単語連鎖にみる科学技術分野と他分野の英語表現比較

統計数理研究所共同研究リポート239
『ESPコーパスからの特徴表現の抽出』2010 pp. 1-12

単語連鎖にみる科学技術分野と他分野の英語表現比較

小山　由紀江*
水本　篤**

*名古屋工業大学
〒466-8555 名古屋市昭和区御器所町
E-mail: koyama@nitech.ac.jp

**流通科学大学
〒651-2188 神戸市西区学園西町
E-mail: atsushi@mizumot.com

概要　本研究では，6 種類のコーパスで共通して使用される「単語連鎖」（lexical bundles）を比較することにより，分野の違いが明らかになるのかという点を調査した。また，どのような単語連鎖により，その違いが現れるのかを検討した．コーパスは，(1) 高校教科書，(2) Scientific American（一般科学雑誌），(3) Nature（専門科学雑誌），(4) 機械系論文，(5) 映画スクリプト，(6) 創作散文（imaginative prose）の 6 種類を使用し，上位 100 位までの 4-gram を対象にコレスポンデンス分析を行った。その結果，広い範囲のディスコースの違いを単語連鎖によって明らかにできることが確認された．また，ある特定専門分野の 4-gram で抽出した特徴表現を効果的な ESP 教育に用いる手法の可能性が示された．

キーワード　MWE，コーパス，4-gram，単語連鎖，コレスポンデンス分析

Comparison of English Expressions from Science, Technology, and Other Fields by Examining Lexical Bundles

KOYAMA, Yukie*

MIZUMOTO, Atsushi**

*Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan

**University of Marketing and Distribution Sciences
Gakuen Nishi-machi, Nishi-ku, Kobe, 651-2188 Japan

Abstract　This paper examines whether or not it is possible to identify differences

between the use of English expressions in 6 different domains by comparing the lexical bundles extracted from 6 different corpora. The corpora used in this study are (1) the high school textbooks, (2) Scientific American (a popular science magazine), (3) Nature (a science journal), (4) journal papers of mechanical engineering, (5) movie scripts, and (6) imaginative prose. A correspondence analysis was conducted with the highest 100 4-grams from these corpora. The results show the different discourses in varied domains can be explained by the extracted 4-grams. This also suggests that the extraction of 4-grams from a certain domain could lead to an efficient way of teaching domain specific expressions.

Keyword MWE, Corpus, 4-gram, Lexical bundle, Correspondence Analysis

１．始めに

英語教育において English for Specific Purposes (ESP)は習得したスキルが学生の将来設計につながり，学習のモーティベーションを高めるという点からも，社会的ニーズという点からも，現在最も効果的な方法の一つとして広く認められているが，他の分野に比べ科学技術分野のアカデミックな文書作成については英語の重要性がさらに高いものとなっている．その理由は，論文を含めた研究分野での出版言語は一般的に英語の使用割合が高く，その中でも科学分野での英語使用が極めて高いことであり，この傾向は年々進んでいることが挙げられる．J. Swales は *Genre Analysis: English in Academic and Research Settings* (1990)[1] において Baldauf and Jernudd (1983) の調査結果を引用しているが，これによると 1981 年の段階で Biology や Physics の分野では既に 85%以上の学術論文が英語によって要約され，【表 1】に示すように自然科学の多くの分野で英語使用の急速な増加傾向が指摘されている．

【表 1】Percentages of Abstracted English Language Research Articles.

Discipline	1965	1981	gain
Chemistry	50	67	17%
Biology	75	86	11%
Physics	73	85	12%
Medicine	51	73	22%
Math	55	69	14%

Note. From Swales (1990, p. 97)

ESP 教育では教えるべきコンテンツの選定はその成否を決める重大な課題であり，コーパス言語学の発達とともに，高専の学生を対象に作成された理工系の語彙リスト「COCET3300」（2004 年発表），科学技術論文を分析した「名工大 EGST 語彙リスト」(石川・小山，2007)[2] 等がある．しかし近年個々の単語にとどまらず「語」のバウンダリーを越えた「語句」の使用の実態に関心が広がり，言語学者の間で lexical bundle, cluster, chunk 等と呼ばれる Multi-Word Expression（MWE）がその対象として注目を集めている．MWE とは通常平均的に観察されるよりもより多く出現する“cohesive lexemes that cross word boundaries” (Calzolari et al., 2002) [3]つまり「語の境界を越えた一連の語句」であるが，この修得が専門分野に適した英語表現に繋がる重要なポイントと言えよう．この点を Hyland (2008a, p. 5) [4]は次のように述べている．

> These bundles are familiar to writers and readers who regularly participate in a particular discourse, their very ‘naturalness’ signaling competent participation in a given community.

ところがコーパスからの MWE の抽出はそれ程容易ではない．小山（2008）[5] にも述べられているように，MWE が crystal ball, freeze dry のような compounds（複合語）や，give up, put off, のような phrasal verbs （句動詞），kick the bucket, rain cats and dogs 等の idiom を含み，極めて多様であることにより，抽出自体に高度な技術を要するからである．そこで，本研究では MWE を lexical bundle（単語連鎖）としての 4-gram（連続した 4 語）に限定し，その抽出結果を考察することにする．

２．目的

Hyland (2008a) [4] が指摘するように，「単語連鎖」はある分野の中で「自然」な英語使用であるかどうかを決める鍵とも言うべき言語表現であるが，分野の異なる（もしくは似ている）6 種類のコーパスから抽出した 4-gram を考察することによって，それぞれの分野によって特徴的な 4-gram があるかどうか，またあるとすればどのような性質の単語連鎖があるのかを明らかにする．

３．方法

前述のように，MWE は専門分野における特徴的な英語表現を表すものであるが，その中でも，何語かの連続する語である「単語連鎖」（lexical bundles）が，書きことば，話しこ

とばといったディスコースやレジスターを示す重要な役割を担っていると考えられている. (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Cortes, 2004; Hyland, 2008b) [6][7][8][9]

単語連鎖を抽出する場合，何語の単語連鎖を分析の対象とすれば適切かということに関して，近年，いくつかのコーパスの比較研究により 4 語の連続した語（4-gram）という結果が得られている．例えば，Cortes (2004) [8] は，4-gram は 3 語の単語連鎖をその中に含み，しかも頻度が多く，分析すべき形式も多様であるため，その分析が有用であると主張している．また，Hyland (2008a)[4] は，“they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles” (p. 8) と述べており，やはり 4-gram の適切性を支持している．従って，本研究でも 4-gram を分析の対象にする．

コーパスは, (1) 高校教科書コーパス, (2) Scientific American（一般科学雑誌）, (3) Nature（専門科学雑誌），(4) 機械系論文コーパス，(5) 映画スクリプトコーパス，(6) 創作散文（imaginative prose）コーパスの 6 種類を使用した．この 6 種類のコーパスにおける特徴的な 4-gram を調べることで，専門的な論文から，一般雑誌，散文，口語表現の多い映画，そして日本の高校英語教科書まで広い範囲のディスコースにおける単語連鎖の特徴を見ることができる．【表 2】は本研究で使用したコーパスの情報をまとめたものである．

【表 2】分析に使用したコーパスの情報

	総語数（token）	異語数（type）	説明
高校教科書コーパス	424,774	17,858	高等学校英語 II の検定教科書
Scientific American	260,433	21,790	一般科学雑誌
Nature	476,921	15,245	専門科学雑誌
機械系論文コーパス	560,234	15,082	専門分野論文
映画コーパス	1,195,453	36,660	映画（100 本）のスクリプト
創作散文コーパス（Imaginative Prose）	1,038,420	41,515	Brown，Frown，LOB，FLOB のセクション K～R

4-gram は AntConc3.2.1 を使って 6 種類のコーパスそれぞれから抽出した. それぞれのコーパスはサイズが異なるため， 6 つのコーパスの総語数を基に，コーパスサイズが 100 万語になるように換算し，4-gram の調整頻度を算出した．そうして得た 4-gram の合計頻度の上位 100 ケースを対象とし，100 ケースの 4-gram（行）× 6 コーパス（列）の分割表【表 3】を用いてコレスポンデンス分析を行った．分析には R version 2.10.0. を用い，コレスポンデンス分析は，MASS ライブラリーの corresp 関数により実行した.

【表 3】分析に使用した 4-gram（行）× 6 コーパス（列）の分割表の一部

4-gram	High School Textbooks	Nature	Scientific American	Mech	Movie	Imaginative Prose
I don t know	169.50	0.00	4.19	0.00	828.97	151.19
as shown in fig	0.00	468.45	0.00	226.69	0.00	0.00
is shown in fig	0.00	433.89	0.00	235.62	0.00	0.00
as a function of	0.00	414.69	0.00	221.34	0.00	0.00
In the case of	11.77	426.21	8.39	137.44	2.51	2.89
I m going to	80.04	0.00	4.19	0.00	315.36	68.37
On the other hand	42.38	257.26	16.77	105.31	9.20	11.56
I don t think	70.63	0.00	4.19	0.00	261.83	73.19
as well as the	18.83	218.87	14.68	128.52	3.35	9.63
don t want to	87.11	0.00	12.58	0.00	245.93	43.34

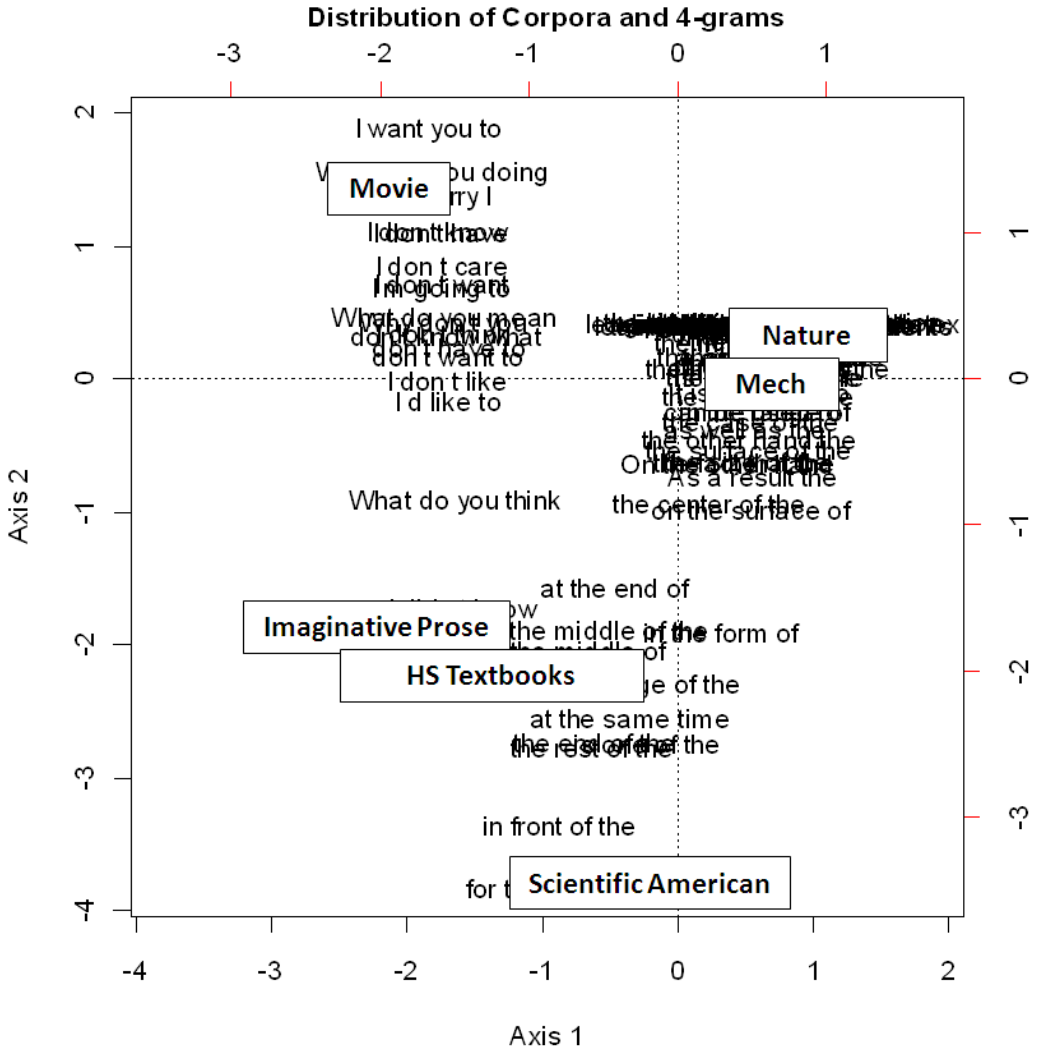
４．結果と考察

【表 4】は 100 ケースの 4-gram（行）× 6 コーパス（列）の相関係数を示したものである．相関係数行列を確認するだけでも，Nature と機械系論文コーパス（r = .96），高校教科書コーパスと創作散文コーパス（r = .89），映画コーパスと創作散文コーパス（r = .83）などの相関係数が高く，これらのコーパス間において使用されている 4-gram はそれぞれ類似度が高いことがわかる．他方，Nature 及び機械系論文コーパスは高校教科書コーパスとそれぞれ r = -.49　r = -.48 と負の相関であり，これは映画・創作散文コーパスとも同様で，互いに異なる 4-gram を使用していることがわかる．ただし，Scientific American は高校教科書コーパス（r = .40），創作散文コーパス（r = .33）とこれらのコーパスともある程度の相関を示しており，この点で Nature や機械系論文のコーパスとは異なる性格のコーパスであることがわかる．

次にコレスポンデンス分析を行った結果，第 1 次元と第 2 次元における行（4-gram）と列（6 つのコーパス）に与えられたアイテムの得点を基にしてプロットしたものが【図 1】である．第 1 次元の寄与率は 82.94%，また第 2 次元の寄与率は 12.10%となり，2 つの次元で 6 つのコーパスにおける 4-gram のデータが 95.04%説明されているため，コレスポンデンス分析によってうまくデータをまとめることができている．

【表 4】6 コーパスの相関係数行列（上位 100 ケース）

	High School Textbooks	Nature	Scientific American	Mech	Movie	Imaginative Prose
High School Textbooks	–					
Nature	-.49	–				
Scientific American	.40	-.09	–			
Mech	-.48	.96	-.11	–		
Movie	.76	-.44	.01	-.42	–	
Imaginative Prose	.89	-.45	.33	-.43	.83	–



【図 1】コレスポンデンス分析の結果

第 1 次元のプラス得点方向には，Nature と機械系論文コーパスが位置しており，マイナス得点方向には映画コーパス，創作散文コーパス，高校教科書コーパスなどがある．また，Scientific American コーパスは第 1 次元ではゼロ付近にある．

【表 5】は第 1 次元におけるプラス・マイナスの得点上位 10 位までの 4-gram であるが，プラスには it is assumed that, it is found that, in the present study, the uncertainty in the 等, 論文によく使われる表現であり理論的文章を書く「書き言葉」, マイナスには I want you to, What are you doing, I m sorry I, I don't know, I don't have 等，日常会話で使われる「話し言葉」であることが明確である．従って，第一次元の軸はプラスが「論文」マイナスが「日常会話」を示していることがわかる．内容は科学的なものであるが，文体としては堅い論文調もあれば，口語的な表現が入る雑誌風の記事もある Scientific American がゼロ付近にあることはこの解釈に合致するものと言えよう．

【表 5】第 1 次元におけるプラス・マイナス得点上位 10 位までの 4-gram

プラス得点上位 10 位（論文的）	マイナス得点上位 10 位（口語的）
it is assumed that	I want you to
it is found that	What are you doing
in the present study	I m sorry I
the uncertainty in the	I don t know
it can be seen	I don t have
the distribution of the	I m going to
the leading edge separation	I don t want
for the case of	I don t care
the right hand side	What do you mean
the effect of the	Why don t you

次に第 2 次元であるが，プラス・マイナスの得点上位 10 位までの 4-gram を【表 6】に示してある．第 2 次元を見ると，プラス得点の方向には映画コーパスがあり，マイナス得点の方向には創作散文コーパス，高校教科書コーパス，Scientific American コーパスがある．寄与率が 12.10%と低いこともあり，解釈は容易ではない．しかし，上位の 4-gram を通覧するとマイナスには for the first time, in front of the, the rest of the, is one of the 等, 場所や時間等を限定する表現が特徴的である．これらの表現は「客観的」にものごとを説明・描写するときに使う表現と考えられる. 他方, プラスの範囲には I want you to, What are you doing 等, 第 1 次元のマイナスと同じ 4-gram がプロットされていて，口語表現が主で

あるため「主観的」な表現と解釈することが可能性である.【表 6】には挙がっていないが,プラス得点第 2 次元の 11 位には it is assumed that, 同 12 位には it is found that という 4-gram がある.（巻末付録表参照）これらも自分の判断を主観的に述べる場合に使われることがあり,「主観性」を表しているとも言えよう．しかし Nature や機械系論文がいずれもゼロ付近に位置することを考えると,「主観性」「客観性」の解釈がやや困難になる．この点からするとむしろマイナス方向は「説明的」という解釈が妥当かもしれず,「主観性」「客観性」はある解釈の可能性を示唆するものであって，他の解釈の可能性を否定するものではない.

【表 6】第 2 次元におけるプラス・マイナス得点上位 10 までの 4-gram

プラス得点上位 10 位　（主観的）	マイナス得点上位 10 位（客観的）
I want you to	for the first time
What are you doing	in front of the
I m sorry I	the rest of the
I don t know	is one of the
I don t have	the end of the
I don t care	at the same time
I don t want	the edge of the
I m going to	in the middle of
What do you mean	in the form of
Why don t you	the middle of the

8．まとめ

以上の分析結果で述べられたように，(1) 高校教科書コーパス，(2) Scientific American（一般科学雑誌），(3) Nature（専門科学雑誌），(4) 機械系論文コーパス，(5) 映画スクリプトコーパス，(6) 創作散文（imaginative prose）は抽出された 4-gram の内容において，コレスポンデンス分析を行ったところ，第 1 次元の差異として「書き言葉としての論文」と「日常的話し言葉」があり，第 2 次元の解釈の一つの可能性としては「主観性」と「客観性」という観点での差異があることが確認された．また，Nature は専門科学雑誌としてカバーする分野は専門分野論文誌に比べて広いものの，4-gram の分析結果を見る限り，専門分野論文誌と極めて類似したディスコースを示すことが明らかとなった．さらに，Scientific American は先端的な科学論文に類する記事から身近な雑誌風の記事まで内容も

文体も多様性に富むコーパスであり，その意味で他の 5 種類のコーパスとは異なった独自の位置を占める結果になった．

以上述べたことから，4-grams のような単語連鎖（lexical bundles）がディスコースやレジスターを示すものであるという，Biber, Conrad, and Cortes (2004) [7] や Hyland (2008a) [4]らの先行研究を支持する結果が得られた．また教育目的の観点では、Nature と機械系論文の 4-gram 分析の結果が類似していることから、科学技術分野の共通した「科学技術表現リスト」作成の可能性が示されたと言える．

MWE の抽出と分析は，N-gram に留まらないことは最初に述べたとおりである．今後さらに不連続の，あるいは可変長の MWE 抽出の研究を進め，科学技術英語の特徴表現を確定して教育に資する「科学技術表現リスト」の作成を実現したいと考えている．

文　献

[1] Swales, J (1990). *Genre Analysis: English in Academic and Research Settings,* Cambridge University Press, Cambridge

[2] 石川有香・小山由紀江(2007).「学術論文読解を目的とした指導語彙の選定」中部地区英語教育学会『紀要』2006, p309-316

[3] Calzolari N., Fillmore C. J., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. (2002) Towards best practice for multiword expressions in computational lexicons. In Proceedings of *LREC 2002*, Las Palmas, Canary Islands, Spain

[4] Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*, 4–21.

[5] 小山由紀江（2008）.「Multi-Word Expression に関する統計と教育への応用」,『ESP コーパス語彙の頻度と習得困難度に基づく統計尺度』, 統計数理研究所共同研究リポート 216, p39-56

[6] Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*, 263–286.

[7] Biber, D., Conrad, S., & Cortes, V. (2004). If you look at. . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405.

[8] Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397–423

[9] Hyland, K. (2008b). Academic bundles: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*, 41–62.

【附録】4-gram コレスポンデンス分析による第１・２次元得点（第２次元の 50 位まで）

rank	4-grams	第1次元	第2次元
1	I want you to	−1.83316	1.883927
2	What are you doing	−1.80543	1.563999
3	I m sorry I	−1.79422	1.381523
4	I don t know	−1.77993	1.125126
5	I don t have	−1.75811	1.114854
6	I don t care	−1.74139	0.870918
7	I don t want	−1.74477	0.729271
8	I m going to	−1.74605	0.688566
9	What do you mean	−1.73462	0.469623
10	Why don t you	−1.73174	0.430047
11	it is assumed that	0.703934	0.426658
12	it is found that	0.703913	0.425707
13	in the present study	0.703876	0.424019
14	the uncertainty in the	0.703871	0.423786
15	it can be seen	0.703869	0.423666
16	the distribution of the	0.703847	0.422683
17	the leading edge separation	0.70384	0.42235
18	for the case of	0.703839	0.422314
19	the right hand side	0.703839	0.422312
20	the effect of the	0.703834	0.422069
21	it should be noted	0.70382	0.421462
22	the results of the	0.703817	0.421326
23	the static friction coefficient	0.703815	0.421202
24	local heat transfer coefficients	0.703814	0.421163
25	icrh gas turbine based	0.703792	0.420147
26	can be written as	0.703792	0.420147
27	pedras and de lemos	0.703792	0.420146
28	of the tip vortex	0.703792	0.420146
29	leading edge separation vortex	0.703792	0.420145
30	single degree of freedom	0.703792	0.420145
31	is assumed to be	0.703785	0.41985
32	be noted that the	0.703784	0.419797
33	can be expressed as	0.703781	0.419655
34	can be obtained by	0.703781	0.419655
35	is given by where	0.703781	0.419655
36	the variation of the	0.70378	0.419637
37	is assumed that the	0.70378	0.419616
38	as shown in fig	0.70378	0.419604
39	should be noted that	0.703771	0.419187
40	shown in fig b	0.703769	0.419101
41	can be seen that	0.703761	0.418758
42	a function of the	0.703758	0.418639
43	is related to the	0.703758	0.418596
44	are shown in fig	0.703753	0.418394
45	as a function of	0.70375	0.418249
46	is shown in fig	0.703745	0.418007
47	with respect to the	0.703737	0.417678
48	shown in fig a	0.703732	0.417435
49	is defined as the	0.703721	0.416935
50	good agreement with the	0.703721	0.416918