コロケーション強度を用いた中古語の語認定

著者	須永 哲矢
雑誌名	国立国語研究所論集
号	2
ページ	91-106
発行年	2011-11
URL	http://doi.org/10.15084/00000483

国立国語研究所論集 (NINJAL Research Papers) 2: 91–106 (2011)
ISSN: 2186-134X print/2186-1358 online

コロケーション強度を用いた中古語の語認定

須永　哲矢

国立国語研究所 コーパス開発センター 非常勤研究員

要旨

中古和文において，どこからどこまでを一語と認めるかという語認定には，従来明確な尺度がなく，既存の辞書の見出し語をあたっても，立項基準は感覚的・主観的なものであると言わざるを得ない。語と語の結びつきの強さ（コロケーション強度）を具体的な数値で示すダイス係数を取り上げ，「名詞＋あり／なし／よし／あし」の組み合わせを例に，語認定の一つの客観的基準として，ダイス係数が有効であることを論じた*。

キーワード：形態素解析辞書，中古語，コロケーション強度，ダイス係数

1. 古典語辞書における見出し語認定の問題

中古和文テキストを扱う際に，現代語にはない，中古特有の問題が生じることはままあるが，辞書編纂の分野では，どこからどこまでを一語と認定するのか判断しづらい，という問題が一つ挙げられる。例えば，「思ひ言ふ」というような例を複合動詞とみなしてよいのか，あるいは単なる動詞の並列と解釈するべきか。また「甲斐なし」は，全体で一つの形容詞と考えるべきか，あるいは「甲斐」―が―「なし」という主述関係をなす二語なのか。中古語においては，明らかに結合して一語になっていると言いたいものから，ただ並んでいるだけのように見えるものまで，二語以上がまとまって用言相当になっているケースにさまざまなものがあり，現代語に比べてそのようなまとまりを作る自由度が高い。二語以上のまとまりをどう分節すべきか判断しづらい，という例としては次のようなものが挙げられる。

(1)　はかなきさまにておはすらむと思ひ言ひけるを，（蜻蛉日記）
(2)　さては童べぞ出で入り遊ぶ。（源氏物語・若紫）
(3)　足ずりをして泣けどもかひなし。（伊勢物語）
(4)　行幸のことを興ありと思ほして，（源氏物語・末摘花）

(1)の下線部「思ひ言ふ」の例は，「思ひ言ふ」全体で複合動詞ととらえて一語と扱うのか，単に「思ふ」と「言ふ」の二語の並列ととらえるのか。現代語であれば，複合動詞の種類がもともとかなり固定されていること，もし単なる動詞の並列であれば「思って言う」「思ったり言ったりする」のような形をとり，ただ動詞が並ぶだけにはならないことなどから，判断に悩むことはさほどない。(2) の下線部の動詞の連続についても同様に，複合動詞か，あるいは並列か，という判断が

* 本稿の内容は 2011 年 7 月 26 日の NINJAL サロンおよび 2011 年日本語学会春季大会（須永・小木曽 2011）での発表をもとにしている。

難しい。

(3) の下線部「かひなし」のように，「名詞＋形容詞」の組み合わせにおいても，「甲斐なし」全体で一つの形容詞と見て一語と認定するか，「甲斐」―が―「なし」という主述関係と見て二語と認定するかの判断が必要となる。こちらも，もし現代語であれば，格関係にある名詞と形容詞は「甲斐がない」のように，格助詞や係助詞，副助詞など，なんらかの助詞をはさむので，間違えようがない。(4) の場合，名詞に後続する用言が形容詞ではなく動詞「あり」であるが，「あり」がサ変動詞のように一部の名詞を取り込んで一つの動詞「興あり」を構成している，と見ることもできるし，あくまで「興」―が―「ある」という主述関係にある二語，と見ることもできる。

中古語の語認定において大きく問題となるのは，主にこの二つのタイプである。現代語であれば，内省を頼る，またはインフォーマントにあたる，といった手段も使えるが，中古語ではそれもままならない。これらについて既存の辞書をあたったとしても，辞書編纂の現場でも，経験的・感覚的に見出し語を選出するのが現状であり，できればもう少しはっきりした，客観的な根拠がほしいところである。

そのような要請に対し，本稿はダイス係数という統計的指標を語認定の一つの基準とすることを提案するものである。二語以上の意味的まとまりに対し，その結合強度（コロケーション強度）を数値として取り出すダイス係数は，従来の経験的・感覚的語認定に対し，客観的な語認定の根拠となりうる。実際のところ，単語認定にあたっては，感覚に頼ることは重要であるし，また意味的側面なども考慮すべきなのは当然である。しかし本稿では，そのような目に見えないものを一度あえて無視し，外形的に，単純な統計的数値化だけでどこまで単語認定ができるか，という一つの極端な試みを行った。以下にその詳細を述べる。

なお，単語認定という以上，「語」とは何か，どのようなレベルで「語」というものをとらえるのか，という問いに答えなければならないところだが，本稿では最も素朴な「単語」観として，「辞書の見出し語となるような単位」を単語の典型イメージとしておくにとどめる。

2.「中古和文 UniDic」とダイス係数によるコロケーション強度の測定

2.1 形態素解析辞書「中古和文 UniDic」

コロケーション強度を測定するにあたっては，コーパスから語と語のまとまりを抽出しなければならないが，そのためには品詞情報付きのコーパスが必要である。文字列のみのテキストから名詞と形容詞の組み合わせを取り出す，というのはきわめて困難だが，形態素解析を経たテキストを用いればそのような作業も容易となる。形態素解析とは，テキストデータを自動で単語に区切り，図 1 のように読み・品詞・活用形などの情報を付けて出力する自然言語処理技術である。

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形
いづれ	イズレ	イズレ	何れ	代名詞			イズレ
の	ノ	ノ	の	助詞-格助詞			ノ
御	オオン	オオン	御	接頭辞			オオン
時	トキ	トキ	時	名詞-普通名詞-副詞可能			トキ
に	ニ	ナリ	なり	助動詞	文語助動詞-ナリ		ニ
か	カ	カ	か	助詞-係助詞			カ
、			、	補助記号-読点			
女御	ニョーゴ	ニョウゴ	女御	名詞-普通名詞-一般			ニョウゴ
、			、	補助記号-読点			
更衣	コーイ	コウイ	更衣	名詞-普通名詞-サ変可能			コウイ
、			、	補助記号-読点			
あまた	アマタ	アマタ	数多	副詞			アマタ
さぶらひ	サブライ	サブラウ	侍ふ	動詞-一般	文語四段-ハ行	連用形-一般	サブラウ
たまひ	タマイ	タマウ	給う-尊敬	動詞-非自立可能	文語四段-ハ行	連用形-一般	タマウ
ける	ケル	ケリ	けり	助動詞	文語助動詞-ケリ	連体形-一般	ケリ
中	ナカ	ナカ	中	名詞-普通名詞-副詞可能			ナカ
に	ニ	ニ	に	助詞-格助詞			ニ
、			、	補助記号-読点			

図 1　形態素解析のイメージ

この形態素解析により，通常のテキストデータベースではできないような検索が可能になる。「品詞」や「活用形」というレベルで条件を指定しての検索は従来の文字列検索では不可能だったが，このように品詞情報が付与されたコーパスであれば，例えば［動詞（連用形）＋動詞（活用形問わず)］といった検索条件で，複数のテキストにまたがってすべての用例を容易に採取することが可能になるのである。

今回は古典資料専用の形態素解析辞書として開発された「中古和文 UniDic」を利用した。従来の形態素解析辞書は現代語を対象としており，古典資料の解析には無力であったが，「中古和文 UniDic」により古典資料の高精度な解析が実現した（小木曽・小椋・田中・近藤・伝 2010，小木曽・小椋・近藤・須永 2010）ため，品詞情報付きの中古語コーパスの作成もようやく可能となった。

実例調査のコーパスとしては，「中古和文 UniDic」の開発時に作成された学習用コーパスを利用した。使用したコーパスは総語数約 80 万，収録作品は表 1 のとおりである。

表 1　使用コーパス（学習用データ）収録作品

作品名	語数
伊勢物語	14624
大和物語	26462
土佐日記	7948
紫式部日記	20348
更級日記	16656
源氏物語	536798
竹取物語	12136
古今和歌集仮名序	3106
枕草子	78343
大鏡	82829
（計）	799250

※語数は「中古和文 UniDic」の言語単位「短単位」（小椋・小磯・冨士池・宮内・原（2011））による。また，句読点も語数に含まれる。

2.2 コーパスを用いた実例調査と『日本国語大辞典』の見出し語比較

今回は，上で語認定が難しいとしたものの中で，特に「名詞＋形容詞または動詞」の組み合わせ（例：「甲斐なし」「興あり」）を取り上げて検証することとした。［主語－述語］の関係に見える二語の組み合わせ，ということになるが，［述語］すなわち，二語目にあたる用言の方は「なし」「あり」「よし」「あし」の 4 種類を選んで調査した。

この 4 種類の用例数は次のようになった。

表 2　語数

コーパス総語数：799250

	のべ	異なり
名詞＋あり	1690	292
名詞＋なし	2581	272
名詞＋よし	120	31
名詞＋あし	86	20

既存の辞書ではどのような組み合わせを一語と認定しているのか参考とするために，『日本国語大辞典』（以下，『日国』）[1] の立項状況を示したものが表 3 である。

表 3　『日本国語大辞典』立項状況と，コーパス上の実例

	『日国』見出し語数	コーパスに出現した異なり語数（うち，『日国』立項）
名詞＋あり	17	292（4）
名詞＋なし	279	272（83）
名詞＋よし	16	31（12）
名詞＋あし	17	20（8）

この表 3 を見ると，「名詞＋なし」の『日国』見出し語数が他の三者とはけた違いに多いことにまず目が行くが，コーパスに出現した異なり語数を見ると，「名詞＋なし」は用例数自体が多く，見出し語数が多いことも首肯される。むしろ問題となるのは，「名詞＋あり」のグループの立項数が異常に低いことである。「名詞＋なし」は『日国』に親見出し・子見出し含め 279 種類，コーパス上は 272 種類出現し，うち 83 種が実際『日国』の見出し語に採られている。これに対し「名詞＋あり」では，コーパス上出現したのは 292 種類と，「名詞＋なし」と遜色ないにもかかわらず，『日国』の見出し語はわずか 17，「名詞＋なし」の 279 と比べると大きな違いがみられる。

この「名詞＋なし」と「名詞＋あり」は，意味の上では非存在と存在という対称をなしているにもかかわらず，『日国』の立項状況には大きな開きがあるということになる。同じ名詞が「なし」「あり」両方に接し，「○○なし」「○○あり」で対称的な意味を持つペアを作るケースも多い。

[1]『日本国語大辞典』の見出し語調査には，「日国オンライン」（『日本国語大辞典　第二版』のデータ）を利用した。よって，本稿での『日国』は第二版としての扱いとなる。

例えば前出の「甲斐なし」に対し「甲斐あり」という用例もコーパス上多数見つかっている。そのような「名詞＋あり／なし」のペアのうち，『日国』ではほとんどの場合で「名詞＋なし」のみが見出し語として立項され，「名詞＋あり」は立項されない。

表 4 「名詞＋あり／なし」のうち，『日本国語大辞典』では「〜なし」のみが立項されているもの

あいぎょう(愛敬)	すじ(筋)	はえ(映)
あと(跡)	そこい(底方)	びん(便)
いつわり(偽)	たえま(絶間)	へだて(隔)
いとま(暇)	たぐい(類)	ほど(程)
えき(益)	たのみ(頼)	ほい(本意)
おぼえ(覚)	たより(便)	まぎれ(紛)
おもい(思)	ちから(力)	もの(物)
かい(甲斐)	ついで(序)	ゆえ(故)
かぎり(限)	とき(時)	ゆるし(許)
かくれ(隠)	なごり(名残)	よう(様)
くま(隈)	なさけ(情)	よし(由)
ここち(心地)	なにごころ(何心)	よりどころ(拠所)
ことわり(理)	に(二)	
さだめ(定)	のこり(残)	

立項されているものについては，「○○なし」で一語，されていないものは「○○」と「あり」とにわかれる二語，という扱いなのだろうが，このように，「名詞＋なし」と「名詞＋あり」の扱いには一見不平等にしか見えない差がある。これは単に立項漏れであるのか，何か確たる根拠があるのか。現状では，『日国』の単語認定が妥当なものであるのかどうかすら，よくわからない。

2.3 コロケーション強度を測る指標〜ダイス係数〜

「名詞＋あり／なし／よし／あし」の実例数を表 2 に掲げたが，当然のことながら，コーパス上に出現した用例のすべてを一語と認定するわけにはいかない。今回の調査では，名詞に「あり／なし／よし／あし」のいずれかが後続する，という条件を満たしたものがすべて機械的に拾い出されるので，その中には感覚的にも一語とは認めがたく，出現頻度も低いものもいくらでも含まれる（「上達部なし」「傘なし」「法師あり」など）。では，一語と認定してもよさそうな組み合わせをうまく洗い出すことのできる指標とはどのようなものか。

まずまっさきに思い浮かぶのは，「用例数の多いものを見出し語と認定する」というものであるが，これでは不十分である。なぜならば用例数は，その組み合わせの構成要素である，もとの語それぞれの語数の違いに影響されてしまうからである。仮に「名詞 A ＋なし」が 40 例と「名詞 B ＋なし」が 20 例出現したとする。これだけを比べると，「A なし」の方が「B なし」より目立ち，見出し語立項にふさわしそうに見えるが，より正確に判断するには，そのもとになる名詞 A，B が，当該コーパスにおいて，そもそもどの程度現れているのかを考慮に入れなければならない。名詞 A は全体で 10000 例ほどあり，その中でわずかに 40 例だけが「なし」と結びつい

ているのに対し，名詞 B の方はそもそも 25 例しかなく，その中の 20 例が「なし」と結びついているとしたら,「B なし」の方が「A なし」より特別な結びつきだと言うべきだろう。実際「名詞＋あり」という条件を満たすものを取り出してみると，「人あり」「事あり」「物あり」などが上位に来るが，これは明らかに「人」「事」「物」という名詞自体の多さに影響されている。このようなことがあるため，結果的な用例数だけに惑わされることなく，名詞 A と名詞 B それぞれの全用例の中で，該当の組み合わせ「A なし」「B なし」がどの程度の割合を占めているかが，結びつきの強さ（コロケーション強度）を考える上では重要となる。一口に「頻度の高さ」と言っても，ただ用例数を数え上げるのではなく，「もとになる語自体の頻度までを考慮した上での頻度の高さ」を比較する必要があるのであり, そのために統計的指標を用いた数値化が必要となる。

コロケーション強度を測定する統計的指標はいくつかあるが，以下の調査では単純で計算しやすい「ダイス係数」を用いた。ダイス係数の計算式は以下のとおりである。

$$\mathrm{D} = 2 \times \frac{\text{「XY」の語数}}{\text{X の語数} + \text{Y の語数}}$$

式を見てのとおり，ダイス係数は，その組み合わせのもとになる語の語数（分母）と，実際に組み合わさって現れた語の語数（分子× 2，もとの 2 語が合わさったら 1 語になってしまうので，比較できるように 2 倍する）を比較することで，コロケーション強度を算出する，という指標であり，1 から 0 の間の値をとる。仮に語 X と語 Y が，常に「XY」という形でしか用いられないとしたら，X の語数，Y の語数,「XY」の語数はすべて同数となり，結果，D = 1 となる。また，X と Y が組み合わさることは一切ない, という場合,「XY」の語数は 0 のため, D = 0 となる。よってダイス係数は基本的に 1 から 0 の間の小数の値となる。ダイス係数の特長はなにより単純なことである。統計指標は，コロケーション強度測定によく用いられるものに限っても,「t スコア」「MI スコア」など，他にもさまざま挙げられる（石川 2008）が，ダイス係数は特に単純な部類だと言える。参考までに，他によく用いられる指標,「t スコア」,「MI スコア」について簡単に紹介しておく。

t スコア，MI スコアはともに,「XY」という組み合わせが出現する確率上の期待値と，実際の出現語数（実測値）を比較する指標で，t スコアは実測値と期待値の差を，MI スコアは実測値と期待値の比をもとに数値化する。

$$\text{t スコア：}\quad t = \frac{(\text{実測値} - \text{期待値})}{\sqrt{\text{実測値}}} \qquad \text{MI スコア：}\quad I = \log_2 \frac{(\text{実測値})}{(\text{期待値})}$$

「XY」という組み合わせの,

期待値 = コーパス総語数× X が現れる確率× Y が現れる確率
= コーパス総語数×（X の語数／コーパス総語数）×（Y の語数／コーパス総語数）
= X の語数× Y の語数／コーパス総語数

これらの指標と比べても，ダイス係数は単純で計算しやすく，かつ，統計や数式にあまりなじみがなくても，感覚的に理解しやすい指標であることがわかるだろう。ダイス係数は，単純であるがゆえに欠点もいくつかある。まず，t スコアや MI スコアと異なり，コーパスの総語数を問題にしないので，異なる語数のコーパス間を比較する場合には適さない。また，式を見てのとおり，X の語数と Y の語数を足した数と，組み合わさった「XY」の語数（の 2 倍）を比較するだけなので，Y は X の他にもさまざまな語と自由に結びつくのだが，X の側には必ず Y と結びつくという特性がある，というような事情があったとしても，そのような特性は数値に反映されない。例えば，（ア）X10 語，Y990 語（＝X＋Y＝1000）で，X10 語はすべて Y と一緒にしか用いられず，結果「XY」が 10 例得られる場合と，（イ）X500 語，Y500 語（＝X＋Y＝1000）で，「XY」がたまたま 10 例得られた場合とでは，X の性質の違いに注目したくなるが，ダイス係数上の数値では，両者はまったく同じ値にしかならない。

しかしダイス係数は，このように単純な割に，実用面でも比較的有用な指標として知られており，他の複雑な指標よりもかえって良好な結果を示すことも多い。実際，今回の調査範囲のうち，実例数も豊富で，『日国』見出し語数も多い「名詞＋なし」の場合を例に，ダイス係数，t スコア，MI スコアの 3 つの指標をもとにコロケーション強度を算出し，比較してみたところ，辞書立項の有無との相関を最も強く示したのはダイス係数であった。

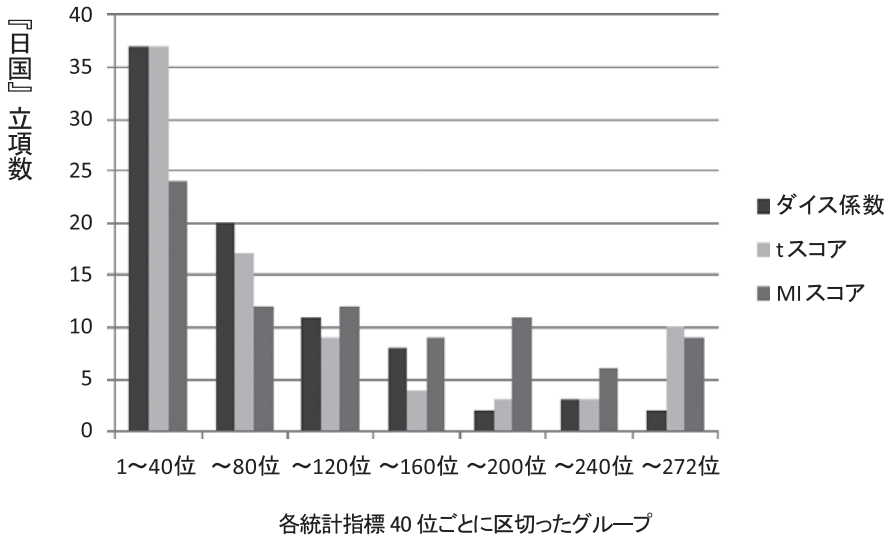

図 2 「名詞＋なし」の統計指標による順位付け（横軸）と，『日本国語大辞典』立項状況（縦軸）

このグラフは，「名詞＋なし」の組み合わせ 272 種を，各統計指標での数値が高い順に並べ，1 位から 40 位まで，41 位から 80 位まで……と 40 語ごとに区切り，『日国』の見出しとの対応関係を調査したものである。縦軸が，各グループごとに 40 語中何語が見出しになっているかを示している。グラフは左からダイス係数，t スコア，MI スコアの順で，いずれの指標でも上位に来る語は辞書見出しになりやすい，という相関がみられるが，特にダイス係数は，上位語ほど多

く辞書に採られ，下位になるにつれ辞書に採られなくなっていく，というきれいな相関をみせた。そこで，手始めとして使いやすい簡単な指標であることに加え，図 2 に示したような実験結果も踏まえ，本稿ではコロケーション強度を算出する指標にはダイス係数を用いることとした。

3.「名詞＋あり／なし／よし／あし」のコロケーション強度調査結果

3.1「名詞＋あり／なし／よし／あし」のダイス係数上位語一覧

「中古和文 UniDic」を用いて，「名詞＋あり／なし／よし／あし」の各用例を洗い出し，ダイス係数順に並べ替えたものが下の表 5 から表 8 である。紙幅の都合上，ダイス係数 0.003 以上のものだけを表出している。なお，参考のために，『日国』に見出し項目として立項されているか否かも表中右端に示した。白地の項目が立項されているもの，黒地が立項されていないものである。

表 5 「名詞＋あり」ダイス係数上位語（ダイス係数 0.003 以上）

順位	名詞	ダイス係数	A＋Bの語数	A：名詞の語数	B：「あり」の語数	『日国』
1	けしき（気色）	0.01023	40	918	6904	子見出し
2	きょう（興）	0.00751	55	7742	6904	項目なし
3	みどころ（見所）	0.00710	32	2104	6904	項目なし
4	たぐい（類）	0.00534	26	2834	6904	項目なし
5	かい（甲斐）	0.00479	81	26923	6904	項目なし
6	きこえ（聞）	0.00477	23	2737	6904	項目なし
7	よし（由）	0.00465	55	16757	6904	子見出し
8	ゆえ（故）	0.00452	38	9900	6904	項目なし
9	おおせごと（仰言）	0.00432	18	1422	6904	項目なし
10	こころばせ（心馳）	0.00411	16	880	6904	項目なし
11	しるし（験）	0.00398	18	2142	6904	項目なし
12	めし（召）	0.00386	14	350	6904	項目なし
13	しょうそこ（消息）	0.00385	21	4011	6904	項目なし
14	おりおり（折々）	0.00373	19	3294	6904	項目なし
15	かぎり（限）	0.00340	89	45433	6904	項目なし
16	へだて（隔）	0.00333	14	1494	6904	項目なし
17	こころばえ（心延）	0.00331	20	5185	6904	項目なし
18	ろう（労）	0.00329	12	387	6904	項目なし
19	たより（便）	0.00310	13	1488	6904	項目なし
20	ざえ（才）	0.00303	12	1008	6904	項目なし

表 6 「名詞＋なし」ダイス係数上位語（ダイス係数 0.003 以上）

順位	名詞	ダイス係数	A＋Bの語数	A：名詞の語数	B：「なし」の語数	『日国』
1	かい（甲斐）	0.09935	183	295	3389	立項
2	限り（限）	0.08090	237	2470	3389	立項
3	あじき（味気）	0.06177	108	108	3389	立項
4	なさけ（情）	0.03877	68	119	3389	立項
5	たぐい（類）	0.03763	67	172	3389	立項
6	くま（隈）	0.03110	54	84	3389	立項
7	なごり（名残）	0.02432	43	147	3389	立項
8	つき（付）	0.02333	40	40	3389	立項
9	に（二）	0.02295	43	358	3389	立項
10	ほい（本意）	0.02219	39	126	3389	立項
11	ひま（暇）	0.02067	36	94	3389	項目なし
12	うら（心）	0.02041	35	40	3389	立項
13	いとま（暇）	0.01964	34	73	3389	立項
14	さだめ（定）	0.01906	33	73	3389	立項
15	よし（由）	0.01783	33	313	3389	立項
16	ほど（程）	0.01670	49	2478	3389	立項
17	びん（便）	0.01565	71	5682	3389	立項
18	へだて（隔）	0.01497	26	84	3389	子見出し
19	ならび（並）	0.01459	25	38	3389	立項
20	あや（文）	0.01406	24	25	3389	立項
21	しずこころ（静心）	0.01370	30	991	3389	立項
22	のこり（残）	0.01322	23	90	3389	立項
23	えき（益）	0.01288	22	27	3389	立項
24	かくれ（隠）	0.01277	22	56	3389	立項
25	たゆみ（弛）	0.01173	20	21	3389	立項
26	おぼえ（覚）	0.01110	20	215	3389	立項
27	ゆるし（許）	0.01049	18	44	3389	立項
28	すべ（術）	0.00938	16	23	3389	立項
29	くもり（曇）	0.00880	15	19	3389	立項
30	なにごころ（何心）	0.00694	17	1512	3389	立項
31	ここち（心地）	0.00687	15	977	3389	立項
32	おも（面）	0.00647	11	11	3389	立項
33	ゆえ（故）	0.00621	12	477	3389	立項
34	ひとげ（人気）	0.00615	11	187	3389	立項
35	こころ（心）	0.00593	20	3359	3389	立項
36	おもいやり（思遣）	0.00584	10	36	3389	項目なし
37	うしろみ（後見）	0.00547	16	2464	3389	項目なし
38	あいぎょう（愛敬）	0.00546	12	1008	3389	立項
39	よう（要）	0.00530	9	9	3389	立項
40	おう（奥）	0.00518	9	84	3389	立項
41	たずき	0.00512	9	128	3389	立項
42	ゆくえ（行方）	0.00510	10	530	3389	子見出し
43	おきどころ（置所）	0.00509	10	538	3389	項目なし
44	まぎれ（紛）	0.00474	10	830	3389	立項
45	ろん（論）	0.00471	8	10	3389	立項
46	よう い（用意）	0.00423	10	1338	3389	項目なし
47	けしき（気色）	0.00418	9	918	3389	項目なし
48	つみ（罪）	0.00400	12	2609	3389	項目なし
49	うたがい（疑）	0.00392	7	182	3389	子見出し
50	ついで（序）	0.00390	16	4823	3389	子見出し
51	よりどころ（拠所）	0.00349	6	48	3389	立項
52	つね（常）	0.00321	34	17782	3389	立項
53	ものおもい（物思）	0.00306	6	537	3389	項目なし

「名詞＋よし」についてはのべ語数 120，異なり語数 31，「名詞＋あし」はのべ語数 86，異なり語数 20 と，そもそも得られた数自体が少なかったが，これらについてもダイス係数上位語を参考までに挙げる。

表 7 「名詞＋よし」ダイス係数上位語（ダイス係数 0.003 以上）

順位	名詞	ダイス係数	A＋Bの語数	A：名詞の語数	B：「よし」の語数	『日国』
1	ここち（心地）	0.04743	35	977	499	立項
2	さま（様）	0.01990	26	2114	499	立項
3	こと（言）	0.01675	5	98	499	立項
4	かおかたち（顔貌）	0.00768	2	22	499	項目なし
5	なからい（仲合）	0.00693	2	78	499	項目なし
6	なか（仲）	0.00542	3	608	499	子見出し
7	いろ（色）	0.00457	2	377	499	立項
8	たけだち（丈立）	0.00394	1	9	499	項目なし
9	き（気）	0.00393	1	10	499	項目なし
10	ひき（引）	0.00388	1	16	499	項目なし
11	まつりごと（政）	0.00382	1	24	499	項目なし
12	かざし（挿頭）	0.00380	1	27	499	項目なし
13	たけ（丈）	0.00380	1	27	499	項目なし
14	こえ（声）	0.00378	5	2145	499	子見出し
15	ね（根）	0.00377	1	31	499	立項
16	かたち（形）	0.00373	12	5936	499	子見出し
17	いろあい（色合）	0.00373	1	37	499	項目なし
18	はかま（袴）	0.00367	1	46	499	項目なし
19	ひ（火）	0.00332	1	104	499	項目なし
20	かお（顔）	0.00305	2	812	499	子見出し

表 8 「名詞＋あし」ダイス係数上位語（ダイス係数 0.003 以上）

順位	名詞	ダイス係数	A＋Bの語数	A：名詞の語数	B：「あし」の語数	『日国』
1	ここち（心地）	0.02884	17	977	202	子見出し
2	さま（様）	0.02579	26	1814	202	立項
3	けしき（気色）	0.01964	11	918	202	子見出し
4	なり（形）	0.01818	2	18	202	項目なし
5	おり（折）	0.01408	6	650	202	立項
6	なか（仲）	0.00978	4	616	202	子見出し
7	きたかぜ（北風）	0.00976	1	3	202	項目なし
8	みだりごこち（乱心地）	0.00893	1	22	202	項目なし
9	おもて（表）	0.00862	1	30	202	項目なし
10	かみ（髪）	0.00837	2	276	202	項目なし
11	あし（足）	0.00755	1	63	202	項目なし
12	ひとめ（人目）	0.00658	1	102	202	項目なし
13	ため（為）	0.00638	3	738	202	項目なし
14	はら（腹）	0.00628	3	754	202	子見出し
15	こよい（今宵）	0.00556	1	158	202	項目なし
16	きょう（今日）	0.00324	1	416	202	項目なし

3.2 ダイス係数と『日本国語大辞典』見出し語立項の比較

ダイス係数の指標と『日国』での見出し語立項基準とは相関していると言えるのだろうか。

表 9　ダイス係数と『日本国語大辞典』見出し語の関係

	ダイス係数			
	0.003 以上	0.004 以上	0.006 以上	0.009 以上
名詞＋なし	53(45)	48(41)	34(33)	28(27)
名詞＋あり	20（2）	10（2）	3（1）	1（1）
名詞＋よし	20（9）	7（5）	5（3）	3（3）
名詞＋あし	16（6）	15（6）	12（5）	7（5）

この表 9 から，ダイス係数 0.009 で区切ると，それ以上の数値を有する組み合わせについては「なし／あり／よし／あし」四者のすべてで，そのほとんどが『日国』で立項されていることがわかる。すなわちダイス係数 0.009 以上であることと，辞書の見出し語になっていることはおおよそ一致していると言える。意味を一切加味せず，語数と用例数のみで算出した純粋に客観的な指標が，主観的・経験的な語認定にある程度沿う結果となっている。

図 2 で示したとおり，ダイス係数と辞書立項には高い相関がみられる（図 2 のうち，ダイス係数のみを図 3 として再掲）。

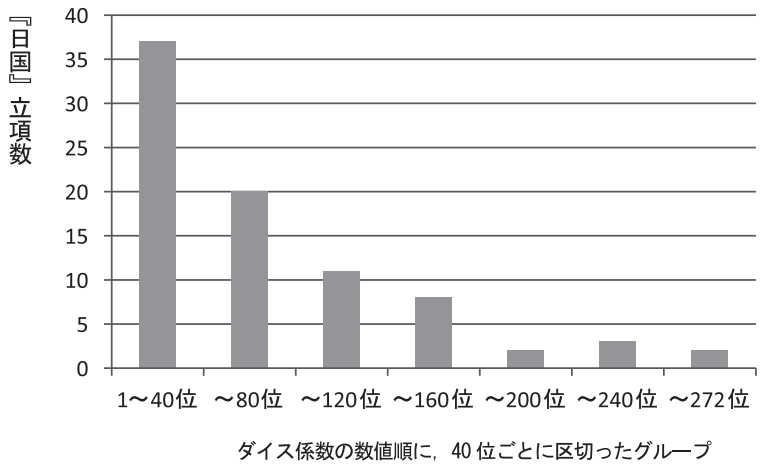

図 3 「名詞＋なし」のダイス係数による順位付け（横軸）と，『日本国語大辞典』立項状況（縦軸）

図 3 をみると，最初のグループ，1 位から 40 位までの 40 語では，実に 37 語が見出し語になっている。ダイス係数の値が高いことと，『日国』の見出し語になっていることとがかなり対応しており，ダイス係数が語認定の指標として有効であることを示している。

2.2 で取り上げた，「名詞＋なし」と「名詞＋あり」で立項状況が大きく異なる，という点につ

いてはどうだろうか。ダイス係数を見てみよう。各グループのダイス係数の値に注目して見ていくと，実は「名詞＋なし」のグループのダイス係数は他を圧倒する高い値であることがわかる。そもそもダイス係数 0.003 以上の用例は，「名詞＋なし」では 53 例も挙げられるのに対し，「名詞＋あり」「名詞＋よし」では各 20 例，「名詞＋あし」にいたっては 16 例しかない。また一口に 0.003 以上と言っても，「なし」とそれ以外では数値に大きな差が出ている。例えば「名詞＋あり」で 1 位になる「けしきあり」のダイス係数は 0.01 だが，この数値は「名詞＋なし」の側では 26 位と 27 位の間に位置する程度の値である。そもそも 0.01 以上をマークしている用例は「あり」「よし」「あし」の三者においてはそれぞれほんの数例ずつしかないが，「なし」の側では 27 位までが 0.01 を超えており，数値上の差は歴然である。

さらに表 4 で，同じ名詞でも「なし」の側だけ見出し語認定され，「あり」の方は見出し語にならない，という例を紹介したが，それらの主なもののダイス係数を比較したのが表 10 である。

表 10 「名詞＋なし／あり」のダイス係数，『日本国語大辞典』立項状況の比較

名詞＋なし／あり		ダイス係数	『日本国語大辞典』
かい（甲斐）	なし	0.09935	立項
	あり	0.00479	項目なし
たぐい（類）	なし	0.03763	立項
	あり	0.00534	項目なし
くま（隈）	なし	0.03110	立項
	あり	0.00167	項目なし
ゆるし（許）	なし	0.01049	立項
	あり	0.00086	項目なし
へだて（隔）	なし	0.01497	子見出し
	あり	0.00333	項目なし
なにごころ（何心）	なし	0.00694	立項
	あり	0.00057	項目なし
ゆえ（故）	なし	0.00621	立項
	あり	0.00452	項目なし

「名詞＋あり」のダイス係数が「名詞＋なし」を上回っているものは一つもなく，それどころか数値が 1 桁違うものも多いことがわかる。数値の差があまりないのは「故なし／あり」程度である。このように数値化してみると，「名詞＋なし」と「名詞＋あり」には大きな違いがあり，『日国』の扱いの差はそれなりに妥当であると言えそうである。

以上を見渡すと，ダイス係数の上位語と辞書の見出し語がある程度対応しており，「名詞＋なし」が立項されているのに対して，同じ名詞でも「名詞＋あり」は立項されないことが多い，ということに対しても，数値上からもある程度の納得がいく。しかし一方で，ダイス係数が高く，辞書見出しに入っていてもよさそうなものが漏れていること，「名詞＋なし」に対し，「名詞＋あり」の項目のほうが少ないのがある程度妥当であるとしても，「名詞＋あり」の項目数が少なすぎるような印象を受けること，など，既存の辞書の問題点も見えてくる。

そのような問題点を補うものとして，ここまでみてきた統計的指標による数値化こそが単語認定の客観的指標として有望だと思われる。

4. ダイス係数を見出し語認定基準に利用する可能性

上述のとおり，ダイス係数の上位語は辞書の見出し語とほぼ対応している。よって，従来の内省的，主観的な「一語」の感覚に沿うものを，統計上の数値という客観的な基準である程度取り出すことができるといえる。この指標を用いることによって，中古語の単語認定の客観的基準を設けたり，既存の辞書の立項漏れを洗い出すことが可能になる。

ある語の組み合わせを辞書項目として取り上げるか否かを判断する際の根拠は，もちろん一つではない。出現頻度の高い組み合わせだから，というのも根拠の一つではあるが，仮に出現頻度は低くても，その組み合わせになると特別な意味になる，というような，意味の面から立項が求められるものもある。例えば「腹悪し」という組み合わせは，ただ「腹が悪い」という意味ではなく，「怒りっぽい」という意味になる。「腹悪し」のダイス係数は 0.006，数値の上でもそれなりの高さを出しているが，このような場合はたとえ出現頻度が低くても，意味の面での変化を重視して辞書見出しに採られるべきであろう。今回の範囲では，出現頻度を基にした外形的な結びつきの強さを根拠に単語認定のありかたを検討したが，それだけでは，ここで述べたような，意味の面から立項が求められる語をすべてすくい上げることは期待できない。よって，今回示したような数値による線引きで，辞書見出しとすべきか否かを一律に判断できるとはいえない。だが少なくとも，頻度の高い，結びつきの強い組み合わせは見出しとしておく，また，この辺りまでは一語としていくという基本的なグループを洗い出す，というような外形面からの要請としては，ダイス係数などの統計的指標は，一語化しているかの認定や，既存の辞書の漏れを洗い出す手掛かりとして有効であると考えている。

以上を踏まえ，辞書編纂の現場においてはダイス係数を最終的な立項基準にするわけではないが，立項候補を洗い出す基準線として用いること自体は効果的ではないかと考える。今回の調査範囲であれば，少し緩めに，ダイス係数 0.004 ぐらいを基準とし，基準値を超えた用例はすべて見出し項目の候補とした上で，最終的な選別は人の経験と判断にゆだねる，というやり方である。

数値だけを絶対的な基準にできないという事情には，ダイス係数は意味の面をすくえない，という点以外にも，資料自体が限られる古典語の場合には，統計上妥当と言えるほどの用例数を確保することが難しく，現代語ほどには信頼できる数値が得られないという問題点もある。このような弱点からしても，最終的には人間による選考はやむないと考えられる。しかし一方で，ダイス係数を導入することで，人間の主観的な選定だけでは見落としがちな用例や，量的に重要な用例を確実にすくい上げることができる。この両方の手段を合わせることで，より確実な編纂につなげることができるのではないだろうか。

なお，ダイス係数 0.004 以上のものを単語の認定候補とすると，今回使用したコーパス内では『日国』の見出し語となっていない以下の 26 例が新たに単語の認定候補として洗い出される。

表 11　ダイス係数 0.004 以上で『日本国語大辞典』に立項されていない語

なし	ひまなし、思いやりなし、後ろ見なし、置きどころなし、用意なし、気色なし、罪なし
あり	興あり、見どころあり、類いあり、甲斐あり、聞こえあり、仰せ言あり、心ばせあり
よし	かおかたちよし、仲らいよし
あし	形(なり)あし、為あし、髪あし
	乱り心地あし、人目あし、おもてあし、今宵あし、北風あし、足あし

※白抜きは用例数 1 例のみ。

「名詞＋よし／あし」の側は，さほど語数が取り出せず，出現頻度が低いものが多く混じっているため，あくまで参考としたい。例えば「北風あし」などは，1 例のみで，名詞「北風」自体が 3 例しかなく，3 例中 1 例が「あし」と結びつく，ということから高い数値が出てしまっている。このようなことがあるため，今後は出現頻度自体が低いものは何らかの基準を設けて排除することが望ましいと考えるが，現代語と異なり，古典語の場合，資料自体が限られるため，統計的に十分といえるほどの例数がそもそも取り出せないことも多い。

5. めざすべき単語認定のあり方と，統計的指標・形態素解析

本稿では，意味的側面など，目に見えないものをあえて無視し，統計的指標のみによる単語認定のあり方を試みた。しかし，これはいわば一つの実験であって，従来の語認定法にかえて，統計的指標のみを語認定の基準にしようと主張するものではない。「名詞＋あり／なし／よし／あし」という組み合わせを，それ以上何の条件をかけることもなく取り出し，単純な統計指標を用いて数値化するだけ，というのは，本来単語認定をするときに考えなければならないことからすると，極めて乱暴な，穴だらけの操作といえよう。しかし，そのような粗い操作でも，従来の「感覚」に沿うものを，外形的・客観的に取り出すことができたということ自体は大きな収穫であったと考える。

意味の面を排除するとしても，一語化しているかを認定する外形的な要因はいくらでもありうる。もともと二語だったものが一語化するということは，語としての性質が変わっているということであり，意味的な面の変化はもちろん，構文的にも変化がみられるはずである。例えば，以下のような場合があげられる。

まず，連体句を受ける場合。「隈なし」は『日国』にも立項されているが，下例（5）のように連体句を受ける場合に限っては，「そのことぞとおぼゆる隈」―「なし」というつながりであると感じられ，「隈なし」で一語とは認めがたい。

(5)　そのことぞとおぼゆる隈なく，愛敬づきなつかしくをかしげなり。（源氏物語・浮舟）

逆に，「名詞＋あり／なし／よし／あし」が「いと」などの副詞によって修飾されている場合，

その組み合わせは一語化しているという感覚が強まる。

(6)　玉淵はいとらうありて，歌などよくよみき。（大和物語）

また，「名詞＋あり／なし／よし／あし」という連続ではなく，間に助詞が入って分断される，以下のような場合も考えられる。

(7)　まどひ歩き給へどかひもなし。（大和物語）
(8)　「名残だになくあさましきこと」と，……（源氏物語・夕霧）

(7)(8)では「甲斐なし」,「名残なし」が，それぞれ係助詞「も」，副助詞「だに」によって分断され，「甲斐もなし」「名残だになし」となっている。このように助詞によって分断される場合がある（，またはその場合が多い）という事実は，その組み合わせが一語として固まりきってはいない，ということの一つの証左となると考えられる。

意味に入り込まず，外形的な特徴に限ってなお，一語化しているかを認定する手がかりはこのようにさまざまにありうる。本来であれば，これらの特徴を洗い出して検討するという作業の方こそが単語認定の操作としては本筋であり，統計指標はあくまで補助，という位置づけだと考えている。

しかし，本節で示したような構文的特徴を基準に調査するというのも，形態素解析により品詞情報が付与されたコーパスがあればこそである。品詞情報が付与されたコーパスがあってはじめて，特定の語と語の連接をセットとして取り出すことが可能になる。このような調査は，平安仮名文学作品を高精度で解析できる「中古和文 UniDic」の開発でようやく現実味を帯びたものであり，経験や感覚に頼るだけの従来の単語認定から，より合理的，客観的な単語認定に向けての第一歩を踏み出したところである。

形態素解析辞書や統計指標を用い，中古語と向き合っていくという試みは始まったばかりであり，例えば本節で示したような構文的特徴にしても，それを何パターン用意し，それら個々の特徴にどの程度のウェイトを置くかなどはまだわからない。また，古典語は資料の数が限られるため，外形的特徴を細かく制限していくと，語自体の性質のためではなく，資料数の少なさから，用例が取り出せない可能性もある。そのようなとき，まず手始めとして，今回のようなごく簡単な操作で，検討候補が洗い出せることが確認できた，ということの意味は大きい。

今回試みた簡単な操作でも，一語化している主なものを洗い出せること，辞書編纂において立項漏れをチェックすることなどに有効であることが確認できた。今回の手法によって洗い出された用例を見渡し，一語化している／していないといえる外形的特徴を探し出し，さらにそれをフィードバックして検索条件を細かくしていく，という流れも考えられる。今回の調査を手始めに，より合理的な単語認定のあり方を探っていきたい。

参　照　文　献

石川慎一郎（2008）「コロケーションの強度をどう測るか―ダイス係数, t スコア, 相互情報量を中心として―」『言語処理学会第 14 回大会チュートリアル資料』40–50.
小木曽智信・小椋秀樹・近藤明日子・須永哲矢（2010）「形態素解析辞書「中古和文 UniDic」とその活用例」『日本語学会 2010 年度秋季大会予稿集』243–248.
小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」『人文科学とコンピュータ』（情報処理学会研究報告）Vol.2010/CH-85: 1–8.
小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版』特定領域研究「日本語コーパス」平成 22 年度研究成果報告書．国立国語研究所.
須永哲矢・小木曽智信（2011）「コーパスとコロケーション強度を用いた中古語の語認定」『日本語学会 2011 年度春季大会予稿集』275–280.

関連 Web サイト

「中古和文 UniDic」「近代文語 UniDic」http://www2.ninjal.ac.jp/lrc/
「日国オンライン」（ジャパンナレッジプラス）http://www.jkn21.com

Word Identification in Early Middle Japanese Using Collocation Strength

SUNAGA Tetsuya

Adjunct Researcher, Center for Corpus Development,
National Institute for Japanese Language and Linguistics

Abstract

It has long been a serious problem for researchers of Early Middle Japanese to determine whether a set phrase like *kai-nashi* should be classified as one word or a combination of separate words. There is no definite criterion, and some phrases are listed in dictionaries as a word while others are neglected, all depending on the judgment of the editor. In this paper, the Dice coefficient is introduced as a solution. The Dice coefficient is an index for estimating collocation strength, i.e., how strongly two words are connected with each other. In combination with a morphological analysis dictionary (*Chuko-Wabun UniDic*), the Dice coefficient works as one criterion for word identification.

Key words: morphological analysis dictionary, Early Middle Japanese, collocation strength, Dice coefficient