「日本語歴史コーパス　室町時代編Ⅰ　狂言」（短単位 Ver.0.9）語彙統計

2015 年 4 月 14 日

1. 目的

ここでは、「日本語歴史コーパス　室町時代編Ⅰ　狂言」短単位 Ver.0.9 の語彙に関する数値の集計結果を示す。なお、長単位については現在整備中であるため、本稿では短単位の情報のみを提示する。

２．粗頻度

巻名・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_v201503.xlsx

３．資料規模

表 1 に、『虎明本狂言集』各類の延べ語数・異なり語数（短単位）を示す。集計に当たり、記号等は除外した。

表 1　述べ語数・異なり語数(短単位)

	延べ語数	異なり語数
脇	27442	8265
大名	38574	10466
聟・山伏	23438	7121
鬼・小名	33733	10230
女	33263	9950
出家座頭	29712	9491
萬集	15556	6717
集	33149	9809
全体	234867	72049

[1] 「記号等」とは、空白・補助記号・解釈不明等の未知語類を指す。未知語の種類に関して「『日本語歴史コーパス　室町時代編Ⅰ狂言』形態論情報（短単位 Ver.0.9）の概要」を参照。

４．語種比率

図１に、各類の語種比率（短単位）を示す（記号等も含める）。

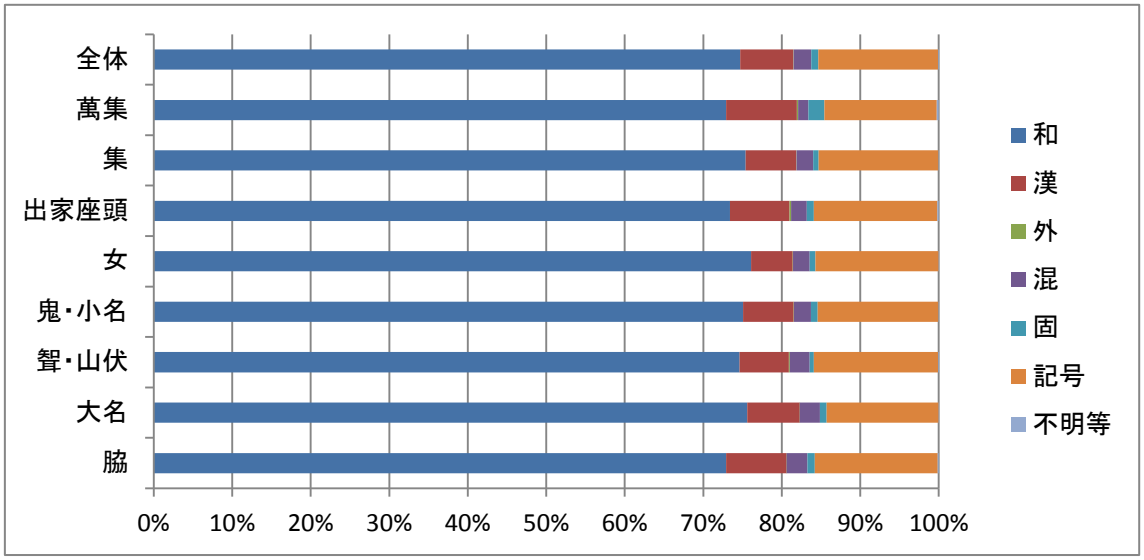

図　１　語種比率（短単位）

短単位　CHJ_SUW_WT.xlsx

５．品詞比率

図２に、各類の品詞比率（短単位）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。

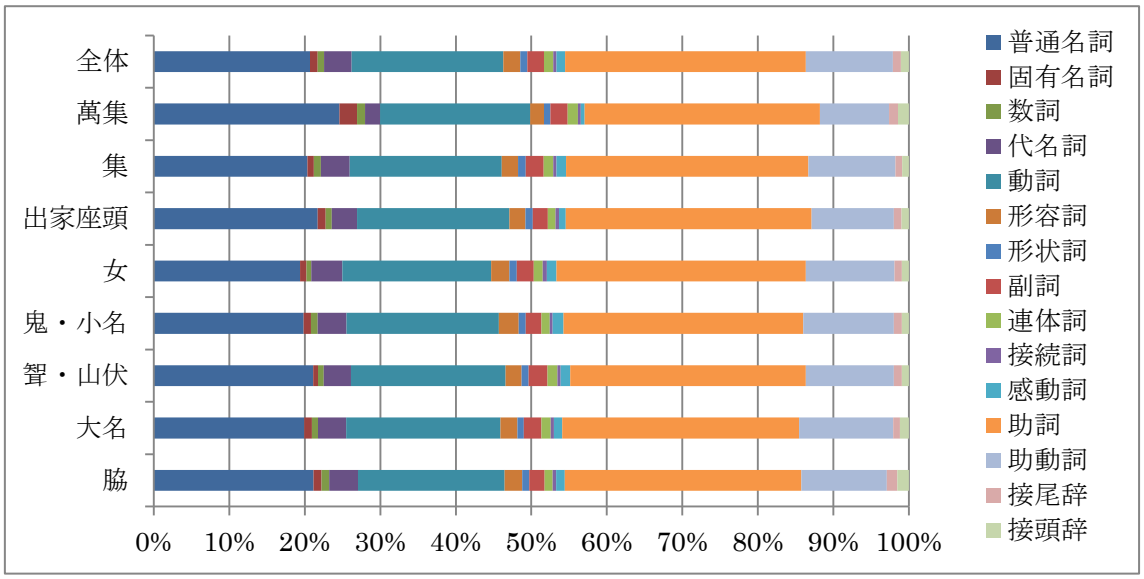

図　２　品詞比率（短単位）

短単位　CHJ_SUW_PR.xlsx

６．高頻度語

自立語の高頻度語上位 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

以上