英文絵本の計量言語学的特徴抽出

伴 浩美

Metrical Linguistic Characteristics of English Picture Books

Hiromi Ban

Picture books play an important role as a material that develops children's linguistic competence. Thus, English picture books can be considered to be indispensable in the children's English study. In this paper, metrical characteristics of some English picture books were investigated, compared with English textbooks for Japanese junior high schools. In short, frequency characteristics of character- and word-appearance were investigated. These characteristics were approximated by an exponential function. Furthermore, the percentage of Japanese junior high school required vocabulary and American basic vocabulary was calculated to obtain the difficulty-level. As a result, it was clearly shown that the English picture books have a similar tendency to literary writings in the characteristics of character-appearance, and some books are more difficult than English textbooks.

Keywords: English for children, English text analysis, metrical linguistics, picture book, statistical analysis, text mining

1. はじめに

絵本は，子どもの言語能力を成長させる材料として大きな役割を果たしている[1]．良い絵本を繰り返し楽しく読み聞かせていると，絵本の中の美しい言語が知らず知らずのうちに身に付いて行く．絵本の中の言葉が自分の言葉として使われるようになり，それを繰り返し真似たり学習したりして，言葉が発達して行く[1]．このことは日本語の場合のみならず，英語運用能力の習得に関しても同様であり，英文絵本は子どもの英語学習において重要な役割を果たすものと考えられる．

本研究では，英文絵本の文体にはどのような特徴がみられるのか，計量言語学的な解析を行った．すなわち，英文絵本の一例として，Dick Bruna 氏の Miffy シリーズの英訳絵本の英語について，文字種や単語種，及びその出現頻度を調査した．

＊ 教養部

2. 解析方法

本研究において解析した試料は以下の通りである.

試料 1: *Miffy* (2003, オランダ語原著 1963) ～

試料 24: *Miffy's Garden* (2005, 原著 2004) の 24 試料

なお, 比較のため, 日本の中学校教科書 *NEW HORIZON English Course 1, 2, 3* (2009, 東京書籍) の本文の解析も行った.

解析プログラムは C++で構成されている. このプログラムからは, 各試料の文字と単語の頻度特性の他に, 文の数, 段落数, 平均単語長など様々な情報が得られるよう配慮されている[2].

3. 解析結果

3.1. 文字頻度特性

まず, 各試料における使用頻度の高い文字の種類とその頻度を調べた. 試料 1～24, 教科書いずれも, 空白が 1 位, e が 2 位となっている. 3 位については, 絵本の場合, t が 11 試料, a が 10 試料, o が 2 試料, h が 1 試料あり, 教科書では, o が 2 試料, a が 1 試料ある. さらに全ての試料について, n, i, s, r が上位に見られ, 上位 10 位については, 順位の違いが多少あるものの, 出現文字種はいずれの試料もほとんど違いが見られない.

各試料の上位 50 位までを頻度の高い順に, 縦軸は頻度の度合い, 横軸は順位で, 片対数でプロットした. 一例として, 試料 1 の結果を Fig. 1 に示す.

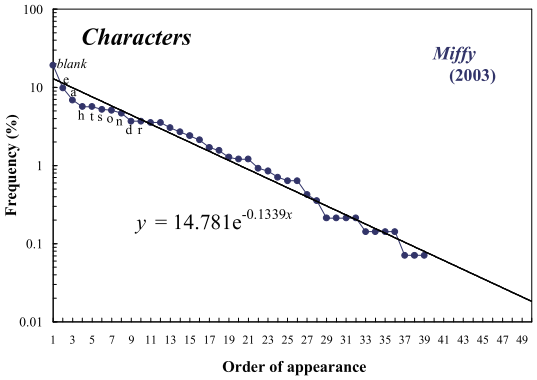

Fig. 1　Frequency characteristics of character-appearance in Material 1.

26 位と 27 位に, 減少度が異なるために生じる変曲点が見られ, 27 位以降は落ち込みが若干大きくなっている. この頻度特性を

$$y = c * \exp(-bx) \tag{1}$$

で指数近似を行った[3]. Fig 1 に示した試料 1 の場合, $c = 14.781$, $b = 0.1339$ という値が得られた.

それぞれの試料について得られた係数 c, b の値を Fig. 2 に示す.

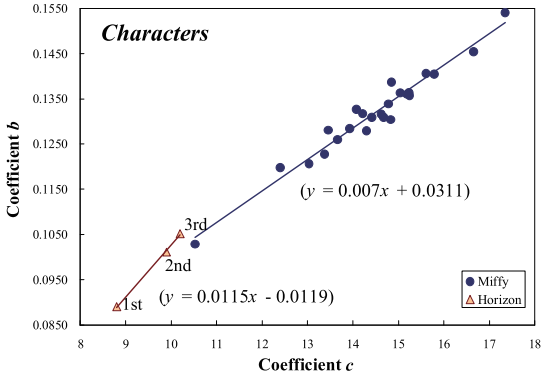

Fig. 2　Dispersions of coefficients c and b for character-appearance.

教科書を含む全ての試料の係数 c と b にほぼリニアな関係が見られ，これらの値は，絵本の 24 試料については $y = 0.007x + 0.0311$，教科書は $y = 0.0115x - 0.0119$ で近似される．絵本の試料は c が 10.525～17.349, b が 0.1029～0.1541，教科書は c が 8.799～10.194, b が 0.0890～0.1052 と，絵本の値の方が全体的に高く，教科書では学年が高くなるにつれて値が高くなっている．前報において著者は様々なジャンルの英文を解析し，それらの係数 c と b には正の相関が見られ，ジャーナリズムや技術英文に近いほど c と b の値が小さく，文学作品に近いほど，それらの値が大きい傾向にあることを示した[4]．従って，絵本の 24 試料は，文学作品に近い傾向があると言える．

3.2. 単語頻度特性

次に単語頻度特性を調べてみた．出現頻度上位 20 位までの単語を，絵本の例として試料 1, 5, 10, 15, 20, 24 の 6 試料，教科書の 3 試料について Table 1 に示す．

Table 1　High-frequency words for each material.

	Miffy 1	Miffy 5	Miffy 10	Miffy 15	Miffy 20	Miffy 24	*Horizon* **1st**	*Horizon* **2nd**	*Horizon* **3rd**
1	and	and	the	the	a	to	I	the	the
2	the	Miffy	and	and	Miffy	and	the	a	a
3	a	the	a	a	the	the	you	I	to
4	her	a	teacher	to	was	a	is	to	and
5	to	look	they	said	you	her	a	you	you
6	bunny	it	was	they	and	so	it's	and	in
7	with	Miff	that	aunt	said	she	we	in	I
8	all	uncle	to	her	aunt	carrots	I'm	it	is
9	chicks	you	then	party	her	Miffy	to	is	of
10	house	I	with	Alice	I	one	do	of	was
11	Mrs	in	school	all	Alice	them	in	but	it
12	said	said	all	danced	ghost	are	my	we	but
13	she	see	her	for	mother	some	have	can	for
14	was	cried	in	guests	sheet	up	yes	he	are
15	baby	oh	said	you	to	very	are	was	she
16	cow	was	she	but	Aggie	all	this	have	people
17	have	what	so	come	how	bunny	at	for	this
18	he	all	there	I'll	I'll	can	can	are	very
19	so	flying	too	it	it	carrot	like	on	have
20	they	just	up	on	like	father	and	about	my

絵本では and が非常に多く用いられており，試料 20 以外の 5 試料において 1 位または 2 位となっている．そのため，教科書において 1 位または 2 位となっている定冠詞の the が，絵本では 1～3 位となっている．また，教科書では I, you という 1, 2 人称代名詞の順位が高いのに対し，絵本では 3 人称代名詞 he, she, her, they の順位が高いものが多くなっている．さらに絵本では，Miffy, Alice という固有名詞や，said, cried, was といった過去形の動詞の頻度が高いという特徴が見受けられる．

先の解析と同様に，各試料の上位 50 位までを頻度の高い順に，縦軸は頻度の度合い，横軸は順位で，片対数でプロットし，(1)式で近似を行った．得られた係数 c と b の値を Fig. 3 に示す．

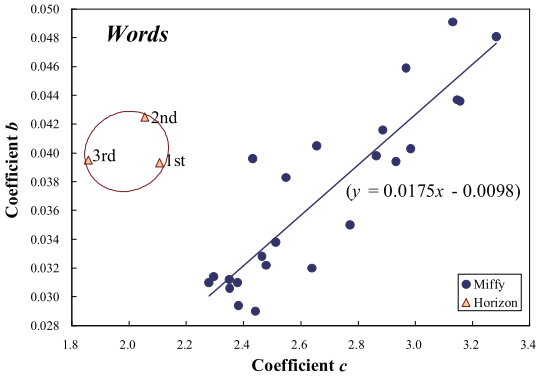

Fig. 3　Dispersions of coefficients c and b for word-appearance.

図より，絵本の 24 試料の係数 c と b に，文字の場合よりも弱い正の相関が見られ，これらの値は $y = 0.0175x - 0.0098$ で近似される．絵本の 24 試料の係数 c は 2.2786～3.2833 と，いずれも教科書（1.8579～2.1067）よりも高くなっており，教科書よりも単語種が少ないことが窺える．係数 b については，教科書の 0.0393～0.0425 と同程度のもの（0.0394～0.0416）が 6 試料，低いもの（0.0290～0.0383）が 13 試料，高いもの（0.0436～0.0491）が 5 試料であり，教科書よりも低い値の試料が多く見られる．一方，教科書の 3 試料は比較的近い値を取っており，Fig. 3 に示したような 1 つのクラスタと見なすことが可能であると思われる．

単語の特徴を表す方法として，統計学者の Udny Yule が 1944 年に，作家の語彙量を量る K 特性値（K-characteristic）と呼ぶ指標を提案し，これを用いて *The Imitation of Christ* の著者の推定を行っている[5]．この K 特性値は，或る作品の中に x_i 回使用された単語が f_i 個あるとすると，$S_1 = \Sigma\, x_i f_i$, $S_2 = \Sigma\, x_i^2 f_i$ として，次のように定義される．

$$K = 10^4 \,(S_2 / S_1^2 - 1 / S_1) \tag{2}$$

各試料について K 特性値を求めてみた．その結果を Fig. 4 に示す．図より，絵本の K 値は 81.502（試料 4）～ 130.255（試料 10）と，50 程度の幅があるが，いずれも 80 以上の値であり，教科書の 3 試料（61.189～73.403）に比べて高くなっている．絵本には 100 を超えるものが 15 試料あり，平均値は 100.924 となり，教科書の平均値 68.317 よりも 32.6 程度高い．

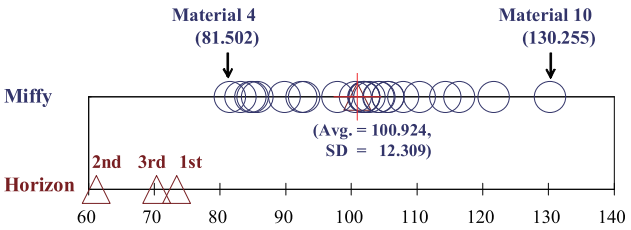

Fig. 4　*K*-characteristic for each material.

絵本の値の方が教科書よりも高いという結果は，先述の文字頻度特性の係数 *c*, *b* や，単語頻度特性の係数 *c* と同様である．この K 特性値と文字や単語の係数との関連性についての検討は，今後の課題である．

3.3. 難易度

単語の種類とその頻度から各試料の難易度を求めてみた．難易度を表すパラメータには，単語種数からの難易度（D_{ws}）と単語数からの難易度（D_{wn}）を考慮した．基準とする語彙には，日本の中学校での必修単語 508 語と，アメリカの 4～8 歳児を対象とした *The American Heritage Picture Dictionary*（American Heritage Dictionary, Houghton Mifflin, 2003）に掲載されている 798 語（以下，「基礎単語」と呼ぶ）を用いた．D_{ws} と D_{wn} の 2 種類の難易度は，全単語数（n_t），全単語種数（n_s），必修［基礎］単語数（n_{rs}），各必修［基礎］単語数（$n(i)$）とすると，

$$D_{ws} = (1 - n_{rs} / n_s) \tag{3}$$

$$D_{wn} = \{ 1 - (1/n_t * \Sigma\, n(i)) \} \tag{4}$$

より求められる[6].

さらに適切な指数を与えるために，D_{ws} と D_{wn} を変量として主成分分析を行った．分散共分散行列を用いて求めた第 1 主成分 z は，必修，基礎単語共に，$z = 0.7071*D_{ws} + 0.7071*D_{wn}$ となった．これより得られた主成分得点をそれぞれ 1 次元で表したものを Fig. 5 に示す．

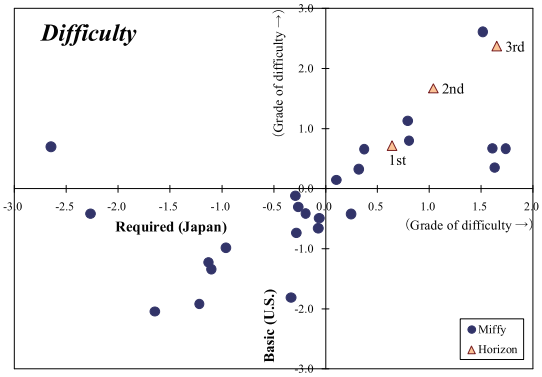

Fig. 5　Principal component scores of difficulty.

図より，必修単語，基礎単語共に教科書の難易度が学年が上がるにつれて高くなっており，必修・基礎単語の種類数とその頻度を難易度を求めるパラメータとすることの妥当性が認められる．絵本の 24 試料の必修単語と基礎単語による難易度については，弱い正の相関が見られる．必修単語を基準とした場合，1 年生の教科書よりも易しいものが 18 試料ある一方，1 年生より難しく 2 年生より易しいものが 2 試料，2 年生より難しく 3 年生より易しいものが 3 試料，3 年生より難しいものが 1 試料あり，絵本の難易度の平均値は-0.1389 である．基礎単語の場合は，1 年生より難しく 2 年生より易しいものが 2 試料，3 年生より難しいものが 1 試料ある他は全て 1 年生よりも易しく，平均値は-0.1985 となり，必修単語の場合よりも若干易しくなっている．全体として，絵本は中学校 1 年生用教科書よりも易しいものがほとんどであるが，3 年生程度のものも僅かながらあることが明らかとなった．

3.4. その他の特徴

各試料のその他の計量的数値を調べた．平均単語長，一文当たりの単語数等についての結果をまとめて Table 2 に記す．表中の “Miffy” の値は絵本 24 試料の平均値である．各試料における前置詞，関係詞等の使用頻度を求めたが，1 語ずつ意味を調べたわけではないため，前置詞，関係詞等とカウントしたものの中に，それ以外の品詞として用いられている単語も若干含まれている．

Table 2　Metrical data for each material.

	Miffy (Avg. of 24 materials)	***NEW HORIZON*** **1st**	***NEW HORIZON*** **2nd**	***NEW HORIZON*** **3rd**
Total num. of characters	1,403	6,621	14,361	13,386
Total num. of character-type	41	68	69	71
Total num. of words	281	1,301	2,877	2,594
Total num. of word-type	151	481	800	764
Total num. of sentences	13	239	395	317
Total num. of paragraphs	12	218	226	176
Mean word length	5.000	5.089	4.992	5.160
Words/sentence	23.982	5.444	7.284	8.183
Sentences/paragraph	1.052	1.096	1.748	1.801
Repetition of a word	1.857	2.705	3.596	3.395
Commas/sentence	1.710	0.272	0.223	0.331
Freq. of prepositions (%)	10.621	8.839	11.786	12.188
Freq. of relatives (%)	2.963	1.768	1.392	1.927
Freq. of auxiliaries (%)	1.687	0.923	1.529	1.119
Freq. of personal pronouns (%)	12.503	17.758	15.503	12.496

まず，平均単語長については平均値が 5.000 文字であり，中学校教科書の中で最も短い 2 年生の 4.992 文字とほぼ同じとなっている．絵本で最も短いものは 4.716 文字である．教科書で最長の 5.160 文字よりも若干多いものが 3 試料あり，5.173, 5.177, 5.412 文字となっている．

一文当たりの単語数は平均が 23.982 語と，かなり多くなっているのが特徴的である．9.852～34.875 語とかなり幅があるものの，最も少ないものでも 3 年生の 8.183 語よりも 1.7 語程度多く，また，29 語以上のものが 6 試料もある．この点から考えると，Miffy シリーズの英文絵本はかなり難解であると思われる．

一文当たりの単語数が多いため，一文当たりのコンマ数も 1.710 と，教科書の 0.223～0.331 と

比べてかなり多くなっている．

関係詞は関係代名詞，関係副詞，関係形容詞を合わせたものである．絵本の平均は 2.963%と，教科書の 1.392～1.927%よりも 1.0～1.5%程度高くなっている．従って，Miffy シリーズの絵本は複文が多く用いられ，読みにくさを感じる可能性が高いと考えられる．

広い意味での助動詞には2種類あり，一つは，進行形･受動態を形成する be，完了形の have，疑問・否定文の do などの時制や態を表すものである．今一つは，話者の気持ちや態度を表す will，can などの法助動詞である[7]．ここでは法助動詞のみを調査の対象とした．その結果，助動詞の頻度の平均が 1.687%と，教科書の 0.923～1.529%よりも高いことが明らかとなった．0.379～3.435%と幅があるが，教科書で最も少ない 0.923%よりも低いものが 5 試料（0.379～0.840%）であるのに対し，教科書で最も多い 1.529%よりも多いもの（1.749～3.435%）が 12 試料もある．従って，絵本の英文は，より多くの助動詞を用いて微妙なニュアンスを表しているものが多く，一方，教科書の英文は断定的な表現が多い傾向があると言える．

さらに，単語長の頻度特性についても調べてみた．その結果を Fig. 6 に示す．これは，単語長を変数として，その頻度を縦軸に取ったものである．

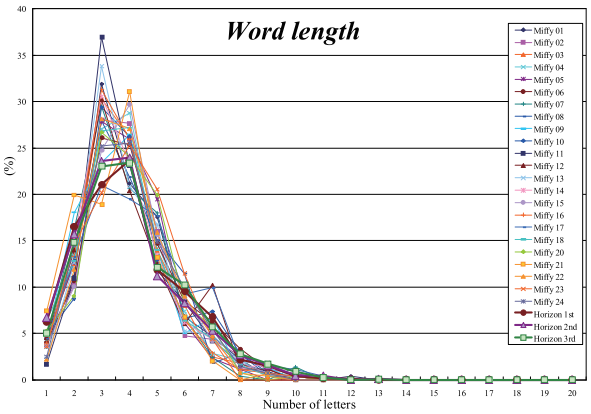

Fig. 6　Word-length distribution for each material.

試料 21 以外の 23 試料の全てが 3 文字あるいは 4 文字の頻度が最も高く，3 文字が 18.919～36.975%，4 文字が 19.485～31.081%となっている．教科書はいずれも 4 文字の頻度が最も高く，23.400～23.983%となっているが，3 文字も 21.061～23.566%と，4 文字と同程度である．絵本の 3 文字，4 文字について，教科書よりも高くなっているものが 3 文字で 18 試料，4 文字で 14 試料ある．一方，教科書の 6～9 文字の頻度が絵本よりも高い傾向が見られ，その結果，教科書の方が平均単語長が若干高くなっているものと考えられる．

また，絵本の 23 試料について，総単語数と総文字数，総単語数と総文数の相関を調べてみた．その結果を Fig. 7 に示す．これは，総単語数を変数として，総文字数を縦軸，総文数を第2縦軸に取ったものである．

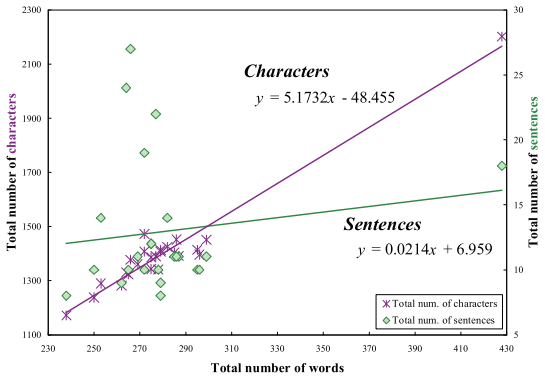

Fig. 7　Correlation of the number of words with the number of characters and sentences.

図より，総単語数と総文字数には強い正の相関が見られ，総単語数と総文数には弱い正の相関が見られる．24 試料の値に対して，Fig. 7 に示したような近似式が得られた．従って，Miffy シリーズの或る英文絵本の単語数が分かれば，$y = 5.1732x - 48.455$ という式を用いて，おおよその総文字数を，また，$y = 0.0214x + 6.959$ という式を用いて，おおよその総文数を求めることができる．

3.5. 各試料の類似度

以上の結果を基に，相関行列による主成分分析を行い，各試料のポジショニングを行ってみた．その結果を Fig. 8 に示す．

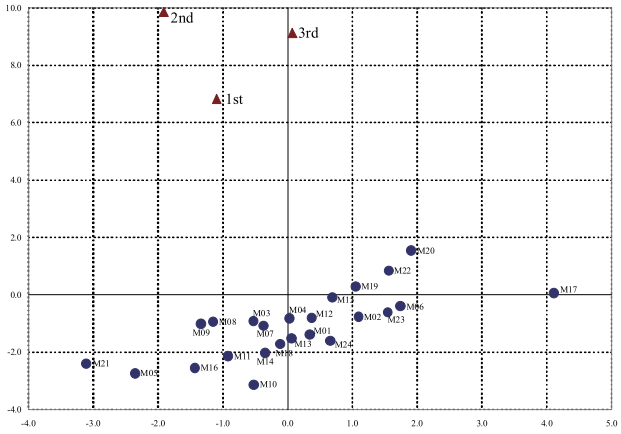

Fig. 8　Positioning of each material.

図より，絵本は中学校 1 年の教科書と近いものが多いことが窺われる．第 1 主成分得点について，絵本の 2 4 3 試料と教科書の 3 試料がそれぞれ近くになっている．教科書は 6.834～9.866 と高く，絵本で 0 を超えているものは試料 17, 19, 20, 22 の 3 試料のみで，得点は 0.069～1.546 と低くなっている．従って，第 1 主成分は教科書か絵本かを示す得点であると解釈できる．一方，第 2 主成分得点については，絵本の 13 試料が 0 を超えており，教科書は −1.918, −1.094, 0.072 と，2 試料

が 0 未満となっている.

4. まとめ

英文絵本の一例として, Miffy シリーズの英文絵本の英語について, 日本の中学校英語教科書と比較をしながら, 文字や単語の頻度特性を調べた. この時, 指数関数の近似式を採用し, 係数 c, b より, 各試料の特徴を抽出した. また, K 特性値を求めてみた. さらに, 試料中に使用されている日本の中学校必修単語 508 語やアメリカの基礎単語 798 語の種類数やその頻度より難易度を求めた. 結果として, 絵本の文字頻度特性には文学作品と似た傾向が見られた. K 特性値については, 教科書よりも高い値が得られた. また, 難易度については, 中学校 1 年生用教科書よりも易しいものが多いが, 中には 3 年生程度の難しいものもあることが明らかとなった.

今後も英文絵本の特徴抽出に関し, さらに研究を重ねていくとともに, 解析結果の児童英語教育への応用についても検討を行う予定である.

[参考文献]

[1] 絵本について http://www.j-k-s.net/kosodate_ehon.html

[2] H. Ban and T. Oyabu: Metrical Analysis of the Speeches of 2008 American Presidential Election Candidates, *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference*, 5 pages (2009)

[3] H. Ban, H. Nambo and T. Oyabu: Linguistic Analysis of English Pamphlets at Local Airports in Japan, *Proceedings of the 20th National Conference of Australian Society for Operations Research incorporating the 5th International Intelligent Logistics System Conference*, M2B, pp.4.1-4.9 (2009)

[4] H. Ban, H. Nambo and T. Oyabu: Metrical Linguistic Characteristics of English Materials for Business Management, *Proceedings of the 3rd International Symposium on Computational Intelligence and Industrial Applications*, 6 pages (2008)

[5] G. U. Yule: *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge (1944)

[6] H. Ban, T. Dederick and T. Oyabu: Metrical Linguistic Analysis of English Materials for Tourism, *Proceedings of the 7th Asia Pacific Industrial Engineering and Management Conference 2006*, pp.1202-1208 (2006)

[7] H. Ban, R. Tabata, K. Hirano and T. Oyabu: Linguistic Characteristics of English Articles on the Noto Hanto Earthquake in 2007, *Proceedings of the 8th Asia Pacific Industrial Engineering & Management System & 2007 Chinese Institute of Industrial Engineers Conference*, Paper ID: 905, 7 pages (2007)

（平成 24 年 3 月 31 日受理）