教育用語彙選定における特徴語抽出及び Wordplot の利用

中條清美 [1]　内山将夫 [2]　中村隆宏 [3]　西垣知佳子 [4]

1 日本大学　2 情報通信研究機構　3 小学館　4 千葉大学

1　背景

英語の教授・学習に使える時間は限られている。英語の語彙学習においてはすべての語を習得することは不可能で，また習得する必要もない。そこで，限られた時間を有効に活用するためには無駄を省いた語彙精選が必要となる。

近年，語彙指導の分野では，効率的な学習・指導には「基本語彙」と「専門語彙」という 2 種類の語彙表で対応すると効果的であるという考え方がある。例えば，最初に 2,000-3,000 語程度の基本語彙を学習した後に，学習者の学問領域や職域に合わせて目標分野を限定し，その特定分野の専門語彙を指導することで学習効率を高めるというものである[1]。

そこで，語彙教材の開発に，基本語彙と専門語彙という 2 種類の教育用語彙の選定作業が不可欠となる。基本語彙は「どの分野にも広く出現する語彙」であり，実際に書かれたり話されたテキストを大量に集めた British National Corpus (BNC)[2] のような一般分野コーパスから高頻度語を選定するという手法が有効である。

一方，ビジネスや科学技術といった「特定分野に特徴的な語彙」である専門語彙の選定に，基本語彙と同様の手法を用いると，中・低頻度の「専門性の高い語」が選定されにくいという問題がある。専門語彙の場合，当該分野の頻度順リストを見るだけでは不十分であり，一般分野の語彙リストのような「特徴のない語彙」と対照させて，両者の出現頻度数に大きな差のある「特徴のある語彙」を抽出する方法が有効である[3]。

さらに，学習者の英語習熟度は一様ではない。そのため教育用の専門語彙選定には，初級・中級・上級といった学習者の習熟度レベルに合致した段階的な専門語彙表が求められる。

以上のような状況に鑑みて，我々は，専門語彙をなるべく簡便で客観的に，しかも学習者の語彙レベルに応じて選定するツールの開発を行ってきた。そのために，1) 9 種の統計指標を用いて特徴語[4]を抽出する方法[1]，さらに語彙の相対的な専門度を視覚的に把握しながら特徴語を抽出する方法として，2) 単語散布図 (Wordplot) を提案した[2]。

本稿では，これら 2 種類の手法を BNC の Commerce 分野と Corpus of Professional English (CPE)[5] に適用して，ビジネス分野および科学技術分野の特徴語を抽出し，その抽出結果の有効性を検討した。

2　研究方法

2.1 特徴語抽出に使用した言語資料

特徴語の抽出には，以下の専門分野および一般

[1] 例えば，Nation (2000:187).
[2] http://www.natcorp.ox.ac.uk/World/HTML/thebib.html
[3] 例えば，竹蓋幸生 (1981)『コンピューターの見た現代英語』東京: エデュカ出版
[4] 本稿では，以下では，頻度分布における何らかの特徴に基づいて抽出された語彙を「特徴語」と定義する。
[5] http://www.perc21.org/

分野コーパスにおける単語の出現頻度付き語彙リストを用いた[6]。

1）専門分野の言語資料

① BNC の Commerce 分野 726 万語より作成した頻度 100 以上の 2,973 語（延べ 5,883,249 語）

② CPE 科学技術コーパス 2,000 万語より作成した頻度 100 以上 4,467 語（延べ 13,527,303 語）

2）一般分野の言語資料

BNC 全体より作成した頻度 100 以上の 13,956 語（延べ 86,008,037 語）[3]

2.2 特徴語の検討に使用した言語資料

1) 英語母語話者の語彙習得学年リスト[7]

2) 基本語彙リストと専門語彙リスト[8]

2.3 9 種の統計指標と特徴語リスト

使用した統計指標は，頻度（Freq），ダイス係数（Dice），コサイン（Cosine），補完類似度（CSM），対数尤度比（LLR），カイ二乗値（Chi2），イエーツの補正公式（Yates），自己相互情報量（PMI），マクニマーのテスト（McNemar）の 9 種である[9]。

以上の統計指標を用いて，Commerce 分野および CPE リストの各語の出現状況を，BNC での当該語の出現状況と比較してその語の特徴度を示す指標値を求めた。指標値にしたがって降順にソートし，上位 500 語から，ビジネス分野と科学技術分野の特徴語リストを 9 種ずつ作成した。

2.4 Wordplot と特徴語リスト

専門分野の英語語彙を一般分野の英語語彙に対照させて可視化する方法として，単語散布図を利用する Wordplot を提案した。この単語散布図では，横軸が一般分野における単語の出現頻度（Fglobal）と割合（Pglobal）を表し，縦軸が専門分野における単語の出現頻度（Flocal）と割合（Plocal）を表す。

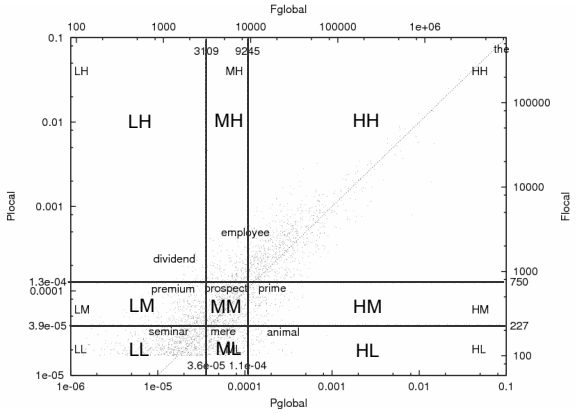

図 1 Commerce 分野の Wordplot（単語散布図）

図 1 は，横軸に BNC における出現頻度を，縦軸に Commerce 分野における出現頻度をとり，単語数によって各軸を 3 等分した 9 領域よりなる Wordplot（単語散布図）である。左上の領域には，一般分野での出現頻度が低く，専門分野での出現頻度が高い単語が位置する。Wordplot の 9 領域に属する語から，Commerce 分野と CPE について 9 種ずつ特徴語リストを得た。

---

[6] 数詞，固有名詞，略語等は人手で除外した。これらの語は，多目的の言語使用に対応する教育用語彙表には不要と考えられる。我々は，外国語としての英語教育における学習語彙は，現代英語を代表する 1 億語の BNC における頻度 100 以上の 13,956 語の範囲内の語彙で十分であると考えている。そこで，専門分野の語彙リストから BNC13,956 語に含まれていない語を除外した。

[7] 学年の決定には 1～3 年は *Basic Elementary Reading Vocabularies* (Harris & Jacobson, 1972), 4～16 年生は *The Living Word Vocabulary* (Dale & O'Rourke, 1981)を参照した。両資料に未収録の語は 17 年生とした。

[8] 詳細は 3.2 参照。

[9] これらの指標はパラメタ a,b,c,d（a：専門リストに単語 X が出現した回数，b：BNC リストに単語 X が出現した回数，c：専門リストの延べ語数－a，d：BNC リストの延べ語数－b）によって計算される。各指標の詳細と定義式は文献[1][4]を参照されたい。

3 結果

ビジネス分野と科学技術分野の特徴語を2種類の手法によって抽出し,抽出結果の有効性を,学年レベル，基本語彙と専門語彙の含有率の観点から検討した。両分野の特徴語抽出結果にはほぼ同様の傾向が見られたので，以下では紙幅の関係で，Commerce 分野について報告する。

3.1 学年レベル

英語母語話者の語彙習得学年リストを用いて，Commerce 分野の特徴語を容易に理解できるのは，米国の何年生ぐらいなのかを調査した結果を表1に示した。表1の左半分は，9 種の統計指標による特徴語上位 500 語の平均学年を指標ごとに示す。表1の右半分は，Wordplot の 9 領域に配分された特徴語の平均学年を領域ごとに示した。

表1 Commerce 分野の特徴語の平均学年

統計指標	対象語数	平均学年
Freq	500	3.3
Dice	500	3.3
Cosine	500	4.3
CSM	500	5.0
LLR	500	6.1
Chi2	500	6.5
Yates	500	6.5
PMI	500	9.0
McNemar	500	10.2

Wordplot 領域	対象語数	平均学年
HH	771	3.5
HM	164	3.1
HL	56	2.4
MH	191	6.2
MM	506	5.8
ML	293	4.8
LH	29	8.2
LM	320	8.9
LL	641	9.0

表1左では，統計指標によって異なる学年レベルの語が抽出されていることが分かる。平均すると，Freq，Dice，Cosine，CSM の特徴語は母語話者の 3～5 年生レベル，LLR，Chi2，Yates は 6 年生レベル，PMI，McNemar は 9～10 年生レベルに相当する語を抽出した。

表1右に示したWordplotも領域ごとに異なる学年レベルの語を抽出している。Wordplot を縦に 3 区分した右列にあたる HH，HM，HL（図 1 の区分参照）は母語話者の 2～3 年生レベル，中列にあたる MH，MM，ML は 4～6 年生レベル，左列にあたる LH，LM，LL は 8～9 年生レベルであった。

以上の結果から，9 種の統計指標と Wordplot の両手法ともレベル別の語彙を適切に選定し分けており，統計指標では Freq から CSM は初級用，LLR から Yates は中級用，PMI と McNemar は上級用に，Wordplot では HH，HM，HL は初級用，MH，MM，ML は中級用，LH，LM，LL は上級用の語彙選定に適することが判明した。

3.2 基本語彙と専門語彙の含有率

各特徴語リストにどの程度，基本語彙と専門語彙が含まれるかの割合を調査した。本稿では基本語彙リストとして次の 4 種を用いた。① Function words（文法的な役割を持つ機能語）[10]，② Basic vocabulary（General Service List[11]のように学習者が最初に学ぶべき基礎的な語彙），③ Textbook vocabulary（日本人学習者が中学・高校教科書を通して習得する語彙）[12]，④ Academic vocabulary（大学生が一般教養科目を理解するために必要な基本語彙）[13]。そして，専門語彙の例として Business dictionary の見出し語リストを Business vocabulary として用いた[14]。

図 2 には，9 種の統計指標によって抽出された特徴語上位 500 語に 4 種類の基本語彙がどの程度含まれるか，また専門語彙である Business dictionary の見出し語との一致度を調べた結果を示した[15]。図 3 には Wordplot の 9 領域に分布した特徴語についての結果を領域ごとに示した。

---

[10] Nation[5]の Function words (pp.430-431)を使用した。

[11] West, M. (1953). *A General Service List of English Words*. Longman.

[12] *New Horizon English Course 1，2，3* と *Unicorn English Course Ⅰ，Ⅱ, Reading* より作成した 3,245 語（延べ語数 38,937 語）

[13] http://www.vuw.ac.nz/lals/research/awl/awlinfo.html

[14] *Longman Business English Dictionary* (Pearson Education Limited, 2000) の見出し語 4,565 語

[15] Freq と Dice，Chi2 と Yates は同じ値であったので，図 2 では同軸上に示した。

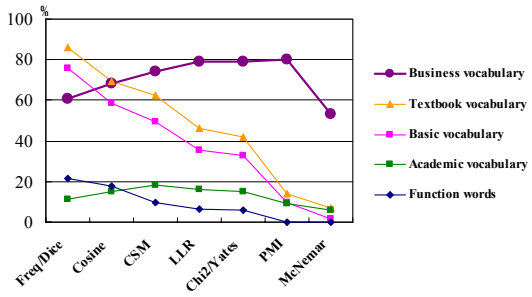

図 2　統計指標による特徴語の基本･専門語彙含有率

図 2 より，LLR，Chi2，Yates，PMI の 4 指標の上位 500 語の約 8 割が Business dictionary の見出し語と一致する。一方，Freq から McNemar にかけて基本語彙の含有率は段階的に減少している傾向が見られることから，PMI に専門性の高い語彙が抽出されていることが推定される。実際，PMI の特徴語上位には，lading，buyout，arbitrage，subcontractor，liquidity，volatility など専門的な経済用語が順位付けられている。

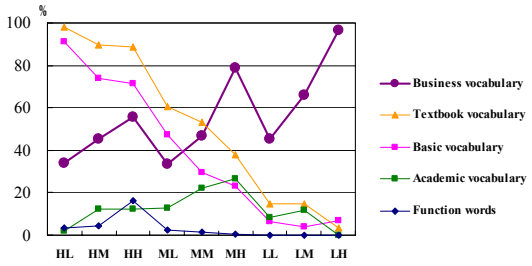

図 3　Wordplot による特徴語の基本･専門語彙含有率

図 3 に見られるように，LH の領域語のほぼすべてが Business dictionary の見出し語と一致する。LH には dividend，equity，discount，monetary や takeover，audit，merger のような経済用語が含まれる。次いで MH，LM の順に一致度が高い。MH に含まれる語は，asset，credit，stock，cash，loan，debt や employee，consumer，investor，buyer，shareholder のようなビジネス用語である。LM では専門語彙の含有率は少し下がるが，premium，aggregate，lease，stake，innovation，monopoly，incentive などの経済英語が含まれる。これらの領域では基本語彙の含有率は低く，高い精度で専門語彙を取得できることがわかる。

3.1 の結果と併せて考察すると，統計指標のうち中級用には LLR，Chi2，Yates が，上級用には PMI が高い精度で専門語彙を選定し，Wordplot では中級用には MH，上級用に LH，LM の領域に入る語を選定すれば高い精度で専門語彙を選定できることが判明した。

4　おわりに

本稿では，ビジネス分野と科学技術分野の専門語彙を対象として統計指標と Wordplot を使用して特徴語を抽出し，その専門語彙抽出の有効性を検証した。本稿で検討した 2 種類の選定手法を適切に選択し利用することで，学習者に応じて，初級･中級・上級といった段階的な専門語彙を効率的に選定できると考える。

参考文献

[1] Chujo, K. and Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, 34 (2), 255-269.

[2] Utiyama, M. and Chujo, K. (2007). Linking word distribution to technical vocabulary. *Journal of the College of Industrial Technology, Nihon University*, 40（in press）.

[3] Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. In: J. Nakamura, N. Inoue, and T. Tabata (Eds.), *English Corpora under Japanese Eyes* (pp. 231–249). Amsterdam: Rodopi.

[4] 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.

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