英語教育のための分野特徴単語の選定尺度の比較

内山　将夫[†]　中條　清美[††]
山本　英子[†]　井佐原　均[†]

特定分野の英語を効率的に学習するためには，その分野に特徴的な語彙を選定し，その語彙を学習するのが効果的である．しかし，人手による語彙の選定は，コストが高く，かつ，その質は，選定者の主観や経験に大きく依存する．そのため，特徴的な単語を機械的に抽出することにより，語彙の選定作業のコストを低くし，かつ，客観的な語彙選定を助けることは重要である．このような背景の下，本稿では，特徴的な単語が抽出可能かという観点から，各種統計的尺度の有効性，および，各種尺度を統合した複合尺度の有効性を比較検討した．その結果，主に以下のことが，明らかになった．まず，英語教育用に選定された単語リストとの一致という観点からは，複合尺度の有効性が示された．また，各種の単独尺度の中では，補完類似度が有効であった．更に，各種尺度により抽出される単語は，それぞれ，異なったレベルにおける学習者に有効であると考えられることが分かった．

キーワード：　特徴単語，統計的尺度，複合尺度，補完類似度

A Comparison of Measures for Extracting Domain-Specific Lexicons for English Education

Masao Utiyama[†] , Kiyomi Chujo[††] , Eiko Yamamoto[†]
and Hitoshi Isahara[†]

Mastery of domain-specific vocabulary in specialized English texts is essential. In order to identify a cost-effective and efficient means to extract domain-specific vocabulary, eight individual statistical measures, and combinations of those measures, were applied to corpora and the resulting lists were then compared to an existing specialized vocabulary control list. It was found that not only was it possible to efficiently produce a list of specialized vocabulary, but a combination of measures created the most comparable data. Due to the complexity of applying combinations of measures, individual measures were also found to be effective and useful for both English teachers and researchers. The complementary similarity measure was ranked as the most effective individual measure. Moreover, each measure created a unique type of word list which has specific pedagogical applications to student proficiency levels and lexicons.

KeyWords: *specialized vocabulary, statistical measure, combination of measures, complementary similarity measure*

† 情報通信研究機構, National Institute of Information and Communications Technology
†† 日本大学, Nihon University

1　はじめに

中学校および高校で学習する英語の語彙は約3000語である．しかし，ビジネス・コミュニケーション等で必要とされる語彙はもう少し大きい．その大きさを，たとえば，Test of English for International Communication[1] (TOEIC) の試験模擬問題16回分に出現した異なり単語数で推定するとすれば約5000語である (中條 2003)．

ここで，この5000語には，その重要性により順位関係がある．そのため，そのような重要性により単語を順位付け，その上位から学習することにより，効率的に語彙を増強できる．

効率的な語彙の増強のために，(中條 2003) では，英語学習者向けにTOEICにおいて重要な語彙を選定し，それが英語教育に有効なこと[2]を確認した．しかし，その選定方法は，2節で述べるように，人手が多くかかるコストの高いものであり，かつ，選定者の専門や経験等に大きく依存するものである．そのため，その方法を，TOEIC以外の分野 (たとえば，情報科学などの分野) の語彙選定に適用するためには，その分野に精通しており，かつ，英語教育の専門家であるような選定者が必要である．しかし，そのような選定者をすべての分野に求めることは困難である．

そのため，選定者に大きく依存する方法ではなく，なるべく簡単で客観性のある方法で，特定分野における重要な語彙を選定できる方法があれば，それは意義のあることである．

その方法は，基本的には，以下の2段階によると想定している．

(1)　特定分野における単語をその分野における特徴の強さ (特徴度) により順位付ける．

(2)　その上位から重要な単語を選定する．

このような2段階の方法を想定する理由は，2節で述べるように，各分野における重要な語彙とは，各分野で特徴的な単語から，更に，学習の必要性等の重要性を勘案して選定するため，まず，特徴的な単語を求めることが必要だからである．ここで，2番目の段階である，重要な単語の選定は，選定者の主観や経験等に依存するものである．この段階を機械的に実行することは，我々は想定していない．我々が機械的に精度良く実行したいのは，1番目の段階，すなわち，単語の特徴度による順位付けの段階である．なぜなら，もしこの段階における順位付けが十分に「良い」場合には，順位付けの上位において特徴単語が効率的に抽出されるため，段階2における選定を容易にすることができると考えられるからである．

しかしながら，どのような順位付けの基準が，単語の特徴度として「良い」ものであるかは現時点では明らかではない．そのため，各種の基準を比較し,「良い」基準を探す必要がある．

そのような基準として，我々は，2単語間の関連性の強さ (関連度) の測定 (Church and Hanks 1989; Manning and Schütze 1999; 山本 梅村 2002) や，各単語の特定分野における特徴度の測定 (Hisamitsu and Niwa 2001) に使われている統計的尺度を利用することを考えた．なぜなら，

1 http://www.toeic.or.jp/toeic/index.html

2 端的にいうと，TOEICにおけるスコアの向上に有効なこと．

これら各種の統計的尺度は，それぞれの応用において有効であることが報告されている尺度であるため，今回の応用においても有効であることが予想できるからである．

しかし，どの尺度がどの程度有効であるかは，実験により確かめる必要がある．そのため，本稿では，各種尺度を，特定分野における特徴単語抽出に役立つか (有効か) どうかという観点から比較する．比較の基準は，各種尺度により抽出された語彙と (中條 2003) で選定された語彙との一致の度合である．更に，本稿では，各種の尺度を単独で用いた場合だけでなく，それら尺度を組み合わせた場合についての有効性についても検討する．その組み合せの方法としては，(内山 井佐原 2003) において，専門用語抽出に対して有効性が示されている F_{cum} という方法を用いる．

まとめると，本稿での主要な目標は，各種単独尺度の特徴単語抽出への有効性を調べるとともに，F_{cum} による尺度の統合が，専門用語抽出だけでなく，特徴単語抽出にも有効であることを示すことである．

以下では，まず，2節で，本稿で考察の対象とする「分野」「特徴単語」「重要な語彙」について，その概念を規定する．そして，それ以降の節では，まず，3節で，各種単独尺度について，その説明をし，次に，4節で，それらを統合する方法について説明する．5節では，実験により各種尺度を比較し，それらの有効性について考察する．6節は結論である．[3]

2　分野と特徴単語と重要語彙についての概念規定

本節では，本稿で考察の対象としている,「英語教育のための分野特徴単語の選定」という観点から,「分野」と「特徴単語」と「重要語彙」について，その概念を規定する．

まず，英語教育では，EGP (English for General Purposes) のための教育と ESP (English for Specific Purposes) のための教育とが区別されている (深山, 野口, 寺内, 笹島, 神前 2000; Orr 2002; Noguchi 2002; Douglas 2003) ．そのため，語彙選定においても，語彙を，EGP のための語彙と ESP のための語彙とに区別することが効果的である (Nation 2001)．ここで，EGP とは,「一般的な目的のための英語教育」，すなわち,「何らかの特定の目的というものを定めずに学習する英語」についての教育であり，ESP とは,「目的限定型教育」，すなわち，個々の英語学習者が関わりあう「それぞれの学問領域や職域等で具体的目標を持って使用される英語」につい

3 本稿の内容は (中條 内山 2004) と重なる部分があるが，以下の 3 点が，主に，異なる．

(1)　(中條・内山 2004) と本稿とでは補完類似度 (後述) の適用の仕方が異なり，その結果，(中條・内山 2004) の実験では，8 つの単独尺度中で 7 位であった補完類似度の平均精度が，本稿の実験では，1 位になった．

(2)　補完類似度の適用の仕方が異なる結果として，各種尺度の統合において，補完類似度を利用した複合尺度の実験結果も異なる．

(3)　(中條・内山 2004) では各種尺度についての説明の大部分を省いているが，本稿では，それらの説明をするとともに，これまで 2 単語間の関連度の測定に使われていた，2 ベクトル間の関連度の尺度を，特徴単語の抽出に使うときにおけるパラメタの解釈を示した．

ての教育である (深山他 2000; Orr 2002; Noguchi 2002).

本稿では，ESPにおける語彙選定を想定しているので，そのときの「分野」とは,「それぞれの学問領域や職域等」のことである．つまり，この場合の分野とは，ESP教育における目的とする英語の対象により定まるものである．そのため，分野は,「genre」「register」「domain」「category」「topic」「field」「discipline」などを包括する大きな範疇を表す．たとえば,「分野」は,「工学」「語学」「医学」「農業」等という大きな専門分野を表すこともある．また，より小さな分野として,「TOEIC」や「TOEFL (Test of English as a Foreign Language) [4]」のように，大きくは「英語検定試験」という同一分野に分類されるものであっても，試験としての具体的目標がそれぞれ異なる結果として，内容的には，より下位の，小さな異なる分野であることもある (Chujo and Nishigaki 2003).

このように定義された分野，すなわち,「それぞれの学問領域や職域等」ごとに，その使用語彙が異なることは，これまでの研究で明らかにされている (中條 1991)．たとえば,「ビジネス」(石川, 田中, 高橋, 竹蓋 1987) 「医学」(Chung and Nation 2003)「経済学」(Sutarsyah, Kennedy, and Nation 1994)「計算機科学」(竹蓋, 高橋, 星野 1987) といった大きな分野で使われる語彙が異なること，更に,「TIME」「Business Week」「Newsweek」などの同じ「定期刊行物 (periodical)」の分野に属する下位の分野間でも，やはり，異なる語彙が使われていることが，定性的・定量的に示されている (中條 竹蓋 1989)．すなわち，各分野においては，分野が大きい場合でも，小さい場合でも，その分野に特有な単語が使用されている (中條 1991).

このように,「各分野に特有な単語」すなわち「それぞれの学問領域や職域等に特有な単語」を「特徴単語」と定義する．このように定義された特徴単語は，ESPにおいて学習する必要性が高い単語である．

一方，特定分野に限らず一般的に必要性の高い単語としては，たとえば，British National Corpus[5] (BNC) や Brownコーパス[6] などの，広い分野のテキストを集めた一般的なコーパスにおける頻出単語や，あるいは，中学・高校の学校英語教科書に出現する単語がある．ここで，BNCやBrownコーパスにおける頻出単語は，特定分野に限らない，一般の英語を反映した語彙であり，学校英語教科書の語彙は，学習者が将来進む分野に関わらず役立つことを目的とした語彙である．これらは，非特徴単語の典型例であり,「白色語彙」(竹蓋 1981) とも呼ばれている．白色語彙は，分野を問わずに重要と言えるため，EGP においては，学習する必要性が高い．しかし，白色語彙である単語は，本稿での考察の対象であるESPにおける特徴単語には，以下で述べる「特徴語第II類」以外は，含まれない．

ここで，特徴単語の具体的な定義は，ESPの語彙選定に関わる研究者により異なる．たとえば，(竹蓋 1981) によると，ある分野Aの「特徴語第I類」とは，Aの語彙から白色語彙を除

4 http://www.toefl.org/
5 http://www.natcorp.ox.ac.uk/
6 http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

いた語彙に含まれる単語であり,「特徴語第II類」とは，Aの語彙にも白色語彙にも共通に出現するが，相対的な出現頻度の割合からは，特にAに多く出現する単語である．また，(Nation 2001)は，頻度が高く，かつ，その分野の構成テキストに広く使われている単語(単語が出現するテキストの数をレンジと呼ぶが，そのレンジが広い単語)を特徴単語と定義している．

以上のように定義される特徴単語に基づいた，ESPにおける「重要な語彙」は，以下のように定義される．すなわち，ESPにおける「重要な語彙」とは，ESPにおける「重要単語」からなる語彙であるが，ここでの「重要単語」とは，学習者にとって,「学ぶ必要のある単語」である．具体的には，本稿の場合には，主な対象と想定している学習者は，大学生や社会人の英語初級者・中級者であるので,「中学・高校で学ばなかった単語」でかつ「学習者が必要としている特定分野で特有に用いられる単語(すなわち特徴単語)」で，さらに,「その分野の中でも必要性の高い単語」である．

そのような重要な語彙は，特定分野(たとえばTOEIC)のテキストのみから抽出された出現頻度順の単語リストのみからでは，選定することは困難である．その理由は，たとえ特定分野のみから抽出された出現頻度順の単語リストであっても，その上位には，上述した白色語彙，すなわち，どの分野にも用いられる単語が多く含まれるからである[7]．このような白色語彙を取り除き，特定分野のみに重要な語彙を選定するためには，特定分野のテキストのみを参照するのではなく，一般的な分野における語彙の出現状況を調べて，それと比較した上で，語彙を選定する必要がある．さらに，学習者に適した語彙を選定するためには，教育者の主観や経験や専門知識も必要となる．以上より分かるように，特定分野のテキストから抽出された出現頻度に基づく単語リストのみでは重要な語彙の選定には不十分である[8]．

たとえば，1節で述べたように，(中條 2003)で選定され，本稿で，各種尺度の比較の基準として利用される「TOEICにおいて重要な語彙」は，上記の(Nation 2001)と(竹蓋 1981)の定義の両方を特徴単語の定義として用い，さらに，教育的配慮など，部分的には選定者の主観による重要性の判定に基づき選定されたものである．すなわち，この語彙を選定するためには，出現頻度順の単語リストに加えて，様々な要素を考慮している．その選定の方法の概略は,

(1)　まず，(Nation 2001)の基準を適用するために,
- 各単語をTOEIC試験模擬問題における出現頻度により順位付けたリストを作成し，その上位の単語を優先し,
- 次に，より多くのTOEIC試験模擬問題に出現するレンジの広い単語を優先する．

(2)　次に，(竹蓋 1981)の観点から,

7 たとえば，本稿の実験においては，5.6節の表6のFreqの欄に，出現頻度順で抽出された上位20位以内の単語があるが，これらは典型的な白色語彙である．

8 ただし，このことは出現頻度順に基づく単語リストが不要であるということは意味しない．出現頻度順に基づくリストは，(Nation 2001)による特徴単語の定義に頻度が利用されていることからも分かるように，有用であるが，それに加えて，様々な要素を考慮する必要があるということである．

- 特徴語第I類として，TOEICでの高頻度単語から，白色語彙(ここでは中学校・高校教科書に出現する単語)を除去したものを優先し，
- 特徴語第II類として，白色語彙とTOEICに共通して使用されるが，TOEICにおいて特に顕著に使用される語彙を優先する．

(3)　更に，教育的配慮から，

- 学習の容易性
- 日本人学習者にとっての必要性

等の重要性を考慮して

選定するというものである[9]．

このように選定された語彙は，教育効果が高いものであるが，人手が多くかかるコストの高いものであり，かつ，選定者の専門や経験等に大きく依存するものである．そのため，この方法を，TOEIC以外の分野の語彙選定に適用するためには，その分野に精通しており，かつ，英語教育の専門家であるような選定者が必要である．しかし，そのような選定者をすべての分野に求めることは困難である．

そのため，本稿では，1節に述べたように，選定者に大きく依存する方法ではなく，なるべく簡単で客観性のある方法で，特徴単語を抽出するために，統計的尺度を利用することを考えた．

次節以降においては，各種尺度を比較し，その特徴単語抽出における有効性を検討する．このときの有効性の検討には，前述の，(中條 2003) で選定された語彙(特徴単語リスト)を基準として利用する[10]．すなわち，本稿においては，各種尺度の有効性は，TOEICでの特徴単語リストに基づいて判定される．しかし，本稿で比較検討する尺度自体は，特定の分野に限定されることなく，一般的に利用可能であり，かつ，有効であることを目標としている．

3　単独尺度

本節では，検討対象である8つの単独尺度について説明する．これらの尺度は，基本的な尺度である頻度に加えて，3.2節で説明する，2つの確率変数間の比較に基づく4つの尺度

- 対数尤度比 (Dunning 1993; 池田 1989)
- χ^2 値 (Hisamitsu and Niwa 2001; 池田 1989)
- イエーツ補正 χ^2 値 (Hisamitsu and Niwa 2001; 池田 1989)
- 自己相互情報量 (Church and Hanks 1989; Manning and Schütze 1999)

9 選定された語彙は667語からなるリストである．そのうち20語を無作為抽出したものが以下である：accept, bid, bill, budget, coupon, decrease, delay, estate, guarantee, lease, manufacturer, negotiate, output, passenger, physician, recession, responsibility, sale, storage, taxi. なお，前述の通り，選定された語彙には，「a」とか「the」とかの頻度上位として抽出される白色語彙は含まれない．

10 (中條 2003)で選定された語彙は，上述のように，特徴単語のリストから，更に，重要な単語を抽出したものである．つまり，本稿で対象とする重要単語は，特徴単語の一部である．そのため，以降では，特徴単語と重要単語とを特に区別しないこととし，両者をまとめて特徴単語と呼ぶ．また，それと同様に，重要語彙と特徴単語リストとを区別せずに特徴単語リストと呼ぶ．

と，3.3節で説明する，2つのベクトル間の比較に基づく3つの尺度

- コサイン (Manning and Schütze 1999)
- Dice係数 (Manning and Schütze 1999)
- 補完類似度 (澤木 萩田 1995; 山本・梅村 2002)

である．

これらの尺度は，1節で述べたように，これまで，2単語間の関連性の強さの測定や，各単語の特定分野における特徴度の測定に使われてきたものである．そのため，これらの尺度を比較することは有意義であると考える．

以下では，まず，これらの尺度を計算するために必要なパラメタ a, b,c, d を3.1節で定義し，その後で，各尺度について説明する．

3.1　各尺度に共通するパラメタ

ある特定分野における，単語 α の特徴度を計算するときに，各種尺度で利用するパラメタ a, b, c, d の定義を以下に示す．

$$
\begin{aligned}
a &= \text{特定分野におけるコーパスでの単語}\alpha\text{の頻度} \\
b &= \text{一般分野におけるコーパスでの単語}\alpha\text{の頻度} \\
c &= \text{特定分野におけるコーパスでの単語の総頻度} - a \\
d &= \text{一般分野におけるコーパスでの単語の総頻度} - b
\end{aligned}
$$

なお，$n = a + b + c + d$ とする．これらは表1のように示される．

ここで，特定分野におけるコーパスとは，たとえば，TOEICなどの分野におけるコーパスであり，一般分野におけるコーパスとは，たとえば，BNC のように，分野を限定しないようなコーパスである．

表 1　尺度計算に必要なパラメタ．

	特定分野	一般分野	
単語 α	a	b	
単語 α 以外	c	d	
			n

これらのパラメタを利用して特徴度を計算するという考え方の背景には，もし，単語 α が特定分野において特徴的な単語であれば，その単語の特定分野における出現状況は，一般分野における出現状況よりも顕著であろうという期待がある．そして，そのような顕著性を測定するために，以下で述べる尺度を利用する．まず，基本的な尺度である，特定分野における出現頻度 (Freq) を定義すると，

$$\text{Freq} = a \tag{1}$$

である．Freqが大きいような単語は，特定分野において多く出現しているので，特徴的な単語である可能性が高いと言える．以下，その他の尺度を定義する．

3.2　2つの確率変数間の比較に基づく尺度

まず，用語を定義し，次に，それを利用して各尺度(対数尤度比，χ^2 値，イエーツ補正 χ^2 値，自己相互情報量)を定義する．それらの尺度は，2つの確率変数間の依存性の程度を測定するための尺度であるが，そのうち，自己相互情報量以外は，初めに定義するままでは，特徴度の測定には不都合な点があるので，それを後で補正したものを実際には利用する．

まず，特定分野と一般分野のコーパスにおいて観測された n 個の単語からなる単語列を観測データ $D_0 = v_1, v_2, \ldots, v_n$ とする．ただし，特定分野のコーパスは $1 \leq i \leq a + c$ なる v_i からなり，一般分野のコーパスは $a + c + 1 \leq i \leq n$ なる v_i からなるとする．

次に，単語 α の特徴度を測定したいとして，ある観測された単語 v についての，二つの確率変数 W, T を以下のように定義する．

$$W = \begin{cases} 1 & \text{if } v = \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$T = \begin{cases} 1 & \text{if } v \text{の出現個所が特定分野のコーパスである}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

このように定義すると，単語 v_i に対応する確率変数 W, T の値を，それぞれ，w_i, t_i としたとき，データ D_0 からは，次のような確率変数の値の列 D が得られる．

$$D = \langle w_1, t_1 \rangle, \ldots, \langle w_n, t_n \rangle \tag{4}$$

そして，ある仮設 H を条件として，D の確率を計算すると，以下になる．

$$\Pr(D|H) = \prod_{i=1}^{n} \Pr(W = w_i, T = t_i|H) \tag{5}$$

ただし，2単語 v_i と v_j とは確率的に独立であると仮定する．

次に，2つの仮設 H_{indep}, H_{dep} を考える．それらは以下である．

H_{indep} : 確率変数 W と T とは互いに独立である．

H_{dep} : 確率変数 W と T とは互いに依存(従属)している．

H_{indep} の下では次が成り立つ．

$$\Pr(W = w, T = t|H_{indep}) = \Pr(W = w|H_{indep}) \Pr(T = t|H_{indep}). \tag{6}$$

ただし，$w, t = 1, 0$.

最後に，各種の確率の値を，表1を修正した表2に基づき，最尤推定で推定すると以下になる．

$$\begin{aligned}
\Pr(W=1|H_{indep}) = \Pr(W=1|H_{dep}) &= \frac{a+b}{n} \\
\Pr(W=0|H_{indep}) = \Pr(W=0|H_{dep}) &= \frac{c+d}{n} \\
\Pr(T=1|H_{indep}) = \Pr(T=1|H_{dep}) &= \frac{a+c}{n} \\
\Pr(T=0|H_{indep}) = \Pr(T=0|H_{dep}) &= \frac{b+d}{n} \\
\Pr(W=1, T=1|H_{dep}) &= \frac{a}{n} \\
\Pr(W=1, T=0|H_{dep}) &= \frac{b}{n} \\
\Pr(W=0, T=1|H_{dep}) &= \frac{c}{n} \\
\Pr(W=0, T=0|H_{dep}) &= \frac{d}{n}
\end{aligned}$$

表 2　確率変数の値とパラメタの関係.

	$T=1$	$T=0$	
$W=1$	a	b	$a+b$
$W=0$	c	d	$c+d$
	$a+c$	$b+d$	n

以上の準備の下で，2変数間の依存性の度合を測定するための尺度である，対数尤度比，χ^2 値，イエーツ補正 χ^2 値を定義する[11]が，その前に，なぜ依存性の度合を測定することにより特徴度が測定できるかについて述べる．ただし，ここでの依存性の度合とは，2変数が独立でない程度の大きさということであるので，まず，独立の場合の状況を考える．

さて，2変数が独立であるときには，2変数は互いに影響を与えない．そのため，$T=1$ であろうと $T=0$ であろうと，W の値には影響しない，すなわち，$\Pr(W=w|T=1, H_{indep}) = \Pr(W=w|T=0, H_{indep}) = \Pr(W=w|H_{indep})$ である．

これが意味することは，コーパスが特定分野のものであろうと一般分野のものであろうと，単語 α の出現確率は変化しないということである．したがって，独立である場合には，単語 α は，特定分野にも一般分野にも同様に出現するので，このような単語は，特徴的な単語とは言えない．一方，独立でない程度が大きい (依存性の程度が大きい) ときには，単語 α の出現確率は，特定分野と一般分野とで大きく変わるので，そのような単語は，特徴的な単語の候補となりうる[12]．そのため，依存性の度合を測定することにより，単語の特徴度を測定できる．

11 自己相互情報量は，これら3つとは異なる量を測定するので，後で別に説明する．

12 ただし，単に出現確率が変わるだけでは，一般分野において高確率で出現する単語が，特定分野において低確率で出現する場合もありうるので，そのような場合を排除し，一般分野において低確率で出現する単語が，特定分野において高確率で出現するときに高い値を取るようにする必要がある．しかし，対数尤度比，χ^2 値，イエーツ補正 χ^2 値は，これら2つの場合を区別できないので，それを修正する必要がある．そのための手段は後述する．

対数尤度比

対数尤度比 (Log-Likelihood Ratio, LLR_0) は以下のように定義される．

$$\mathrm{LLR}_0 = \log \frac{\Pr(D|H_{dep})}{\Pr(D|H_{indep})} \tag{7}$$

(5) 式と (6) 式および最尤推定された確率に基づき上式を展開すると以下になる．

$$\begin{aligned}
\mathrm{LLR}_0 &= \sum_{i=1}^{n} \log \frac{\Pr(W=w_i, T=t_i|H_{dep})}{\Pr(W=w_i, T=t_i|H_{indep})} \\
&= a \log \frac{\Pr(W=1, T=1|H_{dep})}{\Pr(W=1, T=1|H_{indep})} + b \log \frac{\Pr(W=1, T=0|H_{dep})}{\Pr(W=1, T=0|H_{indep})} \\
&+ c \log \frac{\Pr(W=0, T=1|H_{dep})}{\Pr(W=0, T=1|H_{indep})} + d \log \frac{\Pr(W=0, T=0|H_{dep})}{\Pr(W=0, T=0|H_{indep})} \\
&= a \log \frac{an}{(a+b)(a+c)} + b \log \frac{bn}{(a+b)(b+d)} \\
&+ c \log \frac{cn}{(c+d)(a+c)} + d \log \frac{dn}{(c+d)(b+d)}
\end{aligned} \tag{8}$$

対数尤度比は，(7) 式に示されるように，2変数 W と T とが依存しているという条件 (H_{dep})，および，独立であるという条件 (H_{indep}) の下で，データ D が観測される確率の比の対数である．したがって，対数尤度比が大きいときには，2変数は互いに依存している可能性が高い．そのため，対数尤度比は，特徴度の測定に利用できる．なお，LLR_0 は，H_{indep} が完全に成立するとき[13]には 0 となり，それ以外の場合には 0 より大きくなる．すなわち，$\mathrm{LLR}_0 \geq 0$ である．

χ^2 値

χ^2 値 ($\mathrm{Chi2}_0$) は，H_{indep} と H_{dep} における期待頻度を比較することにより，依存性の度合を測定する．その値は，両者の期待頻度の差 (これは依存性の度合に相当する) が大きいほど大きい値をとるため，依存性の度合の測定，つまり，特徴度の測定に利用できる．

まず，用語を準備する．ある仮設 H を条件としたとき，$\langle W=w, T=t \rangle$ という事象の，n 回の観測における期待頻度 f は以下である．

$$f(W=w, T=t|H) = n \Pr(W=w, T=t|H) \tag{9}$$

このときデータ D から計算される χ^2 値は

$$\mathrm{Chi2}_0 = \sum_{w,t=0,1} \frac{\{f(W=w, T=t|H_{dep}) - f(W=w, T=t|H_{indep})\}^2}{f(W=w, T=t|H_{indep})} \tag{10}$$

である．これを (6) 式と (9) 式および最尤推定された確率に基づいて展開して整理すると以下になる．

$$\mathrm{Chi2}_0 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{11}$$

13 $\Pr(W=w, T=t|H_{dep}) = \Pr(W=w, T=t|H_{indep})$ のとき．

Chi2_0は，H_{indep}が完全に成立するときには0となり，それ以外の場合は0より大きくなる．すなわち，$\text{Chi2}_0 \geq 0$である．Chi2_0が大きいほど依存性の度合は高い．

イエーツ補正χ^2値

χ^2値は，期待頻度が小さいときには，依存性の度合を測定する信頼性が低くなる．そのため，χ^2値をイエーツの公式により補正した値(Yates_0)として，以下がある．

$$\text{Yates}_0 = \frac{n(|ad-bc| - \frac{n}{2})^2}{(a+b)(c+d)(a+c)(b+d)} \tag{12}$$

ここでも，$\text{Yates}_0 \geq 0$である．

上記3尺度の補正

これまで，対数尤度比(LLR_0)，χ^2値(Chi2_0)，イエーツ補正χ^2値(Yates_0)を定義した．これらの尺度は，依存性の度合を測定することはできる．つまり，ある単語について，その単語の出現確率が特定分野と一般分野とで異なる度合を測定することはできる．しかし，これらの尺度は，特定分野で高確率となる場合と低確率になる場合とを区別できない．たとえば，表1の形式で，

10	1
1	10

と

1	10
10	1

とは同じ値となる．これらは，依存性の程度は同じであるが，前者は，特定分野で高確率であり，後者は，特定分野において低確率となっている．

そのため，特定分野において，一般分野よりも，高確率で出現するような場合には正の値をとり，低確率で出現する場合には負の値をとるように補正する(影浦 1997)．ここで，より高確率で出現するとは$\frac{a}{c} > \frac{b}{d}$，つまり，$ad-bc>0$が成立することであるので，以下のように，LLR, Chi2, Yatesを定義し，それらを特徴度の測定に利用する．

$$\text{LLR} = \text{sign}(ad-bc) \times \text{LLR}_0 \tag{13}$$

$$\text{Chi2} = \text{sign}(ad-bc) \times \text{Chi2}_0 \tag{14}$$

$$\text{Yates} = \text{sign}(ad-bc) \times \text{Yates}_0 \tag{15}$$

ただし，

$$\text{sign}(z) = \begin{cases} +1 & \text{if } z > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{16}$$

である．

自己相互情報量

自己相互情報量(Pointwise Mutual Information, PMI)は以下で定義される．

$$\text{PMI} = \log \frac{\Pr(W=1, T=1|H_{dep})}{\Pr(W=1, T=1|H_{indep})} \tag{17}$$

これを (6) 式および最尤推定された確率に基づき展開すると以下になる．

$$\mathrm{PMI} = \log \frac{an}{(a+b)(a+c)} \tag{18}$$

(17) 式より分かるように，自己相互情報量も，前記3尺度と同様に，2変数WとTとの依存性が高いときに大きい値となる．ただし，前記3尺度と異なり，特定分野において低確率となるような場合 (aが小さくなりbとcとが大きくなる場合) には低い値となり，高確率となるような場合 (aが大きくなりbとcとが小さくなる場合) には高い値となるので，補正の必要はない．

自己相互情報量は，2単語間の関連度の測定に使われてきた (Church and Hanks 1989) が，低頻度語間の関連度を過大評価することが知られている (Manning and Schütze 1999)．

3.3　2つのベクトル間の比較に基づく尺度

本節で定義する，コサインと Dice 係数と補完類似度は，2つの 01-ベクトル間の類似度を測定するための尺度である．そのため，まず，類似度計算用の2つのベクトルを定義する必要がある．それには，(4) 式で定義したデータ D を用いる．D における w_i と t_i とをそれぞれの要素とするベクトル $\overrightarrow{W}$ と $\overrightarrow{T}$ とを以下のように定義する．

$$\begin{aligned} \overrightarrow{W} &= [w_1, w_2, \ldots, w_n] \\ \overrightarrow{T} &= [t_1, t_2, \ldots, t_n] \end{aligned}$$

これらを模式図に示すと図1のようになる．

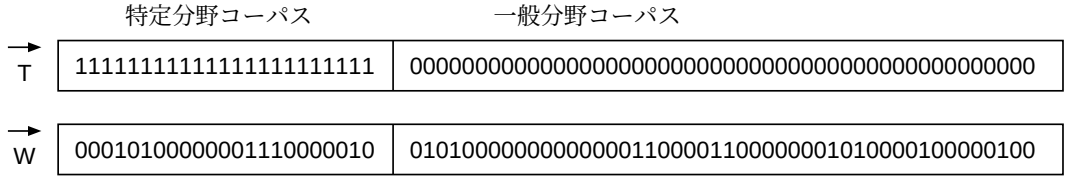

図 1　ベクトルの模式図.

$\overrightarrow{W}$ は，特定分野と一般分野のコーパスにおいて，注目している単語 α が出現した位置が1となり，それ以外が0となるようなベクトルである．つまり，$\overrightarrow{W}$ は α の出現位置を表現している．

一方，$\overrightarrow{T}$ は，特定分野のコーパスにおける全位置が1であり，一般分野における全位置が0であるようなベクトルである．このベクトルは，その特定分野を代表するベクトルと解釈できる．

したがって，本節で定義する尺度は，単語に関するベクトル $\overrightarrow{W}$ と，特定分野の代表ベクトル $\overrightarrow{T}$ との類似度を測定し，それを特徴度とみなしていると解釈できる．

コサイン

コサイン (Cosine) は代表的な類似尺度であり，以下のように定義される．

$$\text{Cosine} = \frac{\overrightarrow{W} \cdot \overrightarrow{T}}{|\overrightarrow{W}||\overrightarrow{T}|} = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{19}$$

コサインは，ベクトル間の角度に応じて値が変化し，その角度が狭いベクトル間において高い値となる．

Dice 係数

Dice 係数 (Dice) は以下の式で定義される．

$$\text{Dice} = \frac{2I(\overrightarrow{W}\&\overrightarrow{T})}{I(\overrightarrow{W}) + I(\overrightarrow{T})} = \frac{2a}{(a+b)+(a+c)} \tag{20}$$

ただし，I(ベクトル) はそのベクトルにおける 1 の数である．また，$\overrightarrow{W}\&\overrightarrow{T} = [w_1 t_1, w_2 t_2, \ldots w_n t_n]$ とする．

Dice 係数は，ベクトルというよりは，むしろ，集合に基づいた尺度であり，共通要素数が多い集合間において高い値となる．

補完類似度

補完類似度 (Complementary Similarity Measure, CSM) は以下の式で定義される．

$$\text{CSM} = \frac{ad - bc}{\sqrt{(a+c)(b+d)}} \tag{21}$$

補完類似度は，元々文字認識の分野で提案された尺度である (澤木・萩田 1995) が，最近，2 単語間の上位下位関係などの 1 対多関係の推定のための尺度としても有効であることが示されている (山本・梅村 2002)．

文字認識における補完類似度の利用法は，テンプレートとなるベクトルを各文字について用意し，与えられた入力文字について，各文字のテンプレートベクトルとの補完類似度を計算し，最も高い値となる文字テンプレートを認識文字とする方法である．この利用法は，入力文字が帰属する文字テンプレートを求める，というものである．

本稿においては，テンプレートとなるベクトルは，特定分野の代表ベクトル $\overrightarrow{T}$ であり，認識対象とみなせるベクトルは，単語についてのベクトル $\overrightarrow{W}$ である．そして，$\overrightarrow{W}$ が $\overrightarrow{T}$ に帰属する度合を補完類似度で測定し，それを特徴度とする．このように，本稿における補完類似度の使い方は，文字認識における使い方とほぼ同様であると解釈できる．

文字認識においては，補完類似度は高性能であることが報告されているので，本稿における

特徴単語の抽出においても高性能であることが期待できる．[14]

4　複数尺度の組み合せ方法

本節では，複数尺度を組み合わせて1つの複合尺度とするF_{cum}という方法について述べる．この方法は，(内山・井佐原 2003)において，専門用語の抽出に有効なことが示されている．本稿での目標の1つは，F_{cum}による複数尺度の統合が，特徴単語抽出にも有効であることを示すことである[15]．

4.1　統合における問題設定

まず，尺度値を割当てたいような単語の集合$X = \{x_1, x_2, \ldots, x_l\}$について，$X$の要素には，特徴度により全順序関係が定義できるとする．そしてある関数gがあり，以下が成立するとする．

$$g(x_i) < g(x_j) \iff x_i \text{よりも} x_j \text{が特徴的である．}$$
$$g(x_i) = g(x_j) \iff x_i \text{と} x_j \text{が同等に特徴的である．}$$

我々の目的は，このgを推定することである．

ここで，我々の手元には，$f_1, f_2, \ldots, f_m$というm個の関数(尺度)があり，それぞれ，gを近似しているとする．つまり，f_iについて，$f_i(x_j) < f_i(x_k)$ ならば $g(x_j) < g(x_k)$が，ある程度は成立しているとする．ここで，ある程度成立しているとは，たとえば，Xをgにより順位付けた場合とf_iにより順位付けた場合とで，それらの間の順位相関がある程度は高いということである．

このようなm個のf_iからgを推定するのが，我々の目的であり，本節で対象とする問題設定である．

---

14 補完類似度と他の7つの尺度とを比べた場合の顕著な相違は，他の尺度においては，表1の形式で，
a	b
c	d

と
a	c
b	d

とが同一の値になるのに対して，補完類似度においては，両者が異なる値となることである．つまり，CSM′を$\mathrm{CSM}' = \frac{ad-bc}{\sqrt{(a+b)(c+d)}}$と定義すると，CSM ≠ CSM′である．CSMは，$\overrightarrow{T}$をテンプレートとみなしたときの補完類似度の式であり，CSM′は，$\overrightarrow{W}$をテンプレートとみなしたときの補完類似度の式である．ここで，文字認識での補完類似度の利用方法からの類推により補完類似度を定義するとすると，CSMを補完類似度の式とすべきである．しかし，(中條・内山 2004)では，本稿とは異なり，特にパラメタの意味を考えずに形式的に補完類似度の式を利用していたため，CSMとCSM′のどちらを補完類似度の式として利用するべきかが不明であった．そのため，(中條・内山 2004)では，$\max(\mathrm{CSM}, \mathrm{CSM}')$を補完類似度の式として利用していた．このように最大値を取るという方法は，2単語間の関連度の測定のように，どちらの単語がテンプレートかが不明な場合には適当であるが，本稿での場合における，注目している単語に関するベクトル$\overrightarrow{W}$と，その帰属先となりうる特定分野のコーパスに関するベクトル$\overrightarrow{T}$のように，$\overrightarrow{T}$がテンプレートとみなせる場合には，不適当な使い方である．そのため，(中條・内山 2004)における補完類似度$\max(\mathrm{CSM}, \mathrm{CSM}')$の平均精度は，単独尺度中7位と低い．しかし，本稿の実験においては，補完類似度の式としてCSMを利用することにより，その平均精度は，単独尺度中1位となった．

15 (内山・井佐原 2003)においては，F_{mul}と呼ばれている統合法も述べられているが，F_{cum}とF_{mul}とでは，5節の実験の結果ではほとんど差がなく，かつ，F_{cum}の方が若干優れていたため，本稿では，F_{cum}による統合についてのみ述べる．

4.2　統合方法

本節で述べる方法は，経験確率に基づいたものであり，簡明なものである．すなわち，以下の (22) 式の $F_{cum}(x)$ をもって，$g(x)$ の推定値とする．

まず，必要な用語を，X と f_i に加えて，定義する．それらは以下のようである．

$$N(x) = x(\in X) \text{ の特定分野における出現頻度}$$

$$N = \sum_{x \in X} N(x)$$

$$R(x) = N(x)/N$$

$$[f_i(x') \leq f_i(x)] = \begin{cases} 1 & \text{if } f_i(x') \leq f_i(x) \\ 0 & \text{otherwise} \end{cases}$$

$R(x)$ は x の経験確率である．このとき，

$$F_{cum}(x) = \sum_{x' \in X} R(x') \prod_{i=1}^{m} [f_i(x') \leq f_i(x)] \tag{22}$$

である．ここで，$\prod_{i=1}^{m}[f_i(x') \leq f_i(x)]$ は，全ての f_i について，$f_i(x') \leq f_i(x)$ のときに，1 となり，それ以外では 0 となる．(22) 式は，その値が 1 となるような x' についてのみ，$R(x')$ を足した値である．

以下では，F_{cum} を例により説明するが，その前に，2 要素 x と x' の比較のための用語を定義する．まず，$x \equiv x'$ とは，全ての f_i について，$f_i(x) = f_i(x')$ であり，$x \prec x'$ とは，全ての f_i について，$f_i(x) < f_i(x')$ であり，$x \preceq x'$ とは，全ての f_i について，$f_i(x) \leq f_i(x')$ であると定義する．そして，$x \preceq x'$ か $x' \preceq x$ のいずれかが成立するような x と x' とは「比較可能」であると言い，比較可能でないときには,「比較不能」であると言う．また，$x \preceq x'$ であるとき，x は x' より「劣位」であり，x' は x より「優位」であると言う．

$F_{cum}(x)$ とは，x より劣位なものについての経験確率の和である．

F_{cum} による統合の性質をみるために，まず，$n = 1$，つまり，尺度が 1 つの場合をみてみる．このとき，

$$F_{cum}(x) = \sum_{x' \in X} R(x')[f_1(x') \leq f_1(x)]$$

である．この場合，X の要素を適当に並べかえて，$i \leq j$ なら，$F_{cum}(x_i) \leq F_{cum}(x_j)$ となるようにできて，そのときの $f_1(x)$ と $F_{cum}(x)$ との関係は図 2 のようになる．

図 2 より分かるように，$F_{cum}(x)$ は，$f_1(x)$ の階段関数である．このとき，$F_{cum}(x)$ は，x より劣位なるものの累積経験確率となっている．したがって，たとえば，$F_{cum}(x) = 0.95$ のときには，$x \prec x'$ なる x' の出現確率は $\sum_{x \prec x'} \Pr(x') = 1 - F_{cum}(x) = 0.05$ である．つまり，x より，尺度 f_1 の観点から特徴的な要素 x' が出現する確率は 0.05 である．

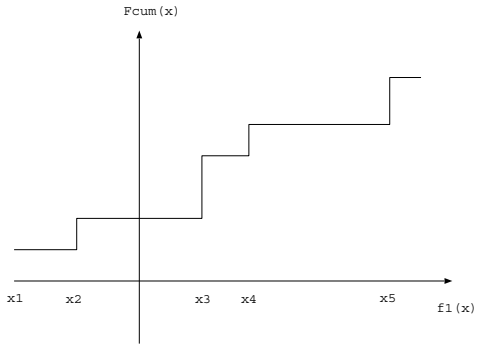

図 2　$f_1(x)$ と $F_{cum}(x)$ との関係.

なお，$n = 1$ のときには，

$$F_{cum}(x) \leq F_{cum}(x') \Longleftrightarrow f_1(x) \leq f_1(x')$$

が成立している．したがって，$n = 1$ のときには，$F_{cum}(x)$ と $f_1(x)$ とは順序尺度の観点からは同等なので，$F_{cum}(x)$ を $f_1(x)$ の代りに利用することは問題ない.

次に，$n = 2$ の場合を考える．まず，図3のように，x_1 と x_2 とが比較可能な場合を考える．このとき，図中の○により X の要素を指すとすると，$x_1 \preceq x_2$ であるので，$F_{cum}(x_1) \leq F_{cum}(x_2)$ である．これは，$i = 1, 2$ について，$f_i(x_1) \leq f_i(x_2)$ であるので，こうなって当然であり，問題はない.

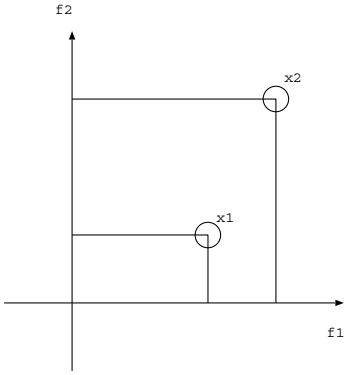

図 3　x_1 と x_2 とが比較可能な場合.

次に，図4のように，x_1 と x_2 とが比較不能な場合を考える．このとき，$f_1(x_1) < f_1(x_2)$ かつ $f_2(x_1) > f_2(x_2)$ である．このように，尺度間において，二つの対象に与える尺度値の順位に矛盾がある場合にも，F_{cum} を利用すれば，一貫した順位を付けることができる．たとえば，図4には，x_1, x_2 の他に，7点が出現し，それぞれの頻度が全て1回ずつだとすると，$F_{cum}(x_1) = \frac{5}{9}$,

$F_{cum}(x_2) = \frac{3}{9}$ である．したがって，$F_{cum}(x_1) > F_{cum}(x_2)$ であるので，F_{cum} の観点からは，x_1 の方が特徴的な要素である．$F_{cum}(x)$が大きい値となるのは，$x' \preceq x$ なる x' が多いときである．

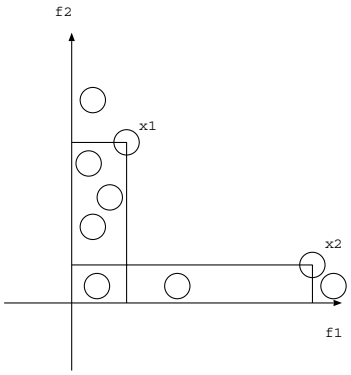

図 4　x_1 と x_2 とが比較不能な場合.

これまでに，$n = 1, 2$ の場合をみたが，$n \geq 3$ についても，同様に F_{cum} を計算できる．F_{cum} を使うことにより，他と比較して優位な要素が，特徴的な要素として抽出される．

5　実験

本節では，3節で述べた単独尺度，および，4節の方法により統合された複合尺度の，特徴単語抽出に関する有効性を調べる．以下では，まず，実験に用いた材料と，各種尺度の有効性の評価基準とについて述べる．次に，各種尺度による単語の順位付けの性質について考察したあとで，単独尺度と複合尺度の有効性を検討する．最後に，各種尺度により抽出される単語の実例を調べることにより，各種尺度が有効な状況について考察する．

5.1　実験材料と評価基準

3節で述べた各種の単独尺度を計算するためには，3.1節で述べたパラメタ a,b,c,d を計算する必要がある．そのためには，一般分野および特定分野のコーパスにおける，単語の出現頻度付きのリストが必要である．そのコーパスやリストとしては，以下を用いた．

- 一般分野のコーパスとしては BNC を用い，そこから，頻度 100 以上の異なり単語 13995 語 (延べ単語数=86112272) を抽出したリスト (Chujo 2004) を用いた．以下では，この単語リストを「BNC リスト」と呼ぶ．
- 特定分野のコーパスとしては，TOEIC 試験模擬問題 16 回分より作成した単語リスト (中條 2003) を用いた．このリストの異なり単語数は 5016 語であり，延べ単語数は 107077 語である．なお，このリストには，BNC リストとは異なり，頻度 1 回以上の単語が含まれ

ている．以下では，この単語リストを「TOEICリスト」と呼ぶ．

なお，これらの単語リストには固有名詞や数詞等は含まれていない．また，同一の基本形を持つ単語については，基本形のみをリストに登録し，その頻度は各屈折形の頻度の合計としている[16]．

次に，各種尺度の精度評価の基準となる単語リストとしては，2節で述べたように，

- (中條 2003)で選定された，TOEIC向けの英語教育用の667語からなるリストを用いた．以下では，この単語リストを「正解リスト」と呼ぶ．なお，正解リストは，TOEICリストに包含される．

ここで，各種尺度の精度を評価するときには，TOEICリスト中の単語を各種尺度で順位付けたリスト(候補語リスト)を作成し，そのリストと正解リストとの一致の度合により評価する．そのときの評価の基準としては，平均精度(Average Precision, AP)を利用した．平均精度は，情報検索の評価(Baeza-Yates and Ribeiro-Neto 1999)や連語抽出の評価(Schone and Jurafsky 2001)にも使われていて，上位に正解が多いほど高い値を取るので，今回の評価の指標として適切である．

平均精度を求めるためには，候補語リストの上位から順に候補語を調べて，それが正解であったときには，そのときの順位精度(その順位をrとしたとき，それまでの順位での正解の個数をlとすると，$\frac{l}{r}$)を求めて，リストの最後の時点で，それまでに得られた正解における順位精度の平均を求めれば良い．その定義式は以下である．

$$\text{平均精度} = \frac{1}{K}\sum_{l=1}^{K}\frac{l}{r_l} \tag{23}$$

ただし，Kは，候補語リスト中の全正解数であり，$\frac{l}{r_l}$ は，l番目の正解が抽出されたときの順位精度であり，r_lは，l 番目の正解が抽出されたときの，候補語リストにおける順位である．なお，同一尺度値の単語が複数ある場合には，無作為にタイ(tie, 結び)を解消した．

5.2　各尺度による順位付けの性質

ここで，どのような単語を抽出すれば正解となるかの目安を得るために，正解リスト(Reference)とTOEICリスト中の単語の頻度分布を比較する(図5参照)．また，図5の各頻度分布において，1/4, 1/2(中央値), 3/4に位置する単語の頻度を表3に示す．

これらの図と表より分かるように，正解リストには，TOEICリストと比べて，あまり低頻度であるような単語は含まれていない．そのため，効率的に正解を抽出するためには，そのような低頻度な単語を避けることができる尺度が必要である[17]．

16 たとえば，基本形 study に対して，屈折形は study, studies, studied, studying があり，それぞれの頻度が 200, 100, 40, 20なら，リストには，基本形のstudyのみを登録し，その頻度は各屈折形の頻度の合計360(= 200 + 100 + 40 + 20)としている．

17 正解リストには，低頻度単語だけでなく，高頻度単語もあまり含まれていないので，高頻度単語を避けることも必要で

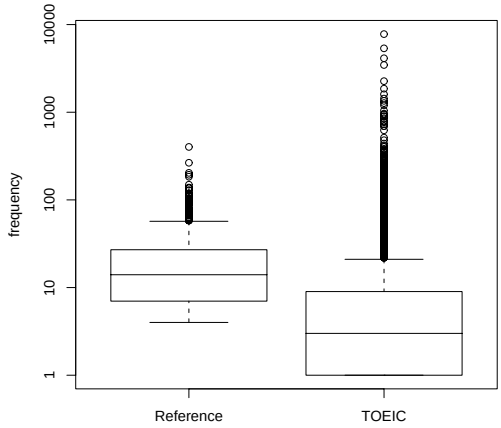

図 5　正解リスト (Reference) と TOEIC リストの頻度分布の比較.

表 3　正解リストと TOEIC リストの頻度分布の比較.

	1/4での頻度	中央値 (1/2)	3/4での頻度
正解リスト	7	14	27
TOEIC リスト	1	3	9

そこで，各種尺度が低頻度単語をどう順位付けるかを調べた．ここで注目する低頻度単語は，表1において，$a = 1, b = 0$ となるような単語である．このような単語は，TOEIC リストにおいて頻度が 1 であり，かつ，BNC リストには存在しない[18]ような単語である．そのような単語は全部で 404 語ある．

ここで，$a = 1, b = 0$ ということは，一般分野の BNC リストに存在しなかったような単語が，特定分野の TOEIC リストには存在する，ということであるので，特徴的な単語の候補とは言える．しかし，TOEIC リストにおける出現頻度が 1 であるので，それを特徴単語と言うには，信頼度が低い[19]．そのため，このような単語は避けるべきである．なお，以下では，$a = 1$, $b = 0$ の単語 (からなる集合) を W_{10} と記す.

あるが，高頻度単語の異なり語数は，低頻度単語の異なり語数よりも，ずっと少ないため，高頻度単語を避けるより，低頻度単語を避けることの方が，相対的に，重要である．

18 BNC リストは頻度 100 以上の単語のみからなるので，BNC リストにないからといって，元コーパスにおける出現頻度が 0 とは限らない.

19 まず，2節で述べた (Nation 2001) による特徴単語の定義として，当該分野における出現頻度が高いことがある．この観点からは，$a = 1, b = 0$ の単語は，当該分野 (TOEIC) での出現頻度 a が 1 であるので，特徴単語と言うには，出現頻度が低い．次に，効率的な学習という観点からは，語彙に加える単語としては，当該分野の多くのテキストに出現する単語の方が，出現頻度の少ない単語よりも望ましい．この観点からも，$a = 1$,$b = 0$ という単語は，当該分野での出現頻度が 1 であるので，たとえ，BNC リストに出現していない ($b = 0$) としても，特徴単語として語彙に追加する必要性は低い．また，$a = 1$,$b = 0$ である 404 語を実際に調べてみたところ，その中には，複合語や派生語など，その意味が類推可能なため，語彙に加える必要性が少ない単語が多かった．たとえば，rear-seat, cookbook, misrepresent, rapidness などの意味は類推可能である．以上より，$a = 1, b = 0$ の単語は,「特徴単語と言うには信頼度が低い」，あるいは,「特徴単語として語彙に追加する必要性が低い」と言える.

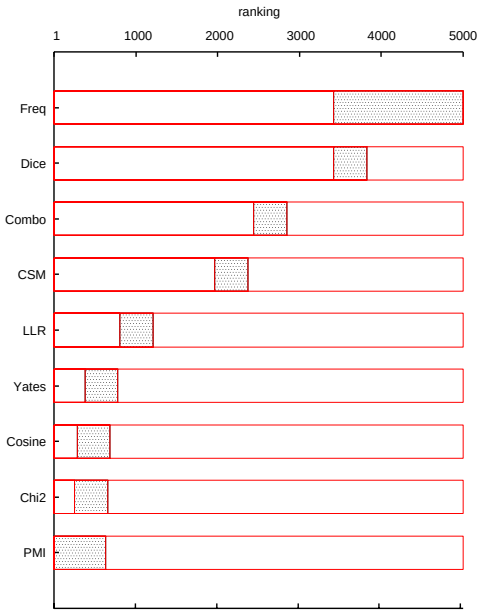

図 6　低頻度単語の順位.

図6は，W_{10}と同スコアである単語の順位を各尺度について示したものである．それらは図において色付けされている部分の順位を占めている．図では，左端が1位で右端が5016位である．また，上から下に，各尺度が並んでいる．ただし，Freq等は，3節で述べた単独尺度であり，Comboは，4節で述べた方法により，5つの尺度(CSM, PMI, Cosine, Dice, Yates)を統合したものである．Comboとしてこれらの尺度の組み合せを選んだ理由は，5.4節で述べるように，任意の尺度の組み合わせのなかで最高の平均精度であったからである．

図6において，FreqとPMI以外は，W_{10}と同スコアのものはW_{10}以外には存在しない．それらは，全順位の8.1%($= \frac{404}{5016} \times 100$)を占めている．一方，Freqにおいては，$a = 1$であるものは同スコアであり，その数は1588 (31.7%)である．また，PMIにおいては，$b = 0$であるものが同スコア[20]であり，その数は630 (12.6%)である．

図6より分かるように，PMIでは，W_{10}のような低頻度語が最高順位となる．一方，図5より分かるように，正解リスト(Reference)にはこのような低頻度語は含まれていない．そのため，PMIは，正解リスト中の単語を抽出するには不向きであると予想できる．

次に，3.2節で説明した，依存性の度合を測定するための尺度である，LLR, Yates, Chi2 に注目すると, LLRにおけるW_{10} の順位が一番低く，Yatesが中間で，Chi2における順位が一番高い．このうち，YatesとChi2については，Yatesは低頻度のときにChi2を補正するためのものであるので，YatesがChi2より低頻度単語の順位を低くすることは説明できる．しかし，こ

20 (18)式において，$b = 0$とすると，PMI $= \log \frac{n}{(a+c)}$であり，nと$a + c$とは定数なので，$b = 0$ の単語はすべて同スコアとなる．また，このとき，PMIは最大値になる．

れは比較的ということであり，LLR に比べると Yates も低頻度単語を高順位としている．また，LLR については，(Dunning 1993) で述べられているように，低頻度のときには Chi2 よりも信頼できる尺度である．図 6 は，それらを反映している．なお，Cosine は，Chi2 同様，低頻度単語を高順位にすると言える．

Freq については，$a = 1$ は最低順位となる．また，Dice についても W_{10} の順位は低い．Combo と CSM は，ちょうど中間に W_{10} がある．このことから，これらの尺度が，Freq や Dice と同様に，出現頻度が低いものは，あまり，過大評価しない尺度であると言える．

図 6 をまとめると，Yates, Cosine, Chi2, PMI は，低頻度単語である W_{10} を，他の尺度に比べて高順位にする，あるいは，過大評価すると言える．そのため，そのような低頻度単語を含まない正解単語を抽出するには不向きな尺度であることが予想できる．このことは，5.3 節で確認する．

最後に，単独尺度間の関係を調べるために，順位相関係数 (Kendall's τ) を計算したところ図 7 のようになった．図 7 では，LLR, Yates, Chi2 は互いの順位相関が 0.91 以上であり，その平均は 0.93 である．また，Dice と Freq の順位相関は 0.92 である．そのため，これらは，それぞれ順位相関の観点からは良く似た尺度であると言える．また，CSM, PMI, Cosine は，それぞれ，ある程度は独立な尺度であると言える．これより，8 つの尺度は，5 つのグループに分けることができる．これらグループ間には，順位相関の平均[21]が 0.5 より大きいものについては実線が引いてあり，そうでないものについては点線が引いてある．なお，これらの線に付いている数字は，グループ間の尺度間の順位相関の平均である．

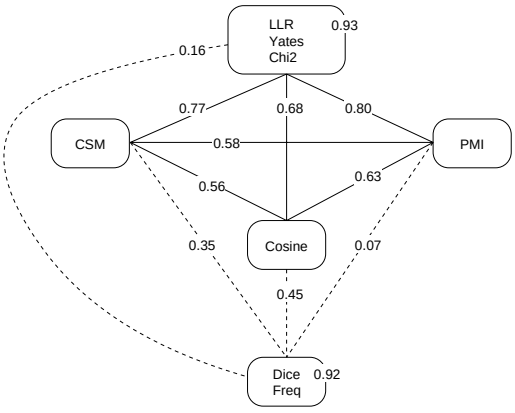

図 7　単独尺度間の順位相関．

21 たとえば，CSM と LLR, CSM と Yates, CSM と Chi2 の順位相関は，それぞれ，0.81, 0.76, 0.75 なので，平均は 0.77 である．

5.3　単独尺度の平均精度

表4には各単独尺度の平均精度を示す．表4から分かるように，Yates, Cosine, Chi2, PMI の平均精度は下位に位置し，その順番は，ちょうど図6における状況と一致している．また，Freq と Yates の間の平均精度の差は0.08程度であり，他の尺度間における精度差と比べて大きい．このことは，5.2節で述べたように，Yates, Cosine, Chi2, PMI は，正解単語を抽出するには不向きな尺度であることを裏付けている．

表 4　単独尺度の平均精度.

尺度	平均精度 (AP)
CSM	0.3685
Dice	0.3212
LLR	0.3104
Freq	0.3042
Yates	0.2284
Cosine	0.2205
Chi2	0.2082
PMI	0.1574

表4においては，Freq と Dice と LLR とは同程度の平均精度である．しかしながら，5.5節で示すように，Freq や Dice と LLR とでは，精度の高い (正解の多い) 順位の範囲が異なる．つまり，Freq や Dice では，LLR と比べて，遅い順位の方で精度が高いのだが，LLR では，早い順位の方で精度が高い．Freq や Dice で遅い順位の方で精度が高い理由は，Freq や Dice においては，早い順位では高頻度単語が多く，それらは特徴単語ではないため精度が高くないが，後の順位の方で中程度の頻度の単語が現われると，正解数が多くなり精度が高くなると考えられる．

CSM については，最高の平均精度である．このことから，CSM が文字認識 (澤木・萩田 1995) や単語の上位下位関係の推定 (山本・梅村 2002) だけでなく，特徴単語の抽出にも役立つことが言える．

なお，平均精度のベースラインとしては，$\frac{\text{リストに含まれる全正解数}}{\text{候補語リストの大きさ}} = \frac{667}{5016} = 0.1330$が適当である．なぜなら，もし，候補単語を無作為に順位付けるとすると，上位から候補単語を調べていったとき，候補語が正解であるような順位の間隔は，平均して1/0.1330であるので，それを (23) 式に適用すると，平均精度は0.1330となるからである．

このベースラインと比べたときには，表4の尺度は，全て，その平均精度がベースラインを上回っている．そのため，どの尺度も，無作為に順位付けた場合と比較すれば，何らかの有効性があると言える．

5.4　複合尺度の平均精度

ここでは，どのような尺度の組み合せが特徴単語抽出に有効かを調べるために，単独尺度の組み合せを網羅的に調べて，その組み合せの精度を比較する[22].

表5には，上位10位の平均精度である複合尺度を示す．表では，同一精度であるような複合尺度は同一順位としている[23].

表 5　上位10位の複合尺度の平均精度.

順位	平均精度	使われた尺度							
1	0.39046	CSM	PMI	Cosine	Dice	Yates			
		CSM	PMI	Cosine	Dice	Yates	LLR		
		CSM	PMI	Cosine	Dice	Yates		Chi2	
		CSM	PMI	Cosine	Dice	Yates	LLR	Chi2	
5	0.39012	CSM	PMI	Cosine	Dice				
		CSM	PMI	Cosine	Dice		LLR		
		CSM	PMI	Cosine	Dice			Chi2	
		CSM	PMI	Cosine	Dice		LLR	Chi2	
9	0.39008	CSM	PMI	Cosine	Dice	Yates	LLR	Chi2	Freq
		CSM	PMI			Yates			Freq

表5と表4とを比べると，単独尺度であるCSM(AP=0.3685)に比べて，1位の複合尺度(AP=0.39046)の平均精度の向上は0.022程度である．これは，CSMとDice(AP=0.3212)における平均精度の差0.047程度に比べると小さいため，統合における平均精度の向上は，適切な単独尺度の選択における精度の向上よりも，小さいことが分かる．しかし，平均精度は向上している．そして，既存の単独尺度を複数組み合わせることは，新たに有効な尺度を探すことに比べると，容易であるので，複数尺度の統合により精度が向上するなら，統合を利用することが好ましい．そのため，複数尺度の統合は有効であると言える．

ここで，表5で同精度で1位である4つの複合尺度は，5つの単独尺度(CSM, PMI, Cosine, Dice, Yates)を共有している．そのため，これら5つの単独尺度の組み合せが有効であると言える．また，9位である，8つの単独尺度全てを使った場合の平均精度は0.39008であり，最適(1位のAP=0.39046)の場合と比べての精度差はわずかである．このことは，4節で述べた複数尺度の統合法が，最適な尺度の組み合せを越えての尺度の追加に対しても，精度が大きく低下することなく頑健(robust)に，これら尺度を統合できることを示している．

なお，本稿で，複合尺度の性質を調べるときには，表5で1位の平均精度である5つの単独尺度(CSM, PMI, Cosine, Dice, Yates)の組み合せからなる複合尺度をComboと呼び，Comboについての性質を調べる．

22 単独尺度の数は8つであるので，それらの可能な組み合せの数は $\sum_{i=1}^{8} \left(\begin{array}{c} 8 \\ i \end{array} \right) = 255$ である.

23 表5における第9位と同精度である複合尺度は16個あるが，そのうち2個のみを示す.

5.5　順位と正解数の関係

5.3節と5.4節とでは，平均精度により各種尺度の有効性を評価した．本節では，各尺度により順位付けたときの順位と，その順位までに抽出された正解数により，各種尺度を比較する．

図8は，各種尺度により順位付けされた候補語リストについて，100位から5000位まで，100位ごとに累積正解数を数えたときのグラフである．ただし，縦軸が累積の正解数であり，横軸が順位である．なお，図8では，Freq, Yates, Chi2が省略されている．その理由は，DiceとFreqが良く似た折線であり，LLRとYatesとChi2が良く似た折線であるため，図を明瞭にするために，それぞれにおいて，表4の平均精度が最大であるDiceとLLRとを表示したためである．なお，これら省略されたものについては図9に示す．

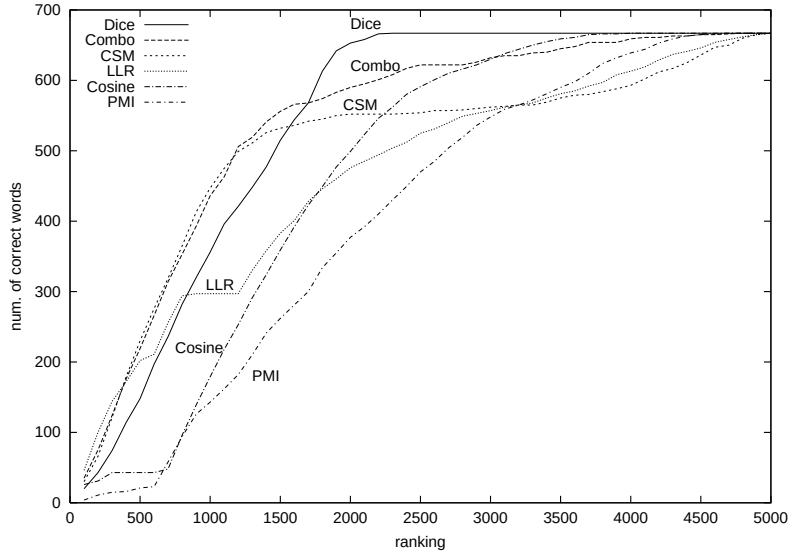

図 8　順位と正解数の関係.

図8より，上位の順位である300位程度までは，LLR が比較的精度良く正解を抽出している．300位の時点での順位精度は，LLRが0.48, Comboが0.41, CSMが0.40である．その後，1100位程度までは，CSMの精度が良い．1100位の時点での順位精度は，CSMが0.43, Comboが0.42である．また，1700位程度までComboの精度が良い．1700 位の時点での順位精度は，ComboとDice共に0.33である．それ以降は，Diceの精度が良くなっている．

図8より，Comboが全区間に渡って，安定して上位の順位精度(累積正解数)であることがわかる．このことは，複数尺度を統合することにより，安定して精度良く正解単語を抽出できることを示している．このことからも，複数尺度を統合することが有効であると言える．

次に，図8に示されていないFreq, Yates, Chi2について，図9に示す．図9 より分かるよう

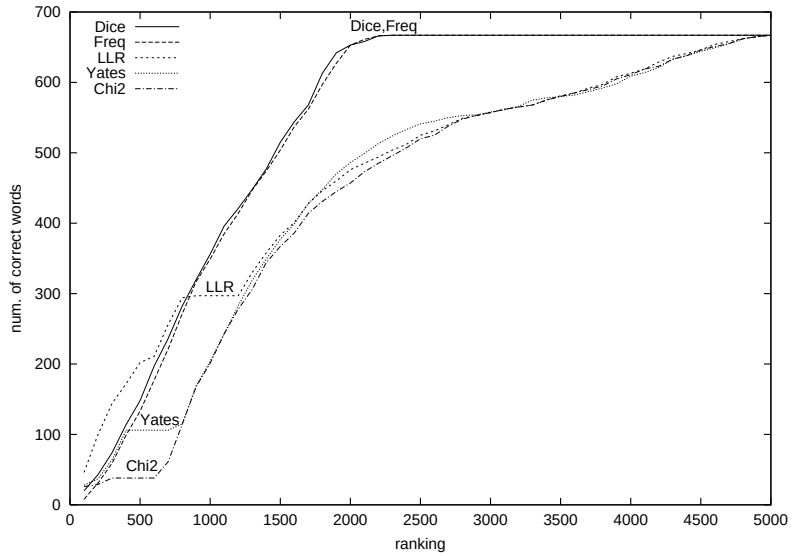

図 9　順位と正解数の関係.

に，Freq と Dice とは互いに良く似た折線である．また，LLR, Yates, Chi2 については，上位の方で LLR の精度が良いが，途中から，良く似た折線となる．上位の方で，LLR > Yates > Chi2 の順で精度が良いのは，5.2節の図6 のところで述べたように，これらの尺度における，依存性の測定の信頼性に対応していると考える．なお，図9の LLR, Yates, Chi2 では，正解数が変化していなく水平になっている部分があるが，これは，ちょうど，図6における色付きの低頻度単語の部分に相当している．また，図9と図7とを比べると，図7において順位相関の高い Freq と Dice，および，LLR, Yates, Chi2 については，その折線の形も似たものとなっていることが分かる．

5.6　上位に順位付けられた単語の例

本節では，各尺度により上位20位に抽出された単語を観察することにより，各尺度の上位における傾向を考察する．

まず，上位20位に抽出された共通単語数により各尺度をグループに分けると，図10のようになる．図では，まず，グループ A は Freq のみからなる．次に，グループ B については，CSM と Dice とで 16 語が共通している．グループ C についても 16 語が共通している．また，グループ D については，グループ内の任意の尺度間で，互いに共通する単語の平均は 18.7 語である．グループ E は PMI のみである．また，グループ間にある線には，それぞれのグループ間の尺度間で共通する単語の平均数が付けられている．なお，線がないグループ間での共通単語数は3以下である．

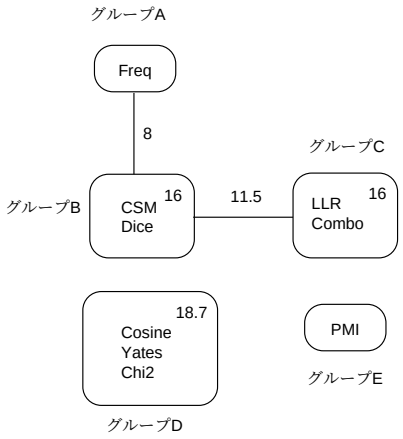

図 10　尺度間で重複する単語数.

表6, 7, 8には，各尺度について上位20位に順位付けられた単語を示す．各表において，1行目には，それぞれの尺度のグループを示し，2行目には，各尺度と，その尺度により得られた単語の頻度 (Freq) の，上位20位における中央値を示す．また，括弧内には，その中央値に対応するFreqによるソートでの順位を示す[24]．たとえば，表6では，Freqにおける頻度の中央値は1272で，その頻度におけるFreqの順位は10位であり，CSMでの中央値は849で，そのFreqにおける順位は18位であり，Diceでの中央値は332で，そのFreqでの順位は34位である．最後に，3行目以降の行には，各尺度により抽出された単語とその頻度が，尺度値の降順で並んでいる．なお，尺度値の表示は省略した．また，これらの表においては，これら同一グループ内の尺度について，その尺度でしか抽出されていない単語には下線が引いてある．なお，グループDについては，YatesとChi2のみに共通な単語は[]で囲んである．

これら中央値に端的に示されるように，各尺度が抽出する単語の頻度には特徴がある．まず，グループAであるFreqは，当然，高頻度語が上位になる．次に，グループB (CSMとDice) も比較的高頻度語が上位になる．グループC (LLRとCombo) は，これらに比べると頻度が低い単語からなる．また，グループE (PMI) は，図6に示したように，低頻度語のみである．最後に，グループD (Cosine, Yates, Chi2) も比較的低頻度語からなることがわかる．また，中央値に対応するFreqでの順位についても，中央値における傾向に対応して，グループA,B,C,D,Eの順に，Freqにおける上位から下位の単語を抽出していることがわかる．

図10のグループに基づいて，それぞれのグループで上位20位以内に順位付けられた単語リストの性質を述べると，以下のようになる．

- グループA (Freq) については，特定分野や一般分野に関わらず，通常の頻度調査により

24 同一頻度の単語が複数あるときには，最高順位の単語についての順位を示した．

表 6　上位20位の単語(グループ A,B).

	グループ A		グループ B			
	Freq	1272 (10)	CSM	849 (18)	Dice	332 (34)
1	THE	7786	A	4115	COMPANY	402
2	BE	5363	WILL	1426	WHAT	849
3	A	4115	BE	5363	WILL	1426
4	TO	3457	YOU	1604	OFFICE	312
5	OF	2263	WHAT	849	QUESTION	332
6	IN	1862	COMPANY	402	REFER	229
7	YOU	1604	AT	901	FOLLOW	318
8	WILL	1426	QUESTION	332	NEW	382
9	HAVE	1350	OFFICE	312	MAN	349
10	FOR	1272	TO	3457	YOU	1604
11	AND	1197	DO	961	EMPLOYEE	193
12	I	1044	FOLLOW	318	WOMAN	278
13	DO	961	WE	893	SERVICE	265
14	ON	948	NEW	382	SALE	204
15	IT	939	MAN	349	A	4115
16	AT	901	FOR	1272	AT	901
17	WE	893	REFER	229	PLEASE	195
18	WHAT	849	WOMAN	278	BUSINESS	212
19	THIS	786	SERVICE	265	DO	961
20	THEY	768	EMPLOYEE	193	HOW	301

表 7　上位20位の単語(グループ C,E).

	グループ C				グループ E	
	LLR	229 (56)	Combo	204 (64)	PMI	2 (2669)
1	OFFICE	312	OFFICE	312	RECONFIGURATION	1
2	REFER	229	EMPLOYEE	193	RENOVATE	5
3	EMPLOYEE	193	REFER	229	ECONOMIZE	1
4	WILL	1426	COMPANY	402	INFORMATION-SYSTEM	2
5	COMPANY	402	QUESTION	332	COATROOM	1
6	QUESTION	332	SALE	204	REORGANIZE	2
7	WHAT	849	HOTEL	155	LUMBERYARD	1
8	SALE	204	PLEASE	195	ATTENDEE	4
9	PLEASE	195	CUSTOMER	149	SEMIANNUALLY	1
10	HOTEL	155	STORE	129	MONOLINGUAL	1
11	CUSTOMER	149	COMPUTER	153	ENCRYPTION	3
12	FOLLOW	318	WHAT	849	PREPAID	2
13	VACATION	57	FOLLOW	318	SHOPLIFTER	1
14	STORE	129	WILL	1426	INTERMISSION	4
15	COMPUTER	153	TRAVEL	125	OBLIGATE	1
16	A	4115	BUSINESS	212	STOCKHOLDER	6
17	SERVICE	265	SERVICE	265	LITHIUM-ION	1
18	BUSINESS	212	HOUR	173	SCREENING	2
19	MAIL	74	ORDER	184	DETAIL-ORIENTED	1
20	WOMAN	278	ROOM	190	DEDUCTIBLE	3

単語を順位付けた場合と同様な単語が上位20 位以内に位置している．すなわち，Freqにより上位20位以内に抽出される単語は，特徴単語とは言えない[25]．

25 また，図9を細かくみると，Freqについては，上位100位までにおける正解数が8 ($\frac{8}{100}$=0.08)であり，200 位までにおける正解数が31 ($\frac{31}{200}$ = 0.155)，300位までにおける正解数が 59 ($\frac{59}{300}$=0.197)であり，かつ，5.3節で述べたように，候補単語を無作為に順位付けるとすると，0.133の割合が正解であるので，上位100位以内に抽出される単語については，正解の割合が無作為に抽出された場合よりも低い．したがって，Freqにより，上位100位程度以内に抽出され

表 8　上位20位の単語(グループD).

	グループD					
	Cosine	10 (1156)	Yates	9 (1225)	Chi2	9 (1225)
1	CHECK-OUT	17	CHECK-OUT	17	CHECK-OUT	17
2	DOWNTOWN	17	DOWNTOWN	17	DOWNTOWN	17
3	E-MAIL	16	E-MAIL	16	E-MAIL	16
4	UPCOMING	15	UPCOMING	15	UPCOMING	15
5	HAMBURGER	13	HAMBURGER	13	HAMBURGER	13
6	COPIER	11	COPIER	11	COPIER	11
7	FERRYBOAT	10	FERRYBOAT	10	FERRYBOAT	10
8	THE	7786	TEAL	9	TEAL	9
9	TEAL	9	BEVERAGE	9	BEVERAGE	9
10	BEVERAGE	9	ACCORDANCE	8	ACCORDANCE	8
11	ACCORDANCE	8	REIMBURSE	8	REIMBURSE	8
12	REIMBURSE	8	INTEROFFICE	8	INTEROFFICE	8
13	INTEROFFICE	8	VACATION	57	PAYLOAD	7
14	BE	5363	PAYLOAD	7	SIGHTSEEING	7
15	VACATION	57	SIGHTSEEING	7	NEWSSTAND	7
16	PAYLOAD	7	NEWSSTAND	7	FORFEIT	7
17	SIGHTSEEING	7	FORFEIT	7	SALESPEOPLE	7
18	NEWSSTAND	7	SALESPEOPLE	7	[ALUMNI]	7
19	FORFEIT	7	[ALUMNI]	7	[REQUISITION]	7
20	SALESPEOPLE	7	[REQUISITION]	7	VACATION	57

- グループB (CSM, Dice) については，Freq以外の他の尺度に比べると，高頻度のものが上位に抽出されていて，特徴単語を学習する前段階として基本語彙の復習が必要な補習レベル学習者に適していると考えられる．
- グループC (LLR, Combo) については，5.5 節において述べたように，上位で効果的に特徴単語を選択している．このグループで抽出された単語には，TOEICの特徴であるビジネス・コンテキストで広く使われる単語，たとえば，会社・人事・オフィス・出張の場面で用いられる business, company, employee, office, computer, travel, hotel, room等の基本語，そして日常生活における sale, store, customer, service, order等買物・購入の場面に分類される基本語がバランス良く抽出されており，グループBで選択される単語よりも一段階上のレベルで学習すると好ましい，初級向けの良質なTOEIC特徴単語を抽出していると判断できる．
- グループD (Cosine, Yates, Chi2) については，グループCよりも少し高レベルのTOEIC特徴単語と言える interoffice, copier, upcoming, reimburse, forfeit や beverage, check-out, downtown, newsstand, hamburger 等，英語圏の文化背景と共に教えると効果的な日常語が現れる．
- グループE (PMI) については，図6にも示したように，低頻度単語を過大評価する[26]．しかし，このことは学習者のレベルによっては欠点とは言えない．なぜなら，TOEIC高得

る単語には，特徴単語と言えないものが(無作為に抽出された場合よりも)多いと言える．

26 グループEにリストされている単語は，図6の説明のところでも述べたように，1位から630位までが同スコアであるので，それらを無作為に順位付けた上位20位をとっている．

点を目指すような上級レベルの学習者にとっては，このような低頻度単語も学習する必要があるからである．PMIは，そのような目的にとっては有効であると言える．[27]

以上をまとめると，各グループにより抽出される単語は，それぞれ異なったレベルにおける学習者に有効であると考えられる．そのため，学習者のレベルに応じて，適切な尺度を選ぶ必要がある．

次に，各尺度により上位に抽出される単語が，単純な頻度であるFreq を利用した場合には，上位には抽出され難いことを確認する．その理由は，もし，各尺度により上位に抽出される単語が，Freqにおいても比較的上位，たとえば，100位以内に抽出されるなら，Freqのみを用いても，各尺度で上位に抽出される単語を抽出できるため，特に複雑な尺度を利用する必要がないからである．そのため，Freq以外の尺度の有用性を確認するために，以下の考察をする．

まず，Freqで上位100位以内に抽出された単語と各尺度の上位100位以内に抽出された単語との共通単語数を求めると表9のようになる．

表 9　上位100位以内におけるFreqとの共通単語数

尺度	Dice	CSM	Combo	LLR	Cosine	Yates	Chi2	PMI
共通単語数	74	52	45	33	9	3	2	0

この表から分かるように，Freqとの共通単語数については，尺度ごとに，その共通する程度に差があるが，どの尺度についても，Freqとは異なる単語を上位100以内に抽出していることがわかる．一例として，Freqと一番共通単語数が多いDiceについて，Diceにより上位100位以内に抽出された単語のうちで，Freqでは上位100位以内に抽出されない単語を抜き出すと以下の26語になる．

store, manager, car, travel, meet, meeting, pay, product, market, receive, increase, offer, today, account, letter, staff, building, price, date, morning, conference, check, notice, speaker, card, request

これらは，表6で，Dice (グループB) により上位20位以内に抽出される単語の性質を述べたときと同様の傾向を示しており,「特徴単語を学習する前段階として基本語彙の復習が必要な補習レベル学習者に適している」と考えられる．更に，その他の尺度についても，各尺度ごとに，Freqでは上位100位以内に抽出されない単語の傾向について確認したところ，表6, 7, 8での20位以内における単語リストでの考察と同様に，異なるレベルにおける学習者に有効であると考えられる単語が上位100位以内に抽出されていた．そのため，Freqだけを使うのではなく，各

27 脚注19では，低頻度の単語については，特徴単語として語彙に加える必要性は低いとしたが，その理由は，2節で述べたように，本稿で主な対象と想定している学習者が英語初級者・中級者であるので，そのような学習者にとっては，特定分野で重要な単語を効率的に習得するという観点からは，教育用語彙に加える必要性が低いからである．一方，上級者にとっては，このような低頻度単語であっても，初級・中級の語彙を越えて，一層の語彙増強を目的とする場合には，学習する必要性の高い単語となる．実際，図6のところで述べた「$a = 1, b = 0$」である404語については，複合語や派生語などではない (意味がその単語だけでは類推できないような) 単語が，約25%を占めていた．これらの単語は，上級者向けの語彙増強のためには，学習する必要性が高い単語である．

尺度を適切に選んで利用することにより，異なったレベルにおける学習者に応じた特徴単語を上位に順位付けて効率的に選定できると考える．

ただし，Freqについても，たとえば，上記実験において正解データとして用いた特徴単語リストが，2節で述べたように，頻度を参考にして選定されていることからも分かるように，Freqが特徴単語の選定に役立つことは既に知られている．それに対して，上記考察で明らかになったことは，Freq以外の尺度を使うことにより，Freq では上位に抽出できないような特徴単語を上位に抽出でき，かつ，それらが，異なったレベルにおける学習者に対応すると考えられることである．このことから，Freqに加えて，本稿で比較検討した尺度を適切に選択し利用することにより，異なったレベルにおける学習者に応じた特徴単語を効率的に選定できると考える．

6　おわりに

特定分野の英語を効率的に学習するためには，その分野に特徴的な語彙を選定し，その語彙を学習するのが効果的である．そのような語彙を選定するためには，まず，その特定分野における単語を，特徴的な順番に並べて，その上位から選定していくのが良い．しかしながら，現状においては,「特徴的な単語」を抽出するためには，どのような尺度を利用して単語を順位付けるべきかが分かっていなかった．

そのため，本稿では，各種尺度を比較し，どの尺度が特徴的な単語の抽出に有効かを調べた．さらに，単独の尺度を利用するだけでなく，それらを統合した複合尺度の有効性も調べた．このときに対象とした分野は，日常生活やビジネスの場面のテキストを多く含むTOEICである．

その結果，主に以下のことが，明らかになった．

- 各種尺度により順位付けられた単語リストと，英語教育用に人手により選定された単語リストとの一致という観点からは，複合尺度(Combo)の有効性が示された．また，単独尺度においては，補完類似度(CSM)が有効であった．
- 各種尺度により抽出される単語は，それぞれ，異なったレベルにおける学習者に有効であると考えられること，そのため，学習者のレベルに応じて，適切な尺度を利用する必要があること，が分かった．

人手による特徴単語の選定は，コストが高いだけでなく，選定者の主観や経験に大きく依存するものである．それに対して，本稿で比較検討した尺度や，別のもっと有効な尺度を探究し利用することにより，それらの選定を，比較的，低コストで，かつ，客観的にできるようになると期待している．

謝辞

情報通信研究機構村田真樹主任研究員との草稿段階での議論が参考になった．

参考文献

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, chap. 3. Addison-Wesley.

中條清美 竹蓋幸生 (1989). “女性向け英語雑誌の語彙.” 時事英語学研究，第28号，pp.73–84.

中條清美 (1991). “英語教育基本語彙の選定に関する研究.” 千葉大学自然科学研究科学位論文.

中條清美 (2003). “英語初級者向け「TOEIC語彙1,2」の選定と効果.” 日本大学生産工学部研究報告 Vol.36, pp.27–42.

中條清美 内山将夫 (2004). “統計的指標を利用した特徴語抽出に関する研究.” 関東甲信越英語教育学会研究紀要第18号 (掲載予定).

Chujo, K. and Nishigaki, C. (2003). “Bridging the Vocabulary Gap: from EGP to EAP.” *JACET Bulletin*, **36**, 73–84.

Chujo, K. (2004). “Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List.” JAECS(英語コーパス学会)10周年記念論文集. 掲載予定.

Chung, T. M. and Nation, P. (2003). “Technical vocabulary in specialised texts.” *Reading in a Foreign Language*, **15** (2), http://nflrc.hawaii.edu/rfl/October2003/chung/chung.html.

Church, K. W. and Hanks, P. (1989). “Word Association Norms, Mutual Information, and Lexicography.” In *Proc. of ACL-89*, pp. 76–83.

Douglas, D. (2003). “English for Specific Purposes vs. English for General Purposes: What can Testing offer Teachers?.” In *Proc. of the KATE (Korea Association of Teachers of English) 2003 Winter International Conference (English for Specific Purposes vs. English for General Purposes in the EFL Context)*, pp. 5–8.

Dunning, T. E. (1993). “Accurate methods for the statistics of surprise and coincidence.” *Computational Linguistics*, **19** (1), 61–74.

深山晶子, 野口ジュディー, 寺内一, 笹島茂, 神前陽子 (2000). ESPの理論と実践. 三修社.

Hisamitsu, T. and Niwa, Y. (2001). “Topic-Word Selection Based on Combinatorial Probability.” In *NLPRS-2001*, pp. 289–296.

池田央（編） (1989). 統計ガイドブック. 新曜社.

石川由紀, 田中貴美枝, 高橋秀夫, 竹蓋幸生 (1987). “ビジネス英語の語彙.” 語学教育研究所紀要，第1号，pp.53–66.

影浦峡 (1997). “文字単位のbigram尺度に基づく複合漢字列の単位切り手法.” 言語処理学会第3回年次大会発表論文集, pp.477–480.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.

Noguchi, J. (2002). "ESP: Where are we and where do we want to go?." In *JACET Summer Seminar Proceedings (New Perspectives in ESP)*, No. 2, pp. 18–24.

Orr, T. (2002). "Twelve ESP Program Models for Study and Reflection." In *JACET Summer Seminar Proceedings (New Perspectives in ESP)*, No. 2, pp. 25–30.

澤木美奈子 萩田紀博 (1995). "補完類似度による劣化印刷文字認識." 信学技報PRU95-14, pp.101–108.

Schone, P. and Jurafsky, D. (2001). "Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?." In *EMNLP-2001*, pp. 100–108.

Sutarsyah, C., Kennedy, G., and Nation, P. (1994). "How useful is EAP vocabulary for ESP? A Corpus-Based Study." *RELC Journal*, **25**, 34–50.

竹蓋幸生 (1981). コンピューターの見た現代英語. エデュカ出版.

竹蓋幸生, 高橋秀夫, 星野昭彦 (1987). "計算機科学の語彙—コンピュータを英語で学ぶために." 千葉大学教育工学研究，第8号，pp.27–40.

内山将夫 井佐原均 (2003). "複数尺度の統計的統合法とその専門用語抽出への応用." 情報処理学会自然言語処理研究会研究報告, 2003-NL-157, pp.69–76.

山本英子 梅村恭司 (2002). "コーパス中の一対多関係を推定する問題における類似尺度." 自然言語処理, **9** (2), 45–75.

略歴

内山 将夫: 1992年筑波大学第三学群情報学類卒業．1997年筑波大学大学院工学研究科博士課程修了. 博士(工学)．1997年信州大学工学部電気電子工学科助手．1999年郵政省通信総合研究所非常勤職員．2001年独立行政法人通信総合研究所任期付き研究員．2004年情報通信研究機構（旧：通信総合研究所）研究員．言語処理学会，情報処理学会，ACL，人工知能学会，大学英語教育学会，英語コーパス学会，各会員.

中條 清美: 1991年千葉大学大学院自然科学研究科修了．学術博士．1994年千葉大学外国語センター専任講師．1999年Manhattanville College (N.Y.). M.P.S. (Master of Professional Studies) in TESL. 現在，日本大学生産工学部助教授．英語教育．言語処理学会，大学英語教育学会，英語コーパス学会，TESOL，各会員.

山本 英子: 1996年豊橋技術科学大学情報工学課程卒業．2002年同大学大学院工学研究科電子・情報工学専攻博士課程修了．博士(工学)．同年，情報通信研究機構（旧：通信総合研究所）専攻研究員．言語処理学会，情報処理学会，

人工知能学会，各会員．

井佐原 均: 1978年京都大学工学部卒業．1980年同大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995年郵政省通信総合研究所関西支所知的機能研究室室長．2001年情報通信研究機構（旧：通信総合研究所）けいはんな情報通信融合研究センター自然言語グループリーダー．自然言語処理，機械翻訳の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．