大規模コーパスを用いた用例の典型性評価
—大規模コーパスを利用した学習辞書作成のために—

千葉 庄寿（麗澤大学外国語学部）†

How to Evaluate the Typicality of the Corpus Evidence

Shoju CHIBA (Faculty of Foreign Studies, Reitaku University)

１．はじめに

よい辞書にとって，用例は不可欠の要素であることは言を俟たない。用例が提供する文脈情報の重要性はつとに多くの辞書編纂者が述べていることである(Sinclair 1991, Fox 1987)。一方で，辞書には記述スペースの問題があり，全ての用例を辞書に掲載することは難しい。掲載する用例を選択する，ないし適切な用例を編纂者が作成することは，辞書作成作業行程の重要な部分を占めることになるわけである。

用例を吟味する際，その用例をどんな人がどんな目的で参照するかを想定することは重要である。(i) 辞書を使用する人が母語話者かどうか，また (ii) 参照するのが文の理解のためか，それとも文の産出のためかによって，用例に持たせるべき情報の重みは異なるだろう。母語話者にとっては，用例はもっぱら理解，学習者にとっては産出の支援が重要になると思われるが，それに限られるわけではない。

学習者用英語辞典における長い伝統をもつ英語の辞書編纂において，用例にはさまざまな教育的配慮が求められてきた。A. S. Hornby による *Oxford Advanced Learners' Dictionary* 以来，用例は文のテンプレートとしての機能を意識して作られている。一方，Rundell (1998) が指摘するように，このような教育的配慮の結果，しばしば一つの例文に情報を盛り込みすぎ例文が不自然になってしまうこともあった。

学習者むけの日本語辞書において，用例の吟味に関する方針が深く議論されたことはこれまであまりない。優れた国語辞典である『明鏡国語辞典』には，用例の編纂方針として「日常生活で頻繁に用いられる重要語には特に用例を多く載せる」とのみ示されている(北原 2010[2]: ix)。

編集ポリシーとして，用例を実例からとるか，辞書編集者の手による作例を用いるのがよいのかは未だに意見が分かれるところである(Fox 1987, Sinclair 1991 *vs.* Laufer 1992)。しかし，学習辞書の編集にコーパスデータを援用することの意義は，今日の英語学習辞書においてコーパスを使った用例収集が既に標準的な作業となっていることを見るまでもなく明らかであり，事実上「どのようなプロセスで用例に選ばれるのかが異なる」(Rundell 1998) だけである。また，コーパスに基づく辞書であることを謳う学習辞書において，教育的配慮から，コーパスからえられた実例を編集し簡略化することもしばしばあるが，*Collins* 社の *Cobuild* シリーズの辞書のように実例そのままの提示にこだわる辞書も存在する。

† schiba@reitaku-u.ac.jp

２．重要度の認定基準

日本語学習者によって有益な例文とは何であろうか？Fox (1987)は例文を以下のような基準で評価する。

(1) 典型的であること
(2) 自然であること：“the concept of naturalness […] is the well-formedness of sentences not in isolation but in text.” (Fox 1987: 139)
(3) 示唆に富むこと
(4) 典型的な文脈とともに現れること。簡略を期して用例を短縮・修正することは望ましくない (*cf.* Fox 1987:147-149)

また，辞書の他の部品との関係として，

(5) 用例が語義の解説と連動する必要はなく，むしろ用例を語義とは独立して読み，理解することができることが重要である

ことも比較的よく知られている。

では，このような用例の重要度を測定し，大規模なコーパスからできる限り自動で取り出すために有益な情報は何だろうか？上記(2), (4)はコーパスから用例を取得することである程度保証される。一方，(3)は集まった用例を最終的に吟味する段階で確認すべきものと考えられる。本稿では，(5)の語義との連動とともに，(3)については辞書編纂者の手に委ねることを想定し，扱わないことにする。こうして，1 の「典型性」の評価のみが，特に吟味が必要な項目として残ることになる。

本稿では，用例の典型性について，以下の多様なパラメータ(6a-6c, 7)を使い，総合的に評価することを試みる。

(6) 頻度情報：
  a. 典型的な統語パターン：形態統語的情報，連鎖，統語構造
  b. 典型的なコロケーションパターン : 共起頻度とその補正値(ダイス係数, MI スコア)
  c. 典型的な文脈：文のタイプ，ジャンル情報(サブコーパス間の分布)
(7) 比較対象となるコーパスの分布からの逸脱 : BCCWJ のような大規模コーパスから取得した用例を BCCWJ 全体，さらにはサブコーパス間で比較する(対数尤度比など，サイズに影響されにくい指標を用いる, *cf.* 千葉 2011)

本稿では，以下，上述の典型性を測るための頻度情報の一部 [1] と分布の偏りの情報を使い，用例のレポジトリとして大規模なコーパスがどの程度用例の典型性評価に役立つかを示す。また，事例分析を通じて，大規模コーパスを辞書編纂に本格的に活用するための基盤構築の必要性を主張する。

[1] このうち，今回は統語構造および文のタイプに関する考察はおこなわない。

３．事例分析

本節では以下の 6 種類の語彙パターンを考察する。

(8) 動詞「生きる」(動詞-普通，上一段-カ行)
(9) 動詞「足りる」(動詞-普通，上一段-ラ行)
(10) 名詞「信用」(名詞-普通名詞-サ変可能)
(11) 名詞「信頼」(名詞-普通名詞-サ変可能)
(12) 複合助詞「かも」(助詞-副助詞＋助詞-係助詞)
(13) 派生使役動詞「動詞＋せる」(未然形＋助動詞，下一段-サ行)

このうち，(8)〜(11)は特定領域研究「日本語コーパス」言語政策班が作成した BCCWJ の語彙表 [2] において，LB_FL (図書館), PB (書籍), PM (雑誌), PN (新聞), OC (知恵袋), OY (ブログ) の 6 つのサブコーパスのうち 5-6 のサブコーパスにおいてカバー率 (累積頻度) がレベル a (0〜78%) に分類される，いずれも出現頻度が高いものである。(12)は助詞のような文法要素の複合の例，(13)は生産性の高い派生動詞の例である。

今回はテストケースとして，解析ずみの小規模なコーパスを使用して分析をおこなう。データ班が構築した「現代日本語書き言葉均衡コーパス・コアデータ」(特定領域研究「日本語コーパス」研究成果報告 DVD (JC-G-10-03 所収)) を用いる。総形態素数約 130 万語，総文数約 5 万 6 千文という小さなデータであるが，本稿が目指す方向性の出発点としては有効であると考える。

３.１　サブコーパス間の分布

以下に，今回の事例として選んだ 6 つの語彙のサブコーパス間の分布を示す。core 用例数は「現代日本語書き言葉均衡コーパス・コアデータ」での用例数を，WPM は 100 万語あたりの出現頻度を表す。

表１：「生きる」(動詞-普通，上一段-カ行)

	PB (書籍)	PM (雑誌)	PN (新聞)	OC (知恵袋)	OW (白書)	OY (ブログ)	合計
文総数	9,247	11,654	15,672	6,301	5,830	7,272	55,976
形態素総数	234,431	239,877	360,825	110,696	228,272	118,305	1,292,406
形態素数/文	25.35	20.58	23.02	17.57	39.15	16.27	23.09
core 用例数	90	41	72	7	10	22	242
WPM	383.91	170.92	199.54	63.24	43.81	185.96	187.25

表２：「足りる」(動詞-普通，上一段-ラ行)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	10	7	5	5	1	4	32
WPM	42.66	29.18	13.86	45.17	4.38	33.81	24.76

[2] 語彙レベルの詳細については田中(2011)を参照。

表３：「信用」(名詞-普通名詞-サ変可能)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	8	8	17	5	4	5	47
WPM	34.13	33.35	47.11	45.17	17.52	42.26	36.37

*PM のみ語彙レベル b

表４：「信頼」(名詞-普通名詞-サ変可能)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	18	23	35	2	17	5	100
WPM	76.78	95.88	97.00	18.07	74.47	42.26	77.38

*OC のみ語彙レベル b

表５：「かも」(助詞-副助詞＋助詞-係助詞)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	122	101	46	70	0	75	414
WPM	520.41	421.05	127.49	632.36	0.00	633.95	320.33

表６：「動詞＋せる」(未然形＋助動詞，下一段-サ行)

	PB	PM	PN	OC	OW	OY	合計
core 用例数	240	195	313	68	116	58	990
WPM	1,023.76	812.92	867.46	614.30	508.17	490.26	766.01

表から分かるように，平準化した数値をサブコーパス間で比較することで，調査語彙の中で偏りがある場合と，調査語彙間での違いが明らかになることである。例えば，類義語「信用」「信頼」を例にとると，「信用」で OW において出現数が明らかに低い(表３)のに対し，「信頼」では OC, OY など Web 系の使用域において出現数が下がっている(表４)。

さらに興味深いことに，サブコーパス間の出現比率の差は，調査語彙の文法的な抽象度が上がると小さくなる傾向があることである。使役動詞(表６)を参照されたい。

３.２　形態論的情報

以下に，今回の事例として選んだ 6 つの語彙のうち，活用を示す 2 つの動詞の活用形の分布を示す。出現する活用形の分布が大きく異なることが分かる。

表７：「生きる」(*f* = 242) の活用形

意志推量形	仮定形	終止形	未然形		命令形	連体形	連用形	総計
生きよう	生きれ	生きる	生か	生き	生け	生きる	生き	
1	1	13	1	8	3 *	**42**	**173**	242

* 「生きとし**生け**るもの」(1 例); 「**生け**る化石」(2 例)

表８：「足りる」(*f* = 32) の活用形

終止形	未然形	連体形	連用形	総計
足りる	足り	足りる	足り	
1	**29**	1	1	32

３.３　シンタグム (3-gram)

コロケーションは，用例の頻度情報として特に有益である(Sinclair 1991)。以下に，文法的要素としての性格が強い「かも」と「動詞＋せる」の特徴をよく表す連鎖の頻度を示す。

表９：「かも」(f = 414) に後続する連鎖 (1 形態素)

lemma1	lemma2	lemma3	POS3	頻度
か	も	知れる	動詞	319
か	も	。	補助記号	26
か	も	？	補助記号	11
か	も	・	補助記号	7
か	も	ね	助詞	6
か	も	…	補助記号	5
か	も	、	補助記号	5
か	も	」	補助記号	3
か	も	分かる	動詞	3
か	も	！	補助記号	3

* 頻度 3 未満は省略する。

表１０：「(か)も」に後続する連鎖 (2 形態素)

lemma1	lemma2	POS2	lemma3	POS3	頻度
も	知れる	動詞	ない	助動詞	196
も	知れる	動詞	ます	助動詞	116
も	。	補助記号	＃	文境界	24
も	・	補助記号	・	補助記号	7
も	知れる	動詞	ず	助動詞	7
も	？	補助記号	＃	文境界	7
も	ね	助詞	。	補助記号	5
も	。	補助記号	。	補助記号	2
も	」	補助記号	と	助詞	2
も	分かる	動詞	ない	助動詞	2
も	！	補助記号	？	補助記号	2

「かも」については，「知れない」といった否定表現への接続への偏りとともに，「かも」で打ち切りの形で終わる文が多いことが明らかになる。

次に，使役動詞の主動詞となる要素の出現傾向を見る(表１１)。「する」「聞く」といった出現頻度の高い動詞が出ている一方，実に多様な動詞が「せる」をとり出現していることが分かる。低頻度で数多くの種類の動詞が出現することは，この接辞の生産性の高さを示す証拠といえる(図 1, *cf.* Baayen 2001)。このような，生産性の高さに起因する多様な候補ある中で，どのような用例がより典型的と言えるかは，単純な頻度では判断することができない。後述のような，分布の偏り自体を考慮することが望ましい。

表１１：「動詞＋せる」($f = 990$) に現れる動詞

lemma1	POS1	lemma2	頻度	累積
為る	非自立可能	せる	546	546
聞く	一般	せる	19	565
済む	一般	せる	16	581
知る	一般	せる	14	595
思う	一般	せる	13	608
言う	一般	せる	12	620
持つ	一般	せる	12	632
遣る	非自立可能	せる	11	643
行う	一般	せる	9	652
楽しむ	一般	せる	9	661
走る	一般	せる	9	670
行く	非自立可能	せる	8	678
騒ぐ	一般	せる	7	685
膨らむ	一般	せる	7	692
負う	一般	せる	6	698
咲く	一般	せる	6	704
募る	一般	せる	6	710
驚く	一般	せる	5	715
思い出す	一般	せる	5	720
輝く	一般	せる	5	725
滑る	一般	せる	5	730
作る	一般	せる	5	735
飲む	一般	せる	5	740
光る	一般	せる	5	745
待つ	一般	せる	5	750
利く	一般	せる	4	754
困る	一般	せる	4	758
漂う	一般	せる	4	762
出す	非自立可能	せる	4	766
悩む	一般	せる	4	770
覗く	一般	せる	4	774
弾む	一般	せる	4	778
働く	一般	せる	4	782
振り込む	一般	せる	4	786
向かう	一般	せる	4	790
窺う	一般	せる	3	793
歌う	一般	せる	3	796
疑う	一般	せる	3	799
終わる	非自立可能	せる	3	802
通う	一般	せる	3	805
気付く	一般	せる	3	808
食う	一般	せる	3	811
死ぬ	一般	せる	3	814
ちらつく	一般	せる	3	817
尖る	一般	せる	3	820
取る	一般	せる	3	823
滲む	一般	せる	3	826
履く	一般	せる	3	829
挽く	一般	せる	3	832
含む	一般	せる	3	835
休む	一般	せる	3	838
喜ぶ	一般	せる	3	841
笑う	一般	せる	3	844

* 頻度 2 以下の動詞は省略する

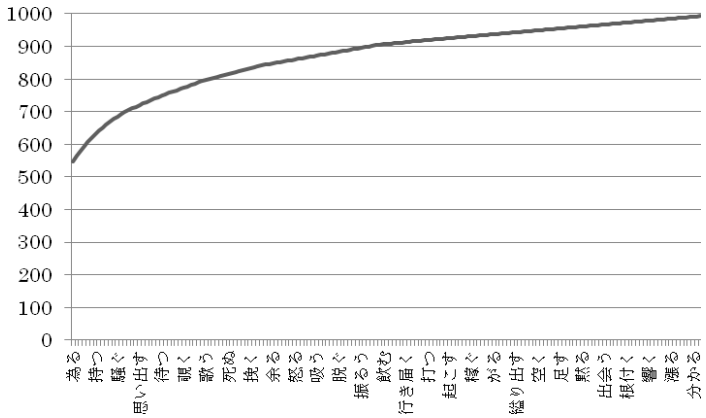

図１：「動詞＋せる」の累積頻度

一方，使役動詞の文末表現のパターンにははっきりと頻度の高い表現群があり，典型的な表現を，通常の動詞(§3.2 参照)と同様あるいはそれ以上に容易に見いだすことができる。テ形接続の一般的な表現パターンとよく似たパターンが現れる一方で，単純な頻度ではなく，後述の分布の偏り自体への考慮(§4 参照)を行うことにより，これらの中から「動詞＋せる」に特徴的な分布を見いだすことができる。

表１２：(動詞＋)「せる」(f = 990) に後続する 2 形態素

lemma1	lemma2	lemma3	POS3	頻度
せる	て	居る	動詞	68
せる	て	頂く	動詞	28
せる	て	呉れる	動詞	20
せる	て	下さる	動詞	15
せる	て	貰う	動詞	13
せる	て	、	補助記号	11
せる	て	仕舞う	動詞	9
せる	て	行く	動詞	9
せる	て	上げる	動詞	6
せる	て	見る	動詞	6
せる	て	欲しい	形容詞	4
せる	て	」	補助記号	3
せる	て	から	助詞	3
せる	て	は	助詞	3
せる	て	ます	助動詞	3
せる	て	来る	動詞	3
せる	て	遣る	動詞	3
せる	て	置く	動詞	2
せる	て	薄皮	名詞	2
せる	て	居る	動詞	2

３.４　用例の特徴を観察・分析するのに有効な情報

これまであきらかになった，典型性を測る調査パラメータと調査語彙との親和性をまとめると表１３のようになる。

表１３：分析パラメータと調査語彙との親和性

	ジャンル	語形	N-gram	語彙頻度	統語構造	係り受け
生きる		◎				
足りる		◎				
信用						
信頼	△					
かも	◎		◎			
動詞＋せる			◎			

「信用」は数十例と得られた用例数が比較的少ないため，どの情報も充分な特徴を観察するものではなかった。「かも」「動詞＋せる」は比較的用例数が多く，容易に特徴を見いだすことができた。

一方，充分な用例数が得られなくとも特徴が判別できる「足りる」のような例もある。これは「足りる」が否定極性に親和性が高いことが，日本語の場合形態素の語形から判断できることが大きいと考えられる。

今後，今回調査の対象から外した統語構造・係り受け構造および文体情報についても引き続き調査を続けていきたい。

４．語彙情報プロファイリングによる分布の分析

千葉(2011)は BCCWJ を比較のサンプルとして他のコーパスの語彙情報を評価する「語彙情報プロファイリング」の手法を開発した。BCCWJ の正式版に基づき，現在オンラインで分析できるプロファイリングシステムを公開準備中である。このシステムを用いて用例の集合の分析を試みる。

語彙情報プロファイリングをおこなうことで，コーパスから特定の語彙パターンを含む用例(コーパスのサブセット)を取得し，その例文集合と BCCWJ (ないしそのサブコーパス)の語彙情報との比較をおこなうことができる。ここではサイズの異なるコーパスの頻度情報を比較するのに有効と考えられている対数尤度比(LLR, 内山ほか 2004)を使用し，bigram (2 形態素の連鎖)の比較をおこなうことにする。

表１４：信用 (f = 47) の bigram の比較

lemma1	POS1	lemma2	POS2	頻度	BCCWJ	LLR *
信用	名詞	為る	動詞	17	494	203.573647
を	助詞	信用	名詞	11	197	142.096939
信用	名詞	出来る	動詞	5	222	55.66316
が	助詞	信用	名詞	4	73	51.508351
の	助詞	信用	名詞	5	357	50.96311
信用	名詞	の	助詞	3	80	36.405926
枠組み	名詞	合意	名詞	2	2	35.928777
た	助動詞	有権	名詞	2	5	33.098369
信用	名詞	を	助詞	3	171	31.912607

* 内山ほか (2004) による補正をおこない用例の特徴のみを正の数値として算出する

用例の集合には当然ながら「信用」と「信頼」という語彙が必ず含まれるため，全体として調査語彙が含まれる連鎖の LLR は高く出る。表１４から，47 例中 17 例(36.2%)が「信用する」という形で，5 例が「信用できる」(10.6%)という連鎖で出現したことがわかる。

一方，表１５で示すように「信頼」の場合，「信頼する」の連鎖は 100 例中 12 例(12%)で，「信用」とは明らかに出現パターンが異なる(LLR 自体は 113.5 と高い)。

表１５：信頼 (f = 100) の bigram の比較

lemma1	POS1	lemma2	POS2	頻度	BCCWJ	LLR *
の	助詞	信頼	名詞	25	793	255.309093
信頼	名詞	を	助詞	21	406	234.783454
信頼	名詞	出来る	動詞	13	241	146.376739
信頼	名詞	が	助詞	10	99	124.676046
信頼	名詞	性	接尾辞	12	441	119.033493
信頼	名詞	為る	動詞	12	557	113.512011
は	助詞	信頼	名詞	7	79	85.516645
信頼	名詞	関係	名詞	8	347	76.735319
信頼	名詞	回復	名詞	5	43	63.666965
科学	名詞	者	接尾辞	6	403	52.369004
で	助詞	信頼	名詞	4	43	49.23556
出来る	動詞	医者	名詞	3	6	45.9852
弾道	名詞	ミサイル	名詞	4	76	44.838018
対する	動詞	信頼	名詞	4	100	42.694878
信頼	名詞	感	名詞	4	107	42.164787

信頼	名詞	の	助詞	4	117	41.4639
関係	名詞	を	助詞	8	3274	41.394149
安全	名詞	で	助詞	3	17	40.535622
た	助動詞	信頼	名詞	3	21	39.359574
者	接尾辞	等	接尾辞	6	1643	35.744162
ポイント	接尾辞	減	名詞	3	40	35.686003
性	接尾辞	の	助詞	8	4818	35.433254

今後, 語彙情報プロファイリングシステムに trigram (3 形態素の連鎖)以上の長さの比較の機能を実装させることで，このような違いはさらに明確に分析抽出できるようになると思われる。

分析対象のコーパスの部分集合である用例リストの語彙情報を分析する意義は，当該語彙の特徴的な組合せを大規模なコーパスの出現比率の観点から判別できることであるが，同時に，用例集合に含まれる分析対象の語彙とは異なる語彙についても観察が可能であることを考えると非常に大きい。明らかに偏った用例集合のもつ LLR の意味，さらにはこのような手法を統計的にどのように位置づけ，典型性の判断に結びつけるかは今後さらに検討していきたい。

５．おわりに

本稿では，頻度情報と分布の偏りの情報を，典型性を測るための情報として用い，コーパスが用例の典型性評価にどの程度役立つかを具体例により検討した。今後，典型性を測るための頻度情報のパラメータを充実させるとともに，より大規模なコーパスデータを用いて検証をすすめ，大規模コーパスを辞書編纂に活用するための基盤の整備をおこなうとともに，辞書の記述とコーパス分析の方法論的な統合のための用例評価のしくみの整備，特に用例の典型性の指標化にむけた研究をすすめていきたい。

一方で，複合的な情報を如何にして組み合わせ，検索された用例の典型性，有用性を測るための指標に練り上げるか，という問題に関して，今後さらに考察する必要がある。特に，指標化はコーパスの中からの典型的用例の抽出にとどまらず，教師や辞書編纂者が作例した文を評価する，といった広い応用可能性をもつ研究成果となる可能性があり，実用化が待たれる。

今後の研究開発の方向性として，まず以下の作業をすすめることが挙げられる：

- 用例評価データベースのプロトタイプの公開
- 語彙レベル情報，統語構造（ごく浅い統語木）や係り受け情報の組み込みと評価
- コーパスの用例を評価するための諸情報の最適な組合せ方法と，指標化およびチューニング方法の検討
- 各種パラメータを総合し，指標化する用例の典型性評価データベースのプロトタイプの開発(*cf.* Kilgarriff *et al*. 2004)

一方，用例の典型性を超えた「よい例文」の判断には，コーパスの情報を使った用例の評価に加え，辞書編集者・教師による判断を併用することが必要になる可能性が高い。「有用」「重要」と辞書編纂者や教師が考える用例を登録し，学習データとして解析・蓄積する機能をもつ用例データベースの開発が望ましい。

また，辞書編纂においては語彙素が包摂する異表記と用例の出現パターンとの関係も自動処理をおこなっておくことが望ましい。例えば，『明鏡国語辞典』(北原 2010[2])では，特殊事例がゴシック体で示されている：

(14) 「重量**挙げ**【上げ・挙げ・揚げ】」「もみじの**ような**手だ【ようだ】」「誇り**高き**騎士【高い】」(北原 2010[2]: xiii)

この種の情報は，コーパスの書字形情報 [3] と語彙素情報を組み合わせることで比較的容易に抽出・分類し，辞書編纂者に事前に提示することができよう。

謝　辞

本研究は、文部科学省科学研究費補助金 基盤研究(A)「汎用的日本語学習辞書開発データベース構築とその基盤形成のための研究」（平成 23～26 年度，課題番号：23242026；研究代表者：砂川有里子）による補助を得ています。

文　献

Baayen, R. Harald (2001) *Word Frequency Distributions*. Dortrecht: Kluwer.

Béjoint, Henri (2010) *The Lexicography of English*. Oxford: Oxford University Press.

千葉庄寿 (2011)「BCCWJ の量的情報の活用：語彙情報のプロファイリングを例に」『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集, pp. 89-92.

Fox, Gwyneth (1987) “The case for examples,” in Sinclair, John M. (Ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT, pp. 137-149.

Heid, Ulrich (2008) “Corpus linguistics and lexicography,” in Lüdeling, Anke, and Merja Kytö (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Walter de Gruyter, pp. 131-153.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrž, and David Tugwell (2004) “The Sketch Engine,” in Williams, G., and S. Vessier (eds.) *EURALEX 2004 Proceedings*. Lorient: Université de Bretagne-Sud, pp. 105-116.

北原保雄(編) (2010[2])『明鏡国語辞典』第 2 版. 大修館書店.

Laufer, Batia (1992) “Corpus-based versus lexicographer examples in comprehension and production of new words,” in Tommola, H., K. Varantola, T. Salmi-Tolonen, and J. Schopp (eds.) *EURALEX 1992 Proceedings*. Tampere: University of Tampere, pp. 71-76.

Rundell, Michael (1998) “Recent trends in English pedagogical lexicography,” *International Journal of Lexicography*. 11/4: 315-342.

Sinclair, John M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

田中牧郎 (2011)「語彙レベルに基づく重要語彙リストの作成―国語政策・国語教育での活用のために―」言語政策班報告書 (JC-P-10-01), pp. 77-87. (特定領域研究「日本語コーパス」研究成果報告 DVD 所収.)

内山将夫，中條清美，山本英子，井佐原均 (2004)「英語教育のための分野特徴単語の選定尺度の比較」『自然言語処理』11/3: 165-197

Walter, Elizabeth (2010) “Using corpora to write dictionaries,” in O’Keeffe, Anne, and Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp. 428-443.

[3] UniDic では書字形情報は基本形ではなく活用していることに注意が必要である。