コーパス検索システムの言語単位『短単位』

岡照晃・石本祐一・大村舞　（国立国語研究所　コーパス開発センター）

短単位とは

- 「国立国語研究所」は、
  国立国語研究所　で、１単語？
  国立＼国語研究所　で、２単語？
  国立＼国語＼研究所　で、３単語？
  国立＼国語＼研究＼所　で、４単語？
  国立＼国＼語＼研究＼所　で、５単語？
- 「青空」は、１単語？　２単語？
- 「茶色」は、１単語？　２単語？

人によっても
使われる状況によっても変わる

用例収集のための
言語単位として

短単位

短単位だと…国立＼国語＼研究＼所　4単語

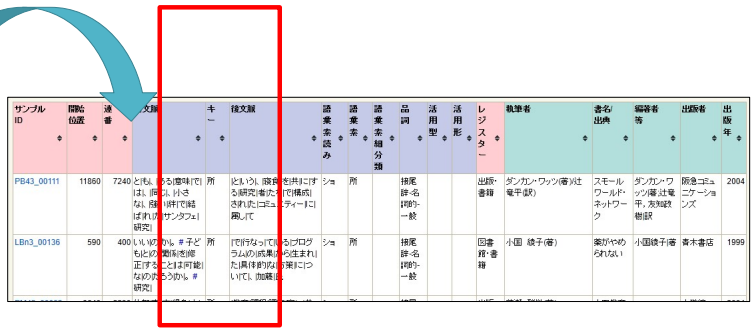
青空　１単語　　茶色　１単語

短単位の使用例

『中納言』・まとめて検索『KOTONOHA』
の単語検索は短単位

まとめて検索
『KOTONOHA』

中納言



短単位の区切りを調べるなら
「中納言」で調べられる

短単位の規定

結合方法をまとめた表

分類		例	結合方法
一般	和語	豊か　大　雨…	２つまで結合可
	漢語	国　語　研　究　所…	２つまで結合可
	外来語	コール　センター　オレンジ…	結合不可
付属要素	接頭的要素	御　各…	結合不可
	接尾的要素	兼ねる　がたい　的…	結合不可
その他	記号	、　・　。　「　」…	結合不可
	数	一　二　十　百　千…	「数」同士で結合可
	人名・地名	星野　仙一　大阪　六甲…	結合不可
	助詞・助動詞	だ　ます　か　から　て　の…	結合不可

意味を持つ最小の単位を結合して『短単位』を組み立てる

1. 和語・漢語は、２最小単位の１次結合体を１短単位とする
   - ◆｜母=親｜
   - ◆｜言=語｜資=源｜
   - ◆｜本=箱｜作り｜
2. 外来語は、１最小単位を１短単位とする
   - ◆｜コール｜センター｜
   - ◆｜オレンジ｜色｜

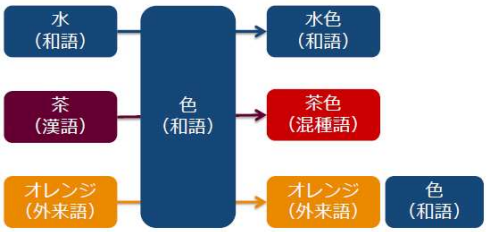
基本ルール

① 和語＋和語の結合は２つまで

② 漢語＋漢語の結合は２つまで

③ 外来語は結合させない

④ 和語＋漢語，漢語＋和語の結合も２つまで（混種語）



まとめて検索『KOTONOHA』　https://chunagon.ninjal.ac.jp/integrated/
中納言　https://chunagon.ninjal.ac.jp/