ID	KK290801
種別	論文
タイトル	初級文法項目の生産性の可視化 —動詞に接続する文法項目の場合—
Title	Visualization of the Productivity of Function Words in Basic Japanese Grammar Syllabus
著者	中俣　尚己
Author	NAKAMATA Naoki
掲載号	29巻8号
発行日	2015年3月19日
開始ページ	275
終了ページ	295
著作権者	計量国語学会

計量国語学 29 巻 8 号（2015 年 3 月）pp.275-295.

論文

初級文法項目の生産性の可視化

―動詞に接続する文法項目の場合―

中俣　尚己（京都教育大学教育学部）

要旨

この論文では，日本語の初級文法項目の中でも多くの動詞と接続して使われるものと，限られた動詞とのみ接続して使われるものとがあることに注目し，その度合いを「生産性指数」として可視化することを目的とする．7 つの指標を候補として立て，BCCWJ から 103 項目を検索したデータを対象にそれぞれの指標を計算して比較するという手順を踏んだ結果，Guiraud Index の式を流用したものが最もよい結果を出すことがわかった．このようにして可視化された生産性のランキング表からは高生産性語は時を表す項目に偏り，また，話し言葉的な項目の方が生産性が低いという傾向が一貫して見られるなどの特徴が見出される．生産性の大小がわかれば，その項目を色々な動詞で練習する必要があるのか，それとも限られた項目だけを重点的に覚えればよいかがわかるため，日本語教育における導入法や指導の比重の議論を行う際などに重要な情報となる．

キーワード：動詞，前接語，生産性，初級文法項目，機能語，Guiraud Index，BCCWJ

1．問題の所在

この論文では初級の日本語教育で扱われる文法項目の「生産性」について論じる．従来，「生産性」は主として接辞研究の文脈で使用されてきた用語である．しかし，この概念は例えば接続助詞とそれに前接する動詞といった関係にも拡張できるものである．「生産性」は一般化すると以下のように定義することができる．

生産性の定義

ある形式 X が一定の関係 R で結びつく要素の多寡の度合い P を生産性と呼ぶ．

この論文においては，X は初級文法項目となり，R をそれに前接する動詞に絞る．その上で，コーパスの調査を元に生産性 P の可視化方法について吟味し，その結果を論じる．

次に，生産性を可視化する意義について述べる．近年，日本語教育の分野でもコーパスを使った研究が盛んになってきており，前接語に注目した研究もある（清水 2009, 田 2013）．また，中には「てある」の前の動詞の 50% 以上が「書く」であるなど，語彙の大きな偏りを指摘した研究もある（中俣 2011）．これらの研究は各項目の導入に際してどのような例文を使えばよいかということについての貴重な情報であり，教育への貢献は大で

あるといえる．しかし，従来の研究はいずれも個別の文法項目に対する単発的な指摘に留まっている．文法項目全体に対して，様々な動詞と結びつくものはどれで，あまり多くの動詞とは結びつかないものはどれか，すなわち生産性がわかれば，重点的に色々な動詞で導入すべきものと，限られた組み合わせだけを提示すればよいものとが明らかになる．これは，指導の比重を議論する際に重要なデータであり，ある項目の重要度や指導順とは異なる側面で日本語教育文法に貢献するデータといえる．そのためには，生産性を数値化・可視化する方法を確立することが第一に必要となる．本研究ではまず生産性の計算方法を７種類の指標を比較することで吟味し，次に 103 項目の初級文法項目を例として生産性について論じる．

本論文の構成は以下のとおりである．まず，２．で生産性に関する先行研究を概観し，これまで提案されてきた計算方法では大規模コーパスに出現した文法項目の計算には適さないことを示す．３．ではその代案として７種類の指標を提示し，実際に『現代日本語書き言葉均衡コーパス』（BCCWJ）の検索結果から 103 項目について指標を計算し，最も優れた指標として Guiraud Index を採用する．４．では 103 項目の生産性について吟味し，生産性の意義について論じる．５．はまとめである．

２．先行研究

2.1　生産性という概念

本研究の主題である「生産性」（productivity）という用語はもともと形態論の分野，特に接辞の研究において使われていた．議論の対象は英語における過去接辞の規則変化（play-played）と不規則変化（sing-sang），あるいは複数接辞の規則変化（cup-cups）と不規則変化（ox-oxen）などであり，規則変化は文法規則による生成，不規則変化はレキシコンへの貯蓄という二とおりの方法で処理されるとする生成文法家（Pinker& Prince 1994）と，ネットワークモデルによる統一的な説明を試みる認知言語学者（Bybee1985, 1995）の論争があった．ここで Bybee（1995）は生産性について重要な指摘を行っている．生産性を決めるのはその形式の Token 頻度ではなく，Type 頻度であるということである．Croft & Cruse（1999）から例を挙げると，過去接辞 -ed は flew や blew などにしか使われない過去接辞 -ew よりもはるかに Type 頻度が多い．そのため -ed のスキーマの方が強く強化され，あらゆる場面に適用され，生産性が高くなるのである．逆に -ew が不規則的な変化に見えるのは，Type 頻度が少ない結果といえる [1]．

このように，欧米の言語の接辞研究の文脈で生産性という概念が用いられてきた．しかし，Type 頻度やスキーマという概念は日本語の種々の文法項目とそれに接続する動詞にも当てはまるものである．本研究は従来の生産性という概念を日本語の初級文法項目にまで拡張して適用する試みである．ただし，先行研究が規則動詞と不規則動詞のような，生産性の高低が直感でわかるような形式から研究が出発しているのに対して，本研究の目的は教育への応用を前提とし，直感では捉えられないレベルの生産性の違いを可視化することにある．そのため，生産性の計算方法についても先行研究以上の工夫が必要となる．次

1　生産性の低いスキーマが生き残っている原因は，個々の形式の Token 頻度が非常に多いからである．一般に不規則動詞のほうが使用頻度が高いことが知られている．

節以降ではこの点について述べる．

2.2　Type 頻度と Token 頻度

Bybee（1995）は生産性の高低に決定的な役割を果たすのは Type 頻度であると主張したが，その具体的な計算方法については述べていない[2]．単純に調査した結果の Type 頻度（以下 Type 粗頻度と呼ぶ）を比較すればよいかというとそうではない．問題となる形式の Token 頻度に差があれば，当然 Token 頻度の上昇につれて Type 粗頻度が上昇するため，その差を補正する必要が出てくる．結果を先に述べてしまうと本調査における Type 粗頻度と Token 頻度との相関係数は 0.9 にも達し，Token 頻度の影響が大きすぎることがわかる．

具体的な例について述べる．本研究における BCCWJ の調査の結果，極めて近い Type 粗頻度を示した項目に「た後」と「ておく」がある[3]．

表 1：「た後」と「ておく」の頻度

	Type 粗頻度	Token 頻度
た後	2,409	9,711
ておく	2,425	31,801

この 2 項目の生産性はほぼ同じであると主張してよいだろうか．この場合は Token 頻度が小さい「た後」の方が生産性が高いと考えられる．

もう少しモデル化して考えると，同じ Type 粗頻度が 10 であっても，Token 頻度が 10 の時の Type 粗頻度 10 と Token 頻度が 100 の時の Type 粗頻度 10 では意味合いが異なる．前者は Token 頻度があと 10 増えれば Type 粗頻度も大きく増加する可能性が高い．一方，後者では Token 頻度があと 10 増えても Type 粗頻度が 2 倍近く増加するということは考えにくいだろう．Type 粗頻度は Token 頻度ともに増加するが，その増加率が生産性であると言える．下式の関数 F に相当するのが生産性と言える．

$$\text{Type 粗頻度} = F(\text{Token 頻度})$$

また，生産性と似て異なる概念に，語彙多様性（Lexical Diversity）という概念があり，これはある項目に接続する項目の多様性ではなく，一定のテキスト内に出現した語彙の多様性を比較するものである[4]．例えば学習者の作文の LD を測定することで，その習熟度を測定するといった試みが行われており，そのために様々な指標が提案され，比較されている（McCarthy & Jarvis 2007, Koizumi 2012）．そこで問題になるのはテキストの長さ，

2　なお，Bybee（1995）も同じような Type 粗頻度であっても，Token 頻度が異なれば生産性は異なると述べている（p.434）ことから，単純な Type 粗頻度の数ではなく，Token 頻度も考慮に入れた判断を行っていると推測されるが，その方法は定式化されていない．

3　用例の収集方法に関しては 3.2 で詳しく述べる．

4　生産性と語彙多様性は Type 粗頻度と Token 頻度の関係を見るという手法においては共通する部分があるが，テキスト内のバラエティと特定の項目との共起のバラエティという異なるものを対象にしており，その増加率も異なると考えられる．

すなわち Token 頻度の影響をいかに小さくしながら Type 粗頻度を比較するかということで，その影響を抑えるために様々な計算式や手法が考案され，比較されている．LD 研究におけるテキストの長さの問題とはすなわちサンプルサイズの問題である．本研究の場合，同一のコーパスを用いて調査しているので，元となるサンプルサイズは同じである．しかし，検索によって得られたサンプル数（=Token 頻度）は異なり，その異なるサンプル数の中で Type 粗頻度を比較するにあたっては何らかの調整は必須である．Type 粗頻度は Token 頻度に連動して変化するため，その連動している Token 頻度に対して調整が必要となるということである[5]．これは大きなコーパスからサブコーパスを作って比較を行う時に，そのサブコーパスの Token 頻度が異なっている場合は何らかの調整を行う必要があるのと同じである．

2.3　Token 頻度の影響を受けすぎてもいけない

Token 頻度の影響を考慮した生産性の計算方法としては Baayen&Lieber（1991）が以下の式を提案している．

$$p = \frac{n_1}{\mathrm{N}}$$

ここで N は該当形式の Token 頻度を表し，n_1 はその項目が結びつく形式のうち，頻度が 1 であるもの[6]の Type の数である．しかし，この式は生産性の大小がある程度明らかな場合に関しては有効ではあるが，大規模コーパスから取得した種々の文法項目の生産性を計算するのには適さない．この式を使って試みに BCCWJ から得た補助動詞「ている」と「てある」の生産性を計算すると下表のようになる．

表２：従来法による「ている」「てある」の生産性の計算

	N（Token 頻度）	n_1（頻度 1 の動詞）	p（生産性）	前接動詞の Type 粗頻度	TTR
ている	985,113	9,547	0.010	20,907	0.021
てある	15,088	662	0.044	1,217	0.081

表２の p の欄を見ると「ている」は 0.010，「てある」は 0.044 であるがこれは直感に反する．そもそも「ている」は自動詞・他動詞の両方に接続可能であるのに対して，「てある」は他動詞にしか接続しない．また，表３に前接動詞の累積比率を示したが，明らかに動詞の分布は「ている」の方が幅広い．よって，「ている」の方が圧倒的に生産性が高くなくてはおかしい．Baayen&Lieber（1991）の式は Token 頻度で割るという処理を行っているが，「ている」は Token 頻度が飛び抜けて多いためにこの処理が過剰に働いてしまっているのである．

５　教育への応用に関しては重要度に関連する Token 頻度も重要であるとの指摘があり，筆者もその考えに賛同する．ただし，Type 粗頻度は Token 頻度と生産性の２つの変数の影響を受けるということであり，本論文の範囲では基礎研究として文法項目そのものが持つ生産性を分離して明らかにすることに主眼を置くということである．Type 粗頻度は生産性と重要度の双方に配慮した指標である可能性も保留しておく．

６　頻度が１の組み合わせは *hapax legomena* と呼ばれる．

表３:「ている」と「てある」の前接動詞の累積比率

	上位 1 動詞	上位 10 動詞	上位 100 動詞	上位 1000 動詞
ている	7.14%	24.91%	50.18%	83.26%
てある	30.57%	55.37%	81.84%	98.56%

同様の理屈で，前接動詞の Type 粗頻度そのものを Token 頻度で割る，いわば TTR と同じ計算式を用いてもうまくいかない．このことを表２の右側で示した．TTR は Token 頻度の影響を受けすぎることがすでに指摘されており（石川 2012），単純に Token 頻度で割るという処理ではうまくいかないのである．

まとめると本研究が目的とする初級文法項目の生産性の指標は（１）Type 粗頻度を比較する．（２）Token 頻度に配慮する．（３）Token 頻度からの影響が強すぎない．の３つの要件を満たすものでなければならない．次節以降ではこの方法を模索していく．

３．生産性指数の提案

3.1　生産性指数の候補

前節の目的を達成するため，本研究ではまず７つの指標を候補として立て，次いで実際にコーパスから得られた 103 の文法項目について各指標を計算し，その結果から最も良い指標を選び出すという手続きをとる．ここではまず以下の７つの候補について紹介する．これらはほとんどが他の分野の研究に使われている指標であり，Type と Token が関係する生産性指数の計算に流用できそうなものを選出した．名称後の（　）内は略称で，以後の表では略称を使うことがある[7]．

（１）Guiraud 値（G.I.）

語彙多様性（LD）の研究において，TTR における Token 頻度の影響を克服するために考えられた指標の一つである．Token 頻度への配慮という点で本研究の目的と合致するため，生産性指数の候補とした．Guiraud 値は下式で表される．Token の違いに配慮した形で Type 粗頻度を比較できる．以下，Token は該当形式の Token 頻度を表し，Type は前接項目の種類の数を表す．

$$R = \frac{\text{Type}}{\sqrt{\text{Token}}}$$

（２）Herdan の C（C）

Guiraud 値と同じく Token 頻度の影響を抑えた TTR のバリエーションである．計算式は以下のとおり．

$$C = \frac{\log_e \text{Type}}{\log_e \text{Token}}$$

（３）共起する上位 10 項目のカバー率（Top10R）

Baayen&Lieber（1991）の式における n1 とは対称的に生産性が高いものは共起する上位 10 項目の割合が低く，生産性が低いものは上位 10 項目の割合が高い．よってこの数値も生産性の指標となりうる．参考までに「ている」の上位 10 項目の割合は 24.91%，「て

7　なお，以下で示す指標は便宜的にその指標と同じ計算式を用いたという意味であり，今回の調査ではその指標そのものを計算しているわけではないということを断っておく．

ある」の上位 10 語項目の割合は 55.37% である．生産性の高低とこの値の高低は逆転していることに注意が必要である．なお，10 という数字に根拠はない．

（4）標準化 TTR（STTR）

Token 数が増えると TTR が正常な値を示さなくなるという弱点を克服するために，サンプルを一定の数ごとに切り分け，それぞれの TTR を平均した指標である．今回はほとんどの項目が 2,000 例以上収集できたので，多くの項目で標準化した値をとれるように，1,000 例ごとに切り分けを行った．2,000 例以下の項目に関してはそのまま TTR を使用している．

（5）修正 Perplexity（Rperp）

Perplexity は情報理論で定義される概念であり，ある項目の次の項目の候補がどのぐらいの量であるかの見積もりに使われる．これは生産性の概念に非常に近い．各前接項目の出現確率の幾何平均のマイナス 1 乗になる．下式で Token と Type はこれまでと同様，Token 頻度と前接項目の種類の数を表す．$\mathrm{Token_i}$ は i 番目の前接項目との共起頻度である．

$$P = \left(\sqrt[\mathrm{Type}]{\prod_{i=1}^{\mathrm{Type}} \frac{\mathrm{Token_i}}{\mathrm{Token}}} \right)^{-1}$$

ただし，今回のデータでは値が 10〜1,000,000 と非常に幅広く分散し，数値の比較が直感的に行えない．そのため，この対数（底は何でもよいが，今回は 10 で計算した．）をとった修正値を利用する．この修正値は変形すると各前接項目の出現確率の対数の算術平均×－1 となり，下式で表される．

$$rP = -\frac{1}{\mathrm{Type}} \cdot \sum_{i=1}^{\mathrm{Type}} \log_{10}\left(\frac{\mathrm{Token_i}}{\mathrm{Token}}\right)$$

（6）ジニ係数（Gini）

ジニ係数は経済学で所得分配の不平等さを測定するのに使われる指標である．右図の横軸は所得の少ない人間から所得の多い人間の順に並べたものであり，縦軸は所得の少ない人間から所得の多い人間の所得を累積していったものが全体の何％にあたるかを示すものである．仮に所得の差が０であれば，左下から右上を結ぶ均等分配線と呼ばれる直線となるが，実際にはそのようなことはあり得ず，その下のローレンツ曲線と呼ばれる曲線を描く．ジニ係数とはこの均等分配線とローレンツ曲線で囲まれた面積の２倍であり，０〜１の値をとる[8]．

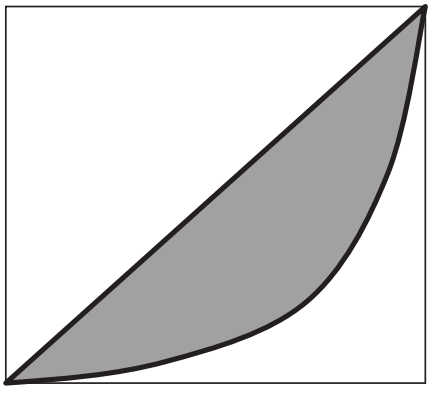

図１：均等分配線とローレンツ曲線

８　実際的な計算は Excel を用い，台形公式により算出した．具体的な手法については以下の URL を参考にした．
ローレンツ曲線の作成とジニ係数の計算 with Excel
http://hitorimarketing.net/tools/lorenz-curve_and_gini-coefficient.html

（7）エントロピー（Entr）

エントロピーも情報理論における基礎概念であり，出来事の情報量を表す．ここでは，該当形式の前接項目が出現する確率を求め，平均情報量と呼ばれる指標を計算した．これは「ある項目に前接しうる項目」という集合が持つ情報量であり，その大小が生産性を表すと読み取れる．計算式は以下のとおりである．

$$H = -\sum_{i=1}^{Type} \log_{10}\left(\frac{Token_i}{Token}\right) \cdot \frac{Token_i}{Token}$$

3.2 調査用のデータについて

本節では，前節で提案した7つの指標のうち，どれが初級文法項目の生産性を計測する上で最も適しているかを考察する．そのために，具体的には103項目の初級文法項目についてそれぞれ7つの指標を計算し，その計算結果から判断していくという手法をとる．

まず，この103項目については庵ほか（2000）より見出し語としてあげられているものの中から動詞に接続しうるものを選出した．つまり，初級文法項目でも今回は格助詞やとりたて助詞は対象としていない．具体的な項目は付表に示したものである．

次に，『現代日本語書き言葉均衡コーパス』（BCCWJ）の検索アプリケーション『中納言1.1.0』[9]を用い，長単位検索モードを使って文法項目の直前に接続する動詞と，文法項目の前に助動詞を1つ挟んで接続する動詞を全て検索，ダウンロードした．具体的には以下のとおりである．

「食べ／ている」 「食べる」は「ている」の前接動詞としてカウント．

「言わ／れ／ている」 「言う」は「ている」の前接動詞としてカウント．

「殴／られ／続け／ている」 「殴る」は「ている」の前接動詞としてカウントしない．

ダウンロードした用例は103項目で延べ4,819,084例である．

長単位検索モードを用いた理由は，「かもしれない」のように，多くの初級文法項目が1単位としてカウントされているため検索が容易であることと，短単位検索では「勉強する」「失敗する」などのサ変動詞が2語に分割され，結果，「する」の数が増えるなどType粗頻度に大きな影響を与えてしまうためである．また，前接語の品詞を動詞のみに絞った理由は名詞を対象に入れると，「ている」のように動詞のみと接続する形式と「かもしれない」のように，動詞，形容詞，名詞＋コピュラと接続する形式で基準が異なってくるからであり，比較に支障をきたすこと，また，名詞は動詞と比較して，頻度が少ない語が大量に出現するという傾向をとるからである．

3.3 測定と評価

続いて，7つの指標を比較していく．生産性の指標に求められる性質は2．節で議論したとおり，（1）Type粗頻度を比較する．（2）Token頻度に配慮する．（3）Token頻度からの影響が強すぎない．の3点である．本節ではこれらに加え，実用における他の観点も加えて1つの指標を決定する．

まず，これらがToken頻度に配慮した結果を出せているかどうかを確認する．以下にToken頻度がほとんど同じでType粗頻度に3倍以上の差があった「た後」と「ておく」

9 https://chunagon.ninjal.ac.jp/login

の計算結果を示す．このような時にはToken頻度が小さい「た後」の方が生産性が高いという結果を出す指標が望ましい．

表４： ７指標による「た後」と「ておく」の計算結果

	G.I.	C	Top10R	STTR	RPerp	Gini	Entr
た後	24.45	0.85	18.33%	0.54	3.72	0.66	8.00
ておく	13.60	0.75	30.17%	0.35	4.10	0.85	7.97

上の７指標のうち修正Perplexityとジニ係数は「た後」の方が小さい値を示しており，これは求める指標の性質を満たさないことがわかる．一方，上位10項目の割合に関しては数値の大小が逆転するため，この値は「た後」の方が生産性が高いことを示している．「た後」と「ておく」の他，（１）Type粗頻度で並べた時に隣接し，（２）Type粗頻度の差が100以下で，（３）Token頻度に２倍以上の差がある２項目，のべ15関係において同様の結果であった．よって修正Perplexityとジニ係数は候補から外してよい．なお，表４において差が最も顕著なのがGuriaud Indexで差が最も小さいのがエントロピーであった．

次に，各指標間の相関，ならびにType粗頻度・Token頻度との相関を示す．

表５： ７指標と頻度の103項目に対する計算結果の相関

	Token	Type	G.I	C	Top10R	STTR	Gini	RPerp	entr
Token	1.000								
Type	0.901	1.000							
G.I.	0.248	0.539	1.000						
C	-0.231	-0.211	0.426	1.000					
Top10R	-0.155	-0.260	-0.648	-0.585	1.000				
STTR	0.244	0.423	0.819	0.514	-0.764	1.000			
Gini	0.356	0.456	0.056	-0.445	0.285	-0.024	1.000		
RPerp	0.629	0.842	0.622	-0.119	-0.274	0.410	0.589	1.000	
entr	0.308	0.483	0.811	0.508	-0.895	0.806	-0.077	0.520	1.000

この論文の目的としては生産性についての議論に的を絞るため，できるだけToken頻度からの影響は少ない方が望ましい．この観点から見ると修正perplexity（0.629）はToken頻度の影響が強すぎることになる．他方，Guiraud Index，C，上位10項目の割合，標準化TTRはそれほど相関は強くない．中でも最もToken頻度からの影響が小さいのは上位10項目の割合（-0.155）である．しかし，この指標は生産性が高いものほど数値が低いという特性があるため，実用面においてやや難があるという問題点がある．なお，Type粗頻度とToken頻度の相関は0.9と極めて高く，Type粗頻度は生産性の指数としては適さないことがわかる．

他方，生産性の主たる決定要因であるType粗頻度と強い相関があるのは修正perplexity（0.842）であり，ついでGuiraud Index（0.539）である．

最後に，それぞれの計算結果がどのような分布を見せるのかを確認するために，平均値，中央値，最大値，最小値，尖度，歪度を比較する．

表6： 7指標の103項目に対する計算結果の基本統計量の比較

	G.I.	C	Top10R	STTR	RPerp	Gini	Entr
平均値	15.95	0.78	38.51%	0.37	3.85	0.77	7.64
中央値	16.71	0.79	36.08%	0.37	3.84	0.79	7.81
最大値	27.10	0.90	100.00%	0.55	5.99	1.00	9.77
最小値	0.53	0.40	13.73%	0.03	1.56	0.19	2.02
尖度	0.77	22.07	3.60	0.81	2.16	7.42	4.78
歪度	-0.59	-3.40	1.28	-0.69	-0.02	-2.10	-1.45

指標の評価の基準としては色々あるが，まず，尖度と歪度に注目すると，Guiraud Indexと標準化TTRがともに1に近く，非常に正規分布に近いことがわかる．その上で最大値と最小値の範囲（レンジ）に注目すると，STTRは0.03から0.55と比較的小さい範囲に留まるのに対して，Guiraud Indexは0.53から27.10とかなり広い範囲に値が分布しており，直感的な比較がしやすい．将来，教育への応用を考えた場合，Guiraud Indexであれば「20以上の値は高生産性項目」というようにわかりやすい目安を規定して議論をすることも容易であるが，他の指標では例えば「0.45以上の値は高生産性項目」というような数値をとることになり，統計的な処理はともかく，直感的にはわかりにくいと考えられる．直感的な扱いやすさではGuiraud Indexに軍配があがるといえよう．

ここまで，様々な角度から7指標を評価してきた．結論としては，この論文ではGuiraud Indexを生産性指数として採用する．その理由は下記の（a）～（e）のとおりである．

（a） Token頻度の差を考慮した値を示している．

（b） Token頻度との相関は強くなく，影響を受けすぎていない．その一方でType粗頻度との相関が2番目に強い．

（c） 計算結果が非常に正規分布に近い．また，扱いやすい範囲に数値が分散し，項目間の比較が容易である．

（d） ある形式の前に別の形式が付加された複合形式は，元の形式と同じような数値を示すなど，一貫性が見られる．

（e） Type粗頻度とToken頻度さえわかれば容易に計算できる．

（a）から（c）まではすでに述べたとおりである．（d）は例えば「てください」の順位が91位（G.I.=10.58）であるのに対して，「ないでください」の順位が92位（G.I.=10.08）であるなど，似た形式は似た数値を示すということであり，これらの値が決してランダムではないことを意味する．同様の性質は標準化TTRにも見られるが，（b）の値の範囲という観点，また（e）の計算の容易さという観点からGuiraud Indexが優れた指標であると判断した．

ただし，実際にはGuiraud Indexと同じ計算式で表されるという指標ということである．よって本節において生産性指数を以下のように改めて定義し，以降の議論では「生産性指

数」という名称を用いる．

$$\text{生産性指数} = \frac{\text{共起項目の Type 粗頻度}}{\sqrt{\text{対象項目の Token 頻度}}}$$

４．初級文法項目の生産性

4.1　初級文法項目の生産性の一覧

103 項目を生産性の高いものから順に並べた表を付表に示す．表のうち，◆をつけた 20 項目は例文を収集する際に直前の動詞のみを収集したものである [10]．ただし，１つ前の動詞のみを収集するか，２つ前の動詞まで収集するかという違いの影響は小さい [11]．また，▼をつけた 5 項目は検索の都合上，長単位検索ではなく短単位検索を使用したものである．このことは Type 粗頻度に大きな影響を与えるため，数値はあくまでも参考であり，実際にはもっと高い生産性を示すものであると考えられる．

付表は四分位のところで太線で区切ってある．すなわち 25% ずつに区切っている．生産性の分布は非常に連続的で，例えば隣接する項目でどちらが生産性が高いか，というような議論はあまり意味があるとはいえない．その一方で，上位の項目と下位の項目では確かに差異が認められる．日本語教育への応用を考える際も，上位群どうしの項目を生産性という単一のパラミタで議論することには問題があり，その際はむしろ他の要素を考慮に入れるべきである．しかし，上位群と下位群には生産性の違いがあり，導入や練習の仕方を変える一つの指標となるといえる．

4.2　何が生産性の高低を生んでいるのか？

ここでは，生産性の高低は何を意味しているのか，どのようなパラミタがこの生産性の高低に寄与しているのかを考察する．

まず，生産性の高い 15 項目を表 7 に示す．なお，◆の記号はその項目の１つ前の動詞のみを調査したことを示す．

表 7： 生産性の高い 15 項目

項目	生産性指数	項目	生産性指数	項目	生産性指数
たり	27.10	るとき	22.40	ようになる◆	21.91
た後	24.45	たら	22.29	ている	21.06
のに	23.98	から	22.22	てしまう	20.58
ことになる	23.09	させる	22.13	ながら	20.40
ようにする◆	22.93	たとき	22.03	と	20.36

10　その理由は様々であり，「食べやすい」のように長単位検索では１語となってしまうため，１つ前までしか見られなかったものや，処理時間の関係で１語前までしか検索できなかったものもある．

11　参考までに「ている」の値を書くと１つ前の動詞のみで計算した場合の生産性指数は 20.59 であり，２つ前の動詞まで計算に入れると 21.06 となる．この程度の差異であれば，他の項目との相対的な比較において影響はほとんどない．本研究で示す生産性指数は相対的な大小を計るものであり，数字は絶対的なものではない．そもそもコーパスを変えれば数値は変動するため，相対的な大小のみが問題となる．

最も生産性が高い項目は並列を表す「たり」(27.10) であり，２位の「た後」(24.45) と比べても比較的大きな差がある．並列はその内容に特に意味的制約を持たないため，生産性が高くなることが予測される．また，「P たり Q たり」のように複数で使われることも多くその際 P と Q は異なる動詞でなければならないため，このことも動詞のバラエティの増大につながったと考えられる．なお，中俣 (2009) は「たり」と「し」では「たり」の方が意味的制約が緩いとしており，今回の調査結果でも「たり」(27.10) は「し」(19.68) よりも生産性が高かった．また，「〜るとき」(22.40)「〜た後」(25.45)「たら」(22.29)「ながら」(20.40) など時間に関係する節を作る要素も生産性が高い．これも意味的制約が少ないからである．アスペクト形式「ている」(21.06)「てしまう」(20.58) はやや数値が低いが，これはやや意味的制約を持つためであると解釈できる．「ことになる」(23.09)「ようにする」(22.93)「ようになる」(21.91) も時間の変化に関わる複合形式である．

日本語教育の観点からいえば「たら」(22.29) が「と」(20.36) よりも高い点も注目に値する．書き言葉のコーパスということもあり，Token 頻度は圧倒的に「と」が多いのであるが，「たら」の方が汎用的といえる．これはコーパスに基づいても「たら」の方が基本的な表現であると主張する根拠になる．

また，ヴォイスの「させる」(22.13) は実際には他動詞化標識として使われている（森 2012）こともあり，生産性が高い．なお，同じくヴォイス形式の「られる」は生産性指数が 19.09 である．しかもこの値は可能と受身を区別しておらず，「受身」の「られる」はこれよりも数値が低いと考えられ，「させる」と「られる」では「させる」の方が若干生産性が高いといえそうである．

また，上の表では３位に「のに」(23.98) があるが，これは「送ったのに届かない」のよな逆接の「のに」と「書くのに使う」のような目的の「のに」の区別ができておらず，各用法の実際の数値はこれよりも低い．

まとめると，生産性の高い項目は並列の「たり」，理由の「から」，他動詞化標識の「させる」を除くとほぼすべて時間に関わる表現が並ぶ．時間の表現は動詞の意味内容に関係なく，単に時間軸上の位置を示すだけなので，どのような動詞とも共起すると考えられる [12]．

次に，生産性の低い 15 項目を表８に示す．なお，◆の記号はその項目の１つ前の動詞のみを調査したことを示す．また，▼の記号は短単位検索モードを利用したことを示す．

表８： 生産性の低い 15 項目

項目	生産性指数	項目	生産性指数	項目	生産性指数
てあげる	11.65	ないか◆	9.82	ませんか	7.25
しようか◆	11.64	たことがない◆	9.52	終える◆	6.50
てください	10.58	ましょうか	8.88	お〜になる（尊敬）▼	5.94
ないでください	10.08	すぎる▼	8.05	お〜する（謙譲）▼	2.46
てある	9.91	終わる◆	7.91	やむ◆	0.53

12　今回，接続助詞の「て」と過去・完了の「た」はデータがあまりにも膨大になるため，調査の対象としなかった．しかし，これらはもちろん非常に生産性が高いと考えられる．

生産性が低い文法項目では，「終わる」（7.91）「終える」（6.50）「やむ」（0.53）など終了を表すものが目立つ．これらは終了が問題になるタイプの動詞としか結びつかないため，生産性が低くなっていると考えられる．なお，開始を表す「始める」（17.23）は文法的には「終わる」とほぼ等しいにもかかわらず，生産性指数に関しては対称的に平均値よりも高い数値を示している点も興味深い．

しかし，生産性が低い項目にはもう１つ顕著な共通点がある．「ませんか」（7.25）「ましょうか」（8.88）「しようか」（11.64）「てください」（10.58）「お～になる」（5.94）「お～する」（2.46）は全て対人モダリティ，あるいは具体的な聞き手が必要となる表現であり，全て書き言葉というよりも，むしろ話し言葉でよく用いられる表現であるということである．

このことの傍証として，各用例のサブコーパスの比率に着目してみたい．表７は各項目のサブコーパスごとの比率のうち，話し言葉的な特徴を持つと考えられる「Yahoo! 知恵袋」「Yahoo! ブログ」「国会会議録」とその３つの合計，さらに最も固い書き言葉であると考えられる「白書」の割合を示したものである．表の後半の生産性の低い項目は，表の前半の生産性の高い６項目と比べて，話し言葉的な特徴を持つ３サブコーパスの合計が高く，どれか１つが突出しているものが多い．また，「白書」の割合も低い．

表９： 生産性の高低とサブコーパスの比率

項目	生産性指数	知恵袋	ブログ	国会	合計	白書
たり	27.10	10.6%	17.0%	1.5%	29.1%	0.8%
た後	24.45	10.2%	11.3%	1.8%	23.4%	2.9%
のに	23.98	16.6%	11.4%	1.7%	29.7%	0.6%
ことになる	23.09	4.7%	4.2%	11.8%	20.8%	1.4%
ようにする	22.93	11.5%	4.9%	2.8%	19.2%	2.3%
るとき	22.40	21.3%	11.4%	3.0%	35.8%	0.7%
しようか	11.64	16.1%	23.2%	7.00%	46.3%	0.3%
てください	10.58	58.9%	8.0%	1.5%	68.4%	0.0%
ましょうか	8.88	12.1%	6.0%	25.5%	43.6%	0
ませんか	7.25	54.2%	3.9%	7.3%	65.4%	0.1%
お～になる（尊敬）	5.94	9.0%	4.5%	27.1%	40.6%	0.0%
お～する（謙譲）	2.46	35.4%	15.7%	15.1%	66.2%	0.0%

この傾向は他の文法項目にも見られる．例えば，理由を表す「から」と「ので」は「ので」の方が丁寧であるとされ（日本語記述文法研究会 2008:126），そのために話し言葉に多く用いられると考えられるが，生産性指数は「から」22.22 に対して，「ので」13.78 と「ので」の方が圧倒的に低い．逆接に関しても「が」19.77 に対して，「けど」は 12.54 である．異形態であると考えられる「なければならない」と「なければいけない」も，生産性指数は前者が 19.06，後者が 13.73 でやはり後者の方が生産性が低い．また，モダリティ形式においても「ようだ」18.35 に対して「みたいだ」17.83，「伝聞そうだ」17.40 に対して「らしい」16.76 と，やはり話し言葉的とされるものが数値が低いという結果が一貫

して見られる[13].

これらについてもサブコーパスの比率の比較を行った結果「伝聞そうだ」:「らしい」のペア以外では生産性が低いものが話し言葉的であることが確認された[14].

表 10：類義表現のサブコーパス比率の比較（%）

項目（生産性指数）	知恵	ブロ	国会	合計	項目（生産性指数）	知恵	ブロ	国会	合計
から（22.22）	13.6	12.0	4.9	30.5	ので（13.78）	27.8	17.7	5.6	51.0
が（19.77）	16.9	7.9	8.5	33.3	けど（12.54）	17.4	21.5	21.4	60.3
なければならない（19.06）	2.3	2.5	10.1	14.9	なければいけない（13.73）	14.1	6.6	20.9	41.6
ようだ（18.35）	20.1	18.4	1.4	39.9	みたいだ（17.83）	29.7	25.0	0.8	55.3
そうだ（17.40）	31.5	24.4	2.1	57.9	らしい（16.76）	14.8	20.0	0.3	35.1

これは単に書き言葉コーパスには話し言葉特有の要素が出現しないから，という理由では説明がつかない（生産性指数と Token 頻度の相関は 0.25 である）．生産性の低い項目でも，「謙遜お～する」「てある」「てください」はいずれも出現数が 1 万を越える．他方，生産性指数が 2 番目に高い「た後」の出現数は 9,711 であり，出現数の問題だけでは片付かない．現時点で言えるのは，「書き言葉においては，話し言葉的な性質を持つ項目は，生産性が低くなる」ということである．ただし，これが BCCWJ を使ったためなのかどうかは，今後話し言葉コーパスを用いて検証する必要がある．もっとも，「終える」など，書き言葉的と考えられる項目にも生産性の低いものがあり，単一のパラミタで生産性を説明することはできない．現時点では，複数考えられるパラミタの 1 つに，話し言葉性が存在する，と結論づけたい．

なお，生産性が低い項目のうち「たことがない」は「見る」「聞く」が，「てある」は「書く」が圧倒的に多くを占める項目であり，日本語教育でも語彙項目に近い扱いが可能となる項目である．

4.3 生産性の意義

最初に述べたように，生産性は日本語教育においてその項目を文法項目として重点的に扱い，様々な動詞とともに練習させるのが良いのか，それとも結びつく動詞が限られているため，頻出する組み合わせを語彙として覚えればよいのか，という比重の議論を行うのに役に立つ．

ただし，注意しないといけないのは，生産性が示唆するのは「どのように」教えるのかという点であって，重要度とは別の観点であるということである．生産性が低いというこ

13　日本記述文法研究会（2003）は「みたいだ」を話し言葉的としている．また，伝聞の「そうだ」は話し言葉ではあまり使わないとしている．ただし，「ようだ」と「みたいだ」，「そうだ」と「らしい」は，「が」と「けど」や「から」と「ので」のように機能が一致し，もっぱら文体上の違いのみがあるわけではなく，若干の用法の違いも存在するため，生産性の違いはそれほど大きく現れなかったと考えられる．

14　先行研究では「そうだ」のほとんどは「そうです」という丁寧形をとり，他方，「らしい」は非丁寧形が多いことがわかっており，このことが「そうだ」のサブコーパスの比率に影響していると考えられる（小西 2011，中俣 2014）．

とが初級で教えなくてもよいということを意味するわけではない．生産性が低くても，ある特定の組み合わせが非常によく使われる（=Token 頻度が多い）ということは十分に考えられる．その代わり，生産性が低いということは特定の動詞とよく結びついて使われるということなので，それをよく使うフレーズとして覚えさせるという方略が有効に働く．

例えば，謙譲語の「お〜する」を例にあげる．これは仕事や接客の場面などで必要になる表現で，そのような場面に接するのであれば，初級であろうと必要だと考えるべきである．また，実は「お〜する」のうち約 40% が「お願いする（お願いします）」という形で使われている．これはどのような学習者にとってもよく使うフレーズとして覚えるべき表現である．実際の教育の現場においても謙譲語を導入するまで「お願いします」を扱わない，といったことはありえないだろう．生産性指数のデータはこの「お願いします」のようにフレーズとして扱った方が良い表現が他にもあることを教えてくれる．

逆に，生産性の高い「から」「たり」「ている」などはどのような初級学習者にとっても，生産的に学習する意義のある項目だと考えられる．例えば，「て」形のドリル練習について言えば，生産性が低い「てください」で行っても実際には使わない用例が多く含まれることになる．一方，生産性が高い「ている」で行えば，多くの用例は実際に使われるため，効率的であることが示唆されるのである [15]．

また，本研究では初級の文法項目を対象としたが，生産性指数がより意義を持つのは中上級項目であると考えられる．中上級の文法項目とされるものの中には，自由に動詞と組み合わせることが難しい，いわば語彙に近い項目が多くみられる．そのため，文法的な項目と語彙的な項目を切り分ける必要性は初級よりも切迫してあるが，直感で線を引くのは難しい．今回提案する手法で生産性を可視化すれば，文法と語彙の連続性（Langacker2008）を認めながらも，教育のために便宜的に線を引くことも可能になると考えられる．一例を挙げると同時を表す「動詞＋ついでに」の生産性指数は 7.14，「動詞＋かたわら」の生産性数は 6.94 であるのに対し，「動詞＋がてら」の生産性は 3.48 である [16]．つまり，「動詞＋がてら」の生産性はかなり低く，「ついでに」「かたわら」と同列に扱う必要のない（あるいは扱ってはいけない）ことを根拠を持って示すことが可能になるのである．

15 「てください」の生産性が低いことについては異論があるかもしれないが，今回の調査対象が BCCWJ であることを考慮しても，初級のコースで扱われている「てください」の使われ方が母語話者のそれと乖離している可能性は考えられる．清（2004）は「ないでください」の使用に関して日本語教師とそれ以外の母語話者の間に乖離があることを明らかにしているが，同様の乖離が肯定形である「てください」に起きていることは十分に考えられる．とはいえ，筆者はサバイバル日本語として，あるいは教室での指示のことばとして，「てください」を初期に導入することについては意義があると考えている．ただし，学習者に産出させる練習は「教えてください」「見てください」など学習者が実際に使うようなものに限るべきである．「読んでください」「泳いでください」のような「て」形の練習のためのドリル練習を「てください」で行っても意義は薄いという主張である．「生産性」（productivity）はその名のとおり，産出（production）に関わる概念であると考えられる．

16 「動詞＋がてら」は庵ほか（2001）に「挨拶をしがてら」「試しがてら」といった例で紹介されている（p.445）．しかし，このような指導が本当に必要かという検証も生産性を元に行う必要があろう．

５．まとめ

この論文では文法項目の生産性についての計算式を７つの候補式から吟味し，共起項目の Type 粗頻度÷√対象項目の Token 頻度という生産性指数を提案した．この生産性指数は以下のような特性を持つ．

（１）　0〜30 の間におさまり，直感的に扱いやすい．また計算も容易である．

（２）　高低が記述文法における研究成果と一致する部分がある．

その上で，BCCWJ のデータを元に，初級 103 項目に対して生産性指数を計算し，生産性のランキング表を提示した．この指標が，即日本語教育の現場で活かせるとはいえないだろう．しかし，大規模なシラバスの見直しを行う際には，種々の問題（どれが初級相当なのか？どれが「文法相当」でどれが「語彙相当」なのか？単にフレーズを丸覚えすればよいのかあるいはある程度生産的な練習が必要なのか？）を考える上で，少なくとも粗頻度などの指標よりも参考にする価値のある指標であると考える．

本研究は萌芽的なものであり，今後究明すべき課題は多く残されている．

（３）　他のコーパスで計算しても同じような結果になるのか[17]．

（４）　中級以上の文法項目ではどうか．

（５）　この計算式は言語の他の部分にも適用できるのか．例えば動詞の生産性を共起する名詞の Type 粗頻度・Token 頻度から計算することはできるか．

（６）　生産性は学習難易度と何か関係があるのか．

（７）　他言語ではどうなるのか．生産性の高い項目，低い項目に見られる意味・機能的な偏りは他言語でも成り立つのか．

生産性の特性について更に考察を深め，日本語教育の文法・語彙シラバスに寄与できる成果を出すことが今後の課題である．

謝辞

本研究は，「学習者コーパスから見た日本語習得の難易度に基づく語彙・文法シラバスの構築　第７回共同研究会」（2013.3.2　京都教育大学），「計量国語学会第五十七回大会」（2013.9.28　首都大学東京秋葉原サテライトキャンパス）ならびに「京都言語学コロキアム」（2013.10.26　京都大学）にて発表した内容をまとめたものである．コメントを下さった方々に感謝申し上げる．特に，石川慎一郎氏，荻野綱男氏，森篤嗣氏，山崎誠氏には可視化の方法について提案を頂いた．また，山内博之氏には研究の構想の段階から，張麟声氏，清水由貴子氏，花村博司氏には論文の執筆の段階で助言を頂いた．英語における語彙密度研究に関しては石井卓巳氏に教示頂いた．以上の方々に感謝申し上げる．また，査読者からのコメントにも大いに助けられたことに御礼申し上げる．なお，論文における瑕疵は全て筆者の責任である．

17　話し言葉における言語活動は，書き言葉における言語活動よりもずっと限定された語彙で行われているはずであり，Type 粗頻度は当然少なくなる．Token 頻度もコーパスが変われば変化するが，こちらの影響は小さい．そのため，生産性が高いとされている形式が，もっと低くなることはありそうだが，生産性が低いとされた形式がもっと高くなるということは考えにくい．今回の調査は，BCCWJ を使ったことで，ある意味生産性の「最大値」を見積もったという言い方ができるかもしれない．

文献

Baayen, H., & Lieber, R. (1991) Productivity and English derivation: a corpus-based study, *Linguistics*, 29:801-843.

Bybee, J. L. (1985) *Morphology: A study of the relation between meaning and form*. Philadelphia: John Benjamins.

Bybee, J. L. (1995) Regular Morphology and the Lexicon, *Language and cognitive processes*,10(5):425-455.

Croft, W., & Cruse, D. A. (1999) *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Koizumi, R., (2012) Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens?, *Vocabulary Learning and Instruction*, 1(1):60-69.

Langacker, R., (2008) *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

McCarthy, P. M., & Jarvis. S., (2007) vocd: A theoretical and empirical evaluation, *Language Testing*, 24:489-488.

Pinker, S., & Prince, A. (1994) Regular and irregular morphology and psychological status of rules of grammar, In S.D.Lima et als (eds.) *The reality of linguistics rules*. 353-388. Amsterdam: John Benjamins.

庵功雄ほか（2000）『初級を教える人のための日本語文法ハンドブック』スリーエーネットワーク .

庵功雄ほか（2001）『中上級を教える人のための日本語文法ハンドブック』スリーエーネットワーク .

石川慎一郎（2012）『ベーシックコーパス言語学』ひつじ書房 .

小西円（2011）「使用傾向を記述する―伝聞の［ソウダ］を例に―」森篤嗣・庵功雄（編）『日本語教育文法のための多様なアプローチ』159-181. ひつじ書房 .

清水由貴子（2009）「反復の意味を表す「V1 テハ V2」文の分析―形式的側面を中心に―」『日本語文法』9（1）:4-70.

清ルミ（2004）「「コミュニケーション能力育成の視座から見た日本語教科書文例と教師の“刷り込み”考―『ないでください』を例として」『異文化コミュニケーション研究』16 :1–24.

田昊（2013）「「言いさし」の「けど」類の使用実態に関する一考察―日本語教育文法の観点から―」『日本語教育』156:45-59.

中俣尚己（2009）「日本語並列表現の体系と記述」大阪府立大学博士学位論文 .

中俣尚己（2011）「コーパス・ドライブン・アプローチによる日本語教育文法研究―「てある」と「ておく」を例として―」森篤嗣・庵功雄（編）『日本語教育文法のための多様なアプローチ』215-233. ひつじ書房 .

中俣尚己（2014）「伝聞の「そうだ」が伝えるもの―機能語と実質語のコロケーション研究―」『京都教育大学国文学会誌』41: 左 1-17.

日本語記述文法研究会（2003）『現代日本語文法 4 第 8 部モダリティ』くろしお出版 .

日本語記述文法研究会（2008）『現代日本語文法 6 第 11 部複文』くろしお出版 .
森篤嗣（2012）「使役における体系と現実の言語使用―日本語教育文法の視点から―」『日本語文法』12（1）:3-19.

（2014 年 4 月 28 日受付，2015 年 8 月 21 日再受付）

付表　初級文法項目の生産性指数のランキング表

◆は１語前まで検索.　▼は短単位検索を使用.

順位	項目	Token	Type	生産性
1	たり	73,100	7,327	27.10
2	た後	9,711	2,409	24.45
3	のに	23,524	3,678	23.98
4	ことになる	34,171	4,269	23.09
5	ようにする◆	9,199	2,199	22.93
6	るとき	43,557	4,674	22.40
7	たら	87,198	6,582	22.29
8	から	61,214	5,497	22.22
9	せる／させる	122,746	7,755	22.13
10	たとき	47,316	4,791	22.03
11	ようになる◆	22,858	3,313	21.91
12	ている	985,113	20,907	21.06
13	てしまう	92,258	6,252	20.58
14	ながら	61,489	5,059	20.40
15	と	219,995	9,549	20.36
16	と思う	13,518	2,363	20.32
17	う／よう（意向形）◆	96,339	6,307	20.32
18	はず	17,354	2,661	20.20
19	ていく	66,242	5,187	20.15
20	が	144,300	7,510	19.77
21	ても	88,715	5,885	19.76
22	ば	135,634	7,267	19.73
23	し	28,704	3,334	19.68
24	ために	31,855	3,512	19.68
25	なら	16,504	2,463	19.17
26	れる／られる	781,023	16,868	19.09
27	なければならない	22,679	2,871	19.06
28	だろう	24,606	2,985	19.03
29	てから	24,283	2,945	18.90
30	つもり◆	6,443	1,513	18.85
31	そうだ（様態）	20,090	2,659	18.76
32	ことができる	41,206	3,746	18.45
33	ようだ	14,950	2,244	18.35
34	命令形現代◆	28,399	3,076	18.25
35	てくる	96,994	5,619	18.04

36	まで	11,632	1,925	17.85
37	みたい	5,420	1,313	17.83
38	かもしれない	9,133	1,692	17.70
39	か	62,582	4,419	17.66
40	前に	6,890	1,466	17.66
41	ないで	7,438	1,522	17.65
42	ずに	17,901	2,353	17.59
43	そうだ（伝聞）	6,232	1,374	17.40
44	でしょう	21,637	2,535	17.23
45	始める	18,175	2,323	17.23
46	た後で	1,264	609	17.13
47	たい	115,706	5,815	17.10
48	のです	185,304	7,322	17.01
49	ことがある	5,047	1,200	16.89
50	ことにする	7,434	1,456	16.89
51	らしい	5,275	1,217	16.76
52	ず	54,264	3,892	16.71
53	てやる	6,187	1,311	16.67
54	に違いない	1,706	688	16.66
55	てくれる	59,040	4,042	16.64
56	までに	2,018	746	16.61
57	てもらう	26,469	2,699	16.59
58	ていただく	16,890	2,061	15.86
59	よ	32,113	2,782	15.52
60	やすい◆	20,866	2,212	15.31
61	てほしい	10,009	1,509	15.08
62	うちに	5,407	1,083	14.73
63	たまま	9,983	1,471	14.72
64	ましょう	25,144	2,320	14.63
65	てはいけない	4,365	958	14.50
66	ね	16,391	1,851	14.46
67	のだ▼	28,054	2,390	14.27
68	てもいい	6,233	1,123	14.22
69	続ける◆	13,440	1,634	14.09
70	たことがある◆	11,437	1,485	13.89
71	あいだ	1,923	606	13.82
72	ので▼	56,286	3,269	13.78
73	なければいけない	1,936	604	13.73

74	ておく	31,801	2,425	13.60
75	あいだに	2,204	630	13.42
76	たほうがいい◆	8,383	1,200	13.11
77	てみる	61,440	3,221	12.99
78	つつある	5,092	924	12.95
79	前	677	336	12.91
80	にくい◆	8,162	1,165	12.90
81	な◆	7,316	1,102	12.88
82	てくださる	2,167	590	12.67
83	けど	34,782	2,338	12.54
84	なさい◆	5,523	919	12.37
85	なくてもいい	2,844	657	12.32
86	ているところだ	2,774	648	12.30
87	なくて◆	3,627	719	11.94
88	よね	6,075	911	11.69
89	てあげる	8,449	1,071	11.65
90	しようか◆	9,258	1,120	11.64
91	てください	49,234	2,347	10.58
92	ないでください	1,467	386	10.08
93	てある	15,088	1,217	9.91
94	ないか◆	1,404	368	9.82
95	たことがない◆	8,125	858	9.52
96	ましょうか	1,756	372	8.88
97	すぎる▼	7,766	709	8.05
98	終わる◆	2,012	355	7.91
99	ませんか	7,013	607	7.25
100	終える◆	1,167	222	6.50
101	お〜になる（尊敬）▼	4,284	389	5.94
102	お〜する（謙譲）▼	28,489	415	2.46
103	やむ◆	286	9	0.53

※付表の注

今回，接続助詞の「て」と過去・完了の「た」はデータがあまりにも膨大になるため，調査の対象としていない．

また，34 位の「命令形現代」とは「命令形」での検索結果のうち「〜せよ」「見よ」「寝よ」のように「よ」で終わる語形を除き，「〜しろ」「見ろ」「寝ろ」のような語形のみを集計したものである．

Mathematical Linguistics, Vol.29 No.8 (March 2015) pp.275-295.

Paper

Visualization of the Productivity of Function Words in Basic Japanese Grammar Syllabus: In a Case of Verb and Function Words

NAKAMATA Naoki (Kyoto University of Education)

Abstract:
The aim of this paper is to propose a new way of quantifying the productivity of 103 functional words taught in a Basic Japanese Course. To measure productivity beyond large differences between token-frequency of items, seven candidate measures are prepared. Then, 103 items in BCCWJ are calculated for every 7 indexes. After that, the candidates were evaluated from several perspectives. As a result, the formula equivalent to Guiraud Index is the most suitable: Type divided by square root of Token. And the results are consistent with the preceding insights regarding the descriptive grammar of Japanese. High-productive items mainly correspond to time; since they involve no constraint on verb meanings, they can collocate with any verbs. Low-productive items, in contrast, contain a lot of markers for asking, proposal, and prohibition, as well as honorifics. These all are used more in oral communication than in written words. Finally it is claimed that this new index can contribute to education of Japanese language, because we can know which items can co-occur with many verbs or situation and which not, which is necessary information for Japanese Language Education.

Keywords: verb, co-occurrence, productivity, items in basic grammar syllabus, Guiraud Index, BCCWJ