異なり語数の推定

異 な り 語 数 の 推 定

山崎　誠

1．はじめに

2．異なり語数と異なり表記形数

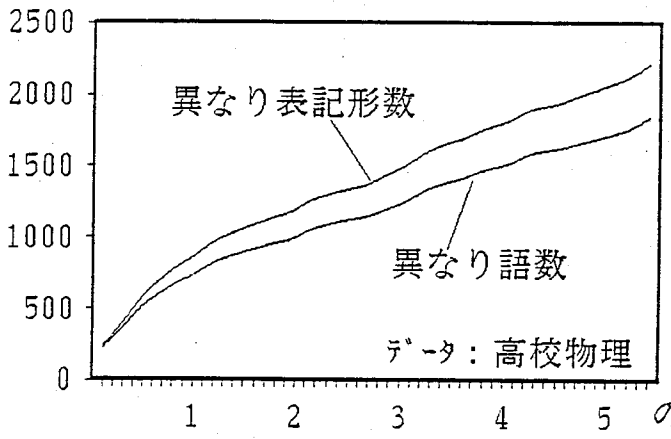
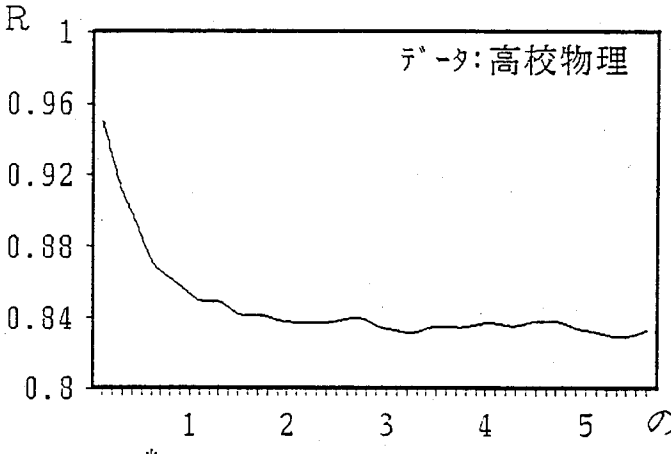


図１　異なり語数と異なり表記形数の推移



$$^{*}R = \frac{\text{異なり語数}}{\text{異なり表記形数}}$$

図２　R^{*}の値の推移

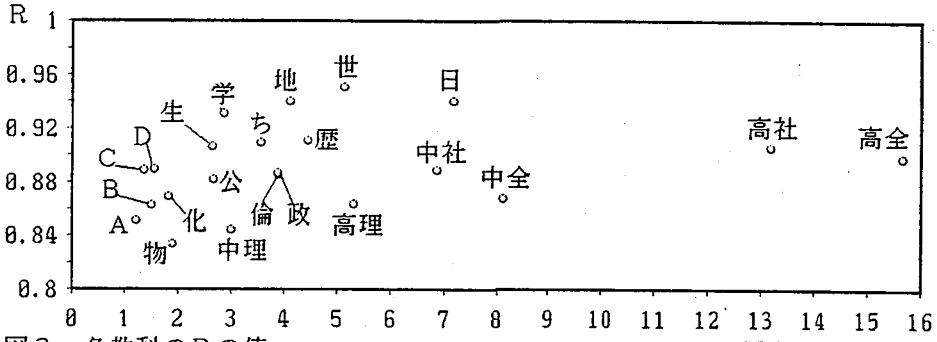


図3 各教科のRの値

高校
物…物理　倫…倫理社会
化…化学　政…政治経済
生…生物　日…日本史
学…地学　世…世界史
　　　　　地…地理

高理…理科系
高社…社会科系
高全…全教科

中学
A…理科1上　公…公民
B…理科1下　ち…地理
C…理科2上　歴…歴史
D…理科2下

中理…理科系
中社…社会科系
中全…全教科

3．1語に対する表記形の数

表1　1語に対応する表記形数の多いもの（高校教科書）

度数	見出し語	類	表記形例
21	とる	2	取れる
18	つくる	2	造り
16	はかる	2	謀り
13	かわる	2	変わる
13	たつ	2	立つ
13	ひく	2	引っ
13	ゆく	2	行く
12	あう	2	合わ
12	おく	2	置け
12	おくる	2	贈る
12	かえる	2	返る
12	かく	2	書ける
11	あらわす	2	表わせる
11	おこなう	2	行なわ
11	おこる	2	興る
11	きる	2	切れ
11	はなす	2	離そ
11	ひらく	2	開こ
11	もつ	2	持て

度数	見出し語	類	表記形例
10	したがう	2	従わ
10	する	2	支
10	たてる	2	立てる
10	つく	2	付く
10	なる	2	成る
10	のぞく	2	除こ
10	はたらく	2	働け
10	はやい	3	速く

「類」は，「分類語彙表」での分類番号の1桁めの数字

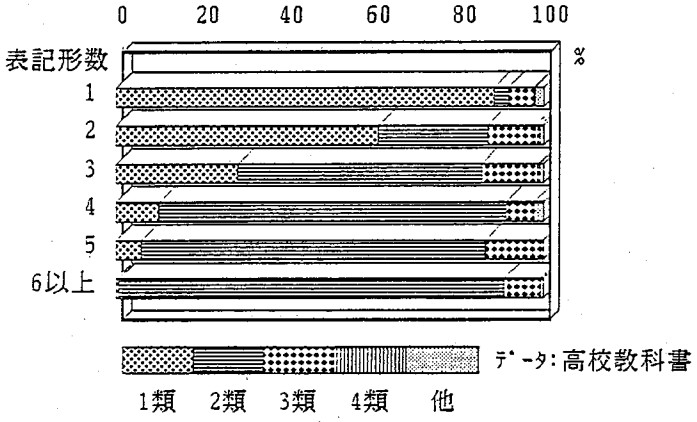
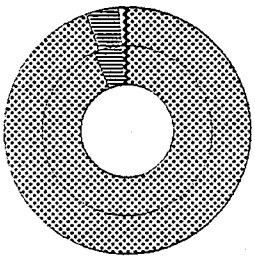


図４　１語あたりの異なり表記形数別の語類構成比

表２　Ｒと語類構成比との相関

	高　校	中　学
１類	0.913	0.680
２類	-0.554	-0.569
３類	-0.463	-0.279
４類	-0.669	-0.309

４．１表記形に対する見出し語の数



対応する語数	高校（外円）	中学（内円）
1語	94.7%	93.7%
2語	4.3%	5.2%
3語	0.8%	0.8%
4語	0.1%	0.2%
5語以上	0.1%	0.1%
合　計	100.0%	100.0%

図５　１つの表記形がいくつの語に対応したか

表３　１表記形に対応する語の数が多いもの（高校教科書）

度数	表　記　形
7	より
6	かけ　つい　下　元　日
5	かた　から　たち　たて　たら　つき　つけ　で　なり　まき　め 金　重　西　中　長　分
4	あい　いき　がけ　さし　さら　し　つぎ　なし　なれ　み　よ 家　間　京　月　原　源　出　上　人　生　土　文　米　本　名
3	あたり　あり　あわ　いっ　うち　え　おい　おき　か　かい かき　かぎ　かく　かっ　かね　かり　がら　き　くま　くみ くり　こめ　さ　さる　しめる　すすめ　すみ　ずみ　せ　そう たえ　たっ　たとえ　つみ　づけ　な　なく　に　のり　ば ふるい　ま　まい　ます　また　まっ　もと　やり　わかれ　ん オレンジ　ガン　サン　バス　ワット　円　縁　王　音　外　楽 巻き　汗　器　基　強　玉　近　形　軽　見　古　後　御　光　公 広　行　高　合　細　三浦　史　子　実　借　主　周　出し　商 床　上り　常　神　世　成功　星　青　石　赤　巣　大　谷　端 値　町　調　鳥　直　東　当　等　頭　白　板　封じ　風　仏　物 文治　平　歩　万　味　密　毛　木　夜　葉　両　刀　林　和

5．表記の種類とRの値

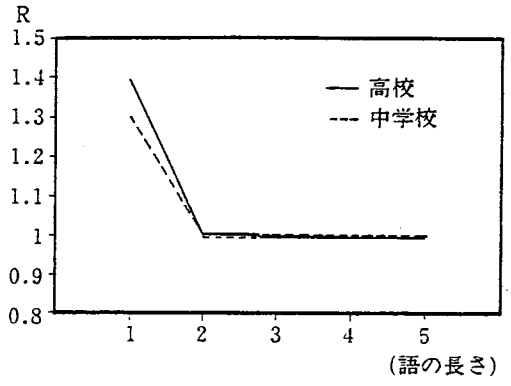


図6　Rの値と語の長さ（漢字表記形）

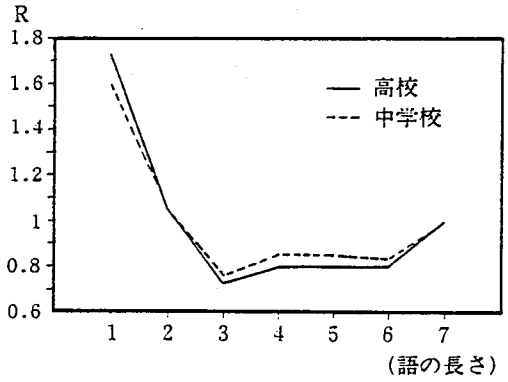


図7　Rの値と語の長さ（平仮名表記形）

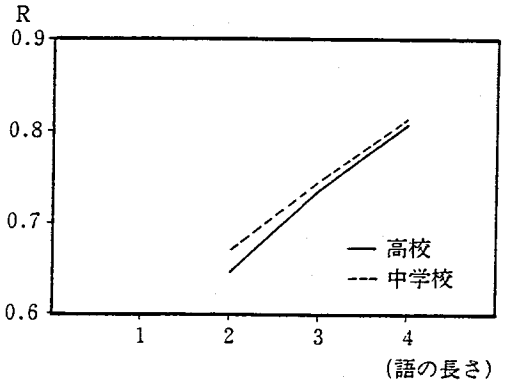


図8　Rの値と語の長さ（混ぜ書き表記形）

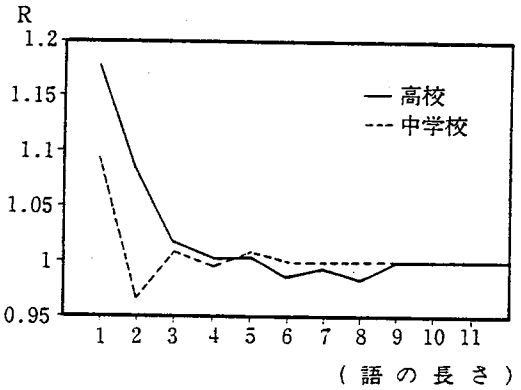


図9　Rの値と語の長さ（片仮名表記形）

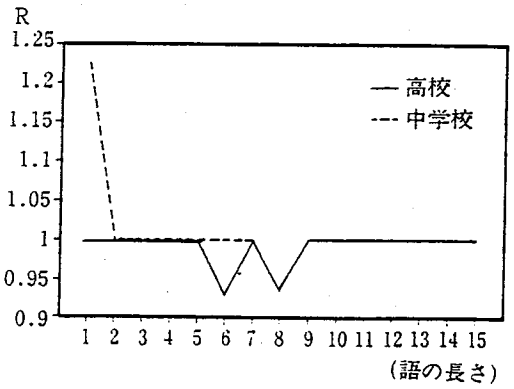


図10　Rの値と語の長さ（その他の表記形）

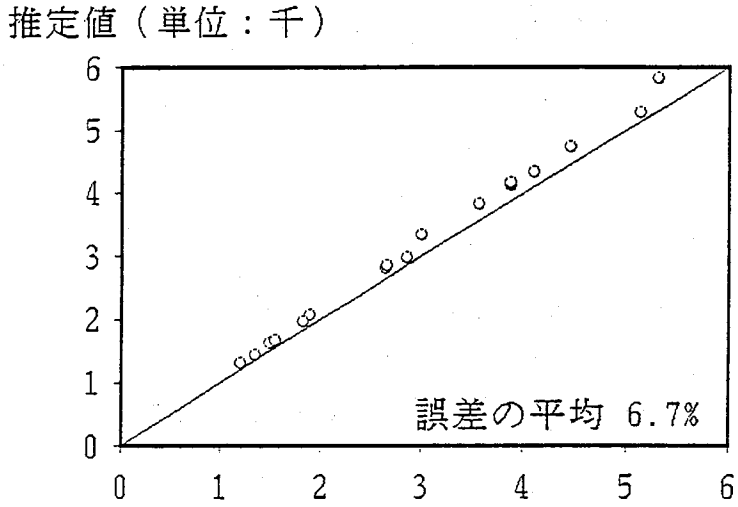
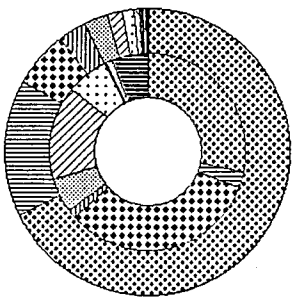


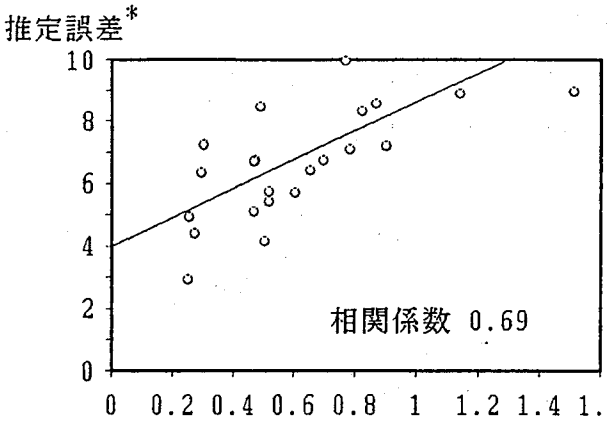
図１１　異なり語数の単純推定値



	異なり語数（外円）	のべ語数（内円）
漢	67.6%	28.2%
片	15.5%	2.6%
平	6.3%	32.5%
混	3.2%	1.4%
平混	2.6%	5.6%
他	1.8%	15.5%
漢平	1.3%	6.7%
漢片	0.6%	0.2%
漢混	0.4%	0.3%
漢平混	0.3%	0.9%
平片	0.2%	0.1%
漢平片	0.1%	6.1%
その他	0.1%	0.0%
合　計	100.0%	100.0%

図１２　１語の持つ表記パターン

ﾃﾞｰﾀ:高校教科書



A…漢字，片仮名，その他の異なり表記形数の和

B…平仮名，混ぜ書きの異なり表記形数の和

図１３　推定誤差と表記種の割合との相関

$$^{*}\text{推定誤差} = \frac{\text{異なり語数推定値} - \text{異なり語数}}{\text{異なり語数推定値}} \times 100$$

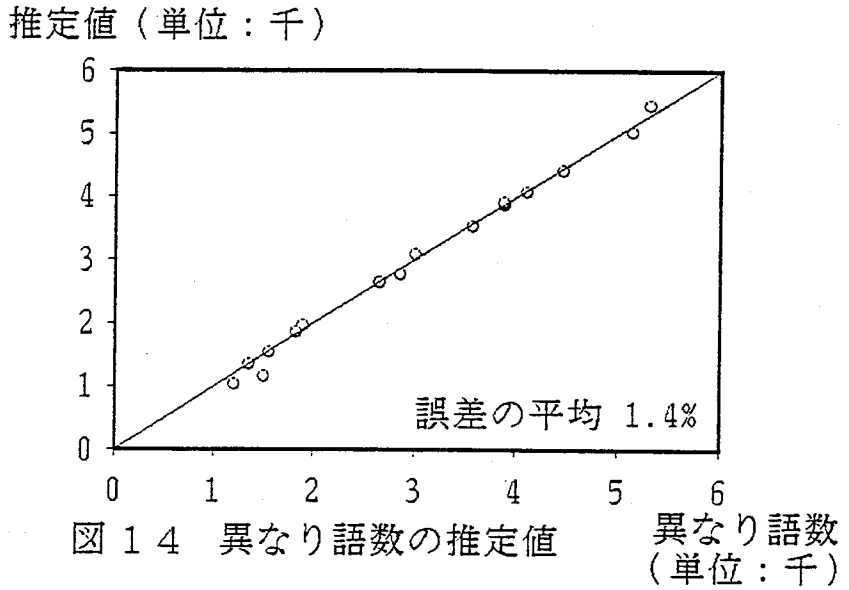


図１４　異なり語数の推定値

$$K' = E - \frac{E}{100}\left[4.647\,\frac{B}{A} + 3.978\right]$$

$$(E = 1.279C_1 + C_2 + 1.513H_1 + 0.992H_2 + 0.731H_3 + 0.648M_2 + 0.735M_3 + 0.845M_4 + 1.01K_1 + 1.02K_2 + K_3 + O)$$

但し，K′：異なり語数推定値，C_1～Bは以下のそれぞれの表記語に対応する異なり表記形数。

C_1：漢字1字　C_2：漢字2字以上　H_1：ひらがな1字　H_2：ひらがな2字

H_3：ひらがな3字以上　M_2：混ぜ書き2字　M_3：混ぜ書き3字

M_4：混ぜ書き4字以上　K_1：カタカナ1字　K_2：カタカナ2字

K_3：カタカナ3字以上　O：その他

A：漢字＋カタカナ＋その他　B：ひらがな＋混ぜ書き

E：異なり語数の単純推定値

6．まとめと課題