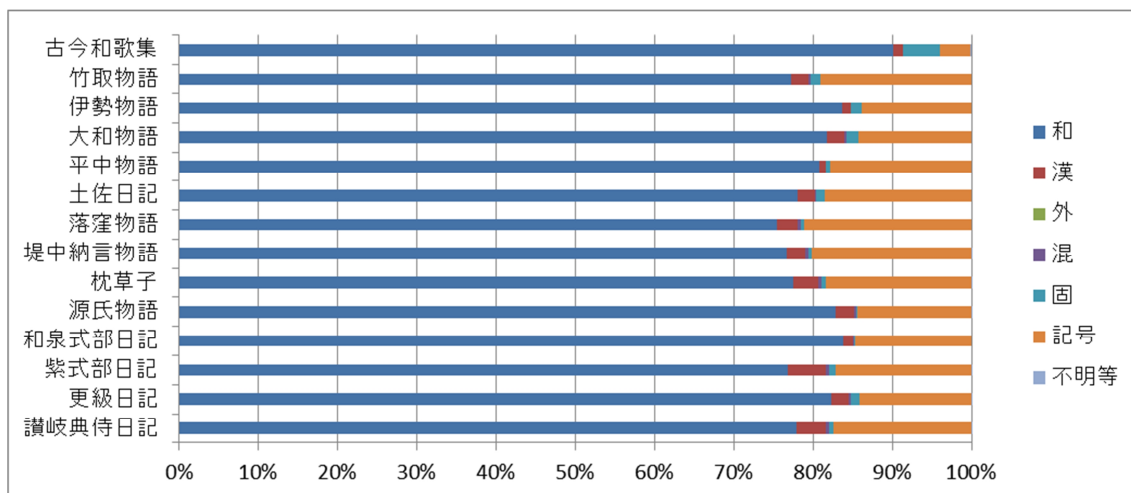


図 1 語種比率（短単位）

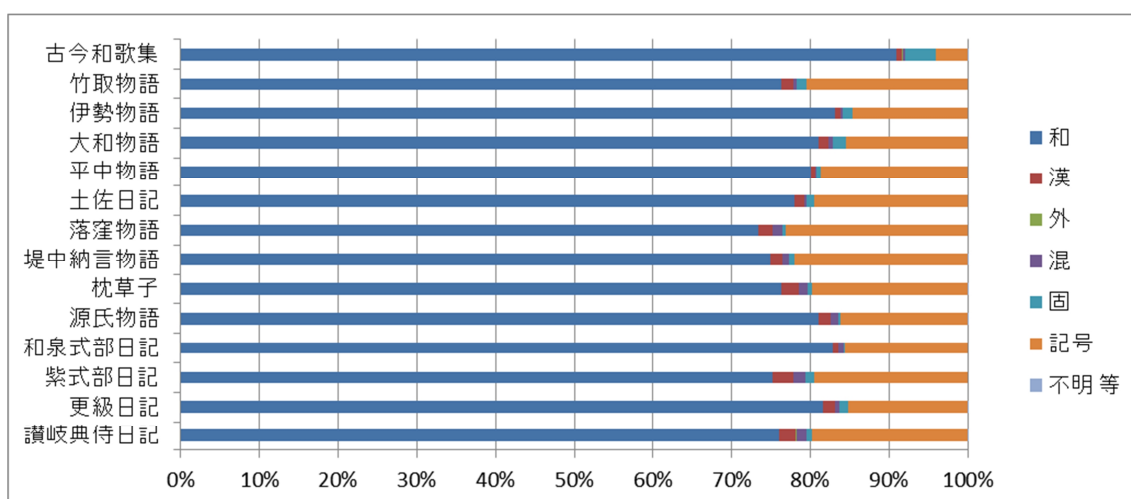


図 2 語種比率（長単位）

短単位 CHJ_SUW_WT.xlsx

長単位 CHJ_LUW_WT.xlsx

5. 品詞比率

図3・4に、中古和文14作品の品詞比率（短単位・長単位）を示す。数値については、中納言 wiki の語彙統計ページからダウンロードできるので、参照されたい。短単位・長単位の品詞は「名詞（大分類）-普通名詞（中分類）-形状詞可能（小分類）」といった階層を持つ。名詞については中分類（普通名詞、固有名詞、数詞）別に、名詞以外の品詞は大分類別に集計し、記号類は除外した。

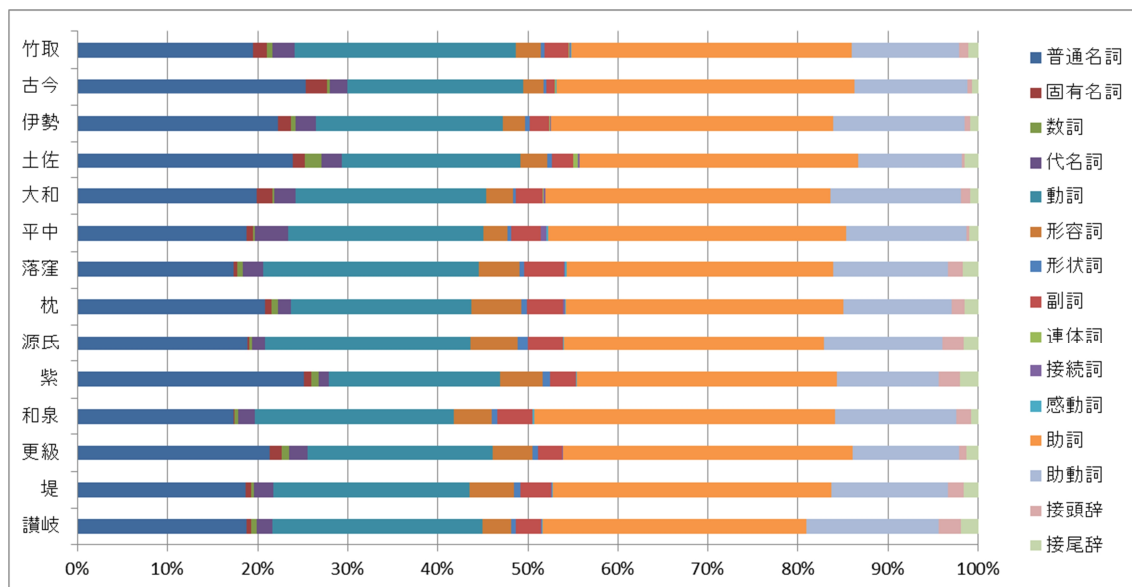


図3 品詞比率（短単位）

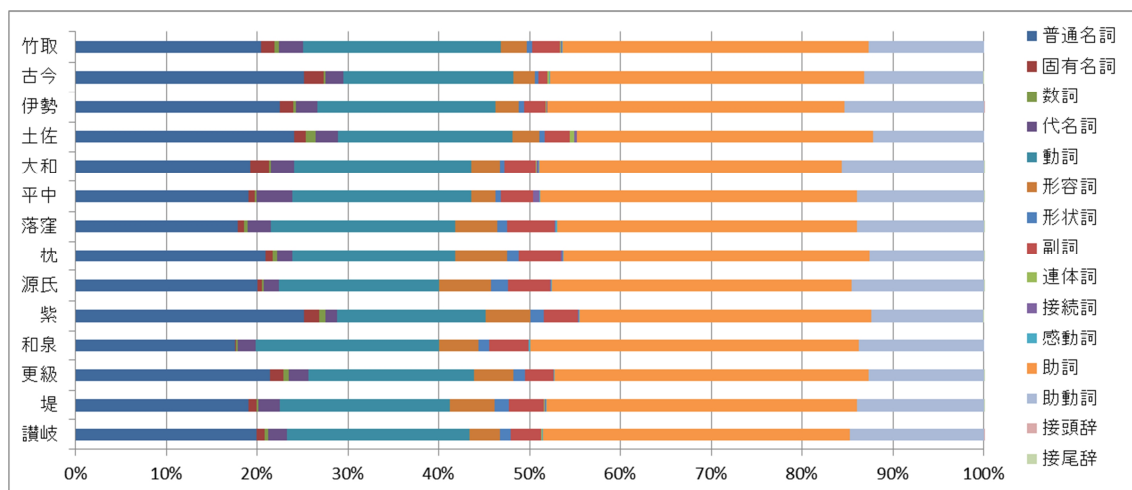


図4 品詞比率（長単位）

短単位 CHJ_SUW_PR.xlsx

長単位 CHJ_LUW_PR.xlsx

6. 高頻度語

自立語の高頻度上位 10 語は、中納言 wiki の語彙統計ページからダウンロードできるので、参照されたい。

短単位 CHJ_SUW_HF.xlsx

長単位 CHJ_LUW_HF.xlsx

以 上