書籍テキストへの分類指標人手付与の試み
―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―

柏野　和佳子[1]　　奥村　学[2]

[1] 国立国語研究所　・　東京工業大学　大学院総合理工学研究科　　[2] 東京工業大学　精密工学研究所

1.はじめに

大規模なコーパスを様々な学術研究や教育に活用するためには，テキストを所望の目的で分類できるように多くの情報が付与されていることが望ましい(EAGLES1996)。『現代日本語書き言葉均衡コーパス』(BCCWJ)[1]の図書館サブコーパスには 10,551 サンプルの書籍テキストが収録されている。そこにはNDC（日本十進分類法）や著者情報，形態論情報などが付与されており，それらを利用して，半自動的に種々の観点から分類することは可能である。しかしながら，EAGLES(1996)がコーパスへ付与することが望ましいと挙げる，A.対象読者に想定される読解レベル（難易度），B.テキストの作成意図，C.さまざまな文体情報の 3 種に関する情報は「Cコード[2]」以外には与えられておらず，それらの観点によるテキストの分類や抽出は困難である。そこで，A.を補う「対象読者（難易）」，B.を補う「主観的・客観的」，C.を補う「硬軟」「丁寧さ」「直接的な語り性の有無」という，あわせて 5 つの分類指標を新たに設計し，現在人手による付与を行っている。

本稿では，提案する分類指標の設計と，試行したアノテーション作業の概要と結果について述べる。そして，我々の分類指標が各種のテキストに対していかなる値を示すかをNDC別に示す。

2．分類指標の設計

分類指標を次のとおり設計した。

(a) 対象読者（難易）

テキストの難易度を想定読者のスケールで測ることとし，次の 5 段階の選択肢を設けた。

1 専門家向き
　読む前提に高度な専門知識が必要なもの
　それを仕事にしているような人向きのもの
2 やや専門的な一般向き
　読む前提に多少の専門知識が必要なもの
3 一般向き
　特に専門的な内容ではないもの
　専門的な内容であっても，読む前提に専門的知識を特に必要とせず，一般向きに書かれているもの
4 中高生向き
　中高生向きに書かれているもの
　専門性の有無にかかわらず，中高生でも読めそうなもの
5 小学生・幼児向き
　明らかに小学生や幼児向きとして書かれているもの

このときの「専門性」はテキストを理解する上での「高度な知識の必要性」の有無と考える。

(b) 主観的・客観的

テキストの作成意図を捉えるための指標を検討した。「論説，随筆，報告文，紀行文，手順書・・・」といったような体系的な分類案を作成し，それに基づいた指標の付与が理想であるが，そういった指標の設計やその判断にかかる負荷が大きいと予測される。よって，その代わりに，作成意図の根本には書き手の態度が「主観的」か「客観的」かとの区別があると考え，その判断付与を行うこととした。次の4段階の選択肢を設けた。

1 とても客観的
2 どちらかといえば客観的
3 どちらかといえば主観的
4 とても主観的

ここで「客観的」とは，主に，事実，観察，論証などが述べてあるもの。誰が読んでも納得できる妥当性が高いもの。「主観的」とは，主に，経験や感想などが述べてあるもの。妥当性は筆者の自由と定義する。なお，これはノンフィクションと判断をしたテキストについてのみ付与する。

(c) 硬軟と丁寧さ

テキストの文体を捉える指標として「硬軟」と「丁寧さ」を設けた。「硬い」とは，かしこまっている感じ，堅苦しい感じであり，「軟らかい」とは，かしこまっていない感じ，親しみやすい感じである。また，「丁寧」とはフォーマルな感じであり，「くだけている」とはフォーマルではない感じである。この時，「硬くてくだけている」というテキストは想定できなかったため，次のとおり選択肢を設けたが，「硬軟」と「丁寧さ」は異なる軸として次の付与作業段階で

[1] 詳細は http://www.tokuteicorpus.jp/。
[2] 日本図書コード。「販売対象」が付されている。

はその選択肢を分ける計画でいる。

1 とても硬くて丁寧
2 どちらかといえば硬くて丁寧
3-1 どちらかといえば軟らかくて丁寧
3-2 どちらかといえば軟らかくてくだけている
4-1 とても軟らかくて丁寧
4-2 とても軟らかくてくだけている

(d) 直接的な語り性の有無

テキストの文体でよく取り上げられるものに,「書き言葉的」か「話し言葉的」かという観点がある。普通は書き言葉テキストと,話し言葉テキストとの差異によって分析されるものである。脚本などを除き,おおよそ書籍テキストは「書き言葉」一つにくくられるのが普通である。しかしながら,書籍テキストの中にも,話し言葉口調,おしゃべり口調の文体のテキストが存在する。そこで,書籍テキストの会話部分以外の地の文において話し言葉口調,おしゃべり口調のみられるテキストを区別するために,試行作業の当初は,次の選択肢を用意した。

1 どちらかといえば書き言葉的
2 どちらかといえば話し言葉的

約2,400テキストについて上記選択肢によって作業を進めた後,作業者にとってこの「話し言葉的」と判断する基準がかなりあいまいであることがわかったため,そのあとの970テキストについては,次の選択肢に改めた(柏野2010)。

1 直接的な語り性あり
2 直接的な語り性なし

「話し言葉的」である判定基準を,「直接的な語り性あり」のものと定めた。たとえば,「あなた」や「みなさん」などの呼びかけ表現や,「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など,読み手に直接的に語りかけているような表現があるか否かで判断するようにした。

3.アノテーション作業の概要

アノテーション作業の概要は,次のとおりである。

●作業目的:人手付与の作業上の問題点の検討,典型例の抽出,分類指標の検証及び基準の検討。
●対象テキスト:BCCWJに収録されている図書館サブコーパス(10,551サンプル)よりランダムに抽出したサンプルのテキスト。(本稿作成時,合計3,324テキストへ付与済み。)
●1テキストの範囲と長さ:コーパス収録テキストの分類指標とするため,その一部を字数を揃えて抽出することはせず,1サンプル全体を範囲とする。1テキストの平均は約3,000語。
●作業ファイル:サンプルを取得した書籍の紙面コピーの電子化ファイルを参照する。
●作業態勢:判断のゆれを検証するために1作業につき,作業者3人を確保した。同一の判定作業を3人がそれぞれ独立して行う。
●作業量:1セット約400〜500の書籍テキストに対する指標付与を延べ約10日で行う。
●作業指示:付与すべき指標の種類をごく簡単な説明のみで指示。

また,作業手順は次のとおりである。

①形式によって区別する。構造的に単純なテキストタイプ(例:章節構造)であれば指標付与の対象とする[3]。
②「対象読者」「主観的・客観的」「硬軟」「丁寧さ」「直接的な語り性の有無」の分類指標を付与する。

4.アノテーション作業の結果の考察

本試行作業の目的である,人手付与の作業上の問題点の検討として,判断のゆれを検証した。また,各分類指標の典型例の抽出を行った。以下,その二点について述べる。

4.1 判断のゆれの検証

3人の作業者の判断のゆれについて検討をした(Kashino and Okumura 2010)。アノテーション試行作業の1セットめより161テキスト(うち,主観的・客観的の対象となるノンフィクションは49テキスト)を抽出した。最初の3人による試行作業をStep1と呼ぶこととする。次に,同じ161テキストについて,別の2人の作業者が,先の3人の判断結果を参照しながらアノテーションを行うという追加作業を試みた。これをStep2と呼ぶこととする。

Step1,Step2において,対象読者,主観的・客観的,硬軟と丁寧さ,話し言葉的・書き言葉的,それぞれについての作業者間の判断の一致度を,一致率,カッパ係数(一致率から偶然の一致率をひいたもの),相関関数で求めた。その結果を表1に示す。

Step1は,明確な判断基準や典型例の例示のない状況下での試行作業であったため,判断の一致は高くないことは予想してはいたが,実際に中〜低度の一致であった。しかしながら,Step2ではいずれの値も飛躍的に改善された。3人の作業結果を見て判断するというStep2は,マニュアル参照に近いことを疑似的に行ったと考えることができる。よって,

[3] 対象外とした形式が特徴的なテキスト(例:対談,Q&A形式,図解,用語解説)については,一定量が分類されてから細分類を検討する予定でいる。

このことから，適切なマニュアルを整備した上でアノテーション作業を進めることにより，アノテーション作業の質を高めることは十分可能であることが確認できた。

表1 Step1，Step2の作業者間の一致度

対象読者	Step1				Step2
作業者の組	1-2	2-3	3-1	Average	4-5
一致率	0.73	0.93	0.73	0.80	0.88
カッパ係数	0.17	0.55	0.18	0.30	0.58
相関	0.33	0.58	0.29	0.40	0.67

主観・客観	Step1				Step2
作業者の組	1-2	2-3	3-1	Average	4-5
一致率	0.49	0.33	0.24	0.35	0.47
カッパ係数	0.46	0.19	0.21	0.28	0.47
相関	0.61	0.29	0.48	0.46	0.70

硬軟	Step1				Step2
作業者の組	1-2	2-3	3-1	Average	4-5
一致率	0.37	0.25	0.16	0.26	0.71
カッパ係数	0.24	0.21	0.07	0.17	0.67
相関	0.59	0.58	0.37	0.51	0.79

話し・書き	Step1				Step2
作業者の組	1-2	2-3	3-1	Average	4-5
一致率	0.73	0.32	0.07	0.37	0.89
カッパ係数	0.03	0.02	0.00	0.02	0.62
相関	0.13	0.06	0.02	0.07	0.63

4.2 典型例の抽出

典型例を抽出するために，各分類指標の各選択肢において，3 人全員の判断が一致したテキスト，2 人が一致したテキストを同定した。分類指標別にその結果を表2〜5に示す。

表2 「対象読者」の判定結果別テキスト数

	1.専門家向き		2.やや専門的な一般向き		3.一般向き		4.中高生向き		5.小学生・幼児向き	
全員一致	1	0.0%	23	0.9%	1904	71.3%	11	0.4%	2	0.1%
2人一致	6	0.2%	138	5.2%	503	18.8%	23	0.9%	6	0.2%
計	7	0.3%	161	6.0%	2407	90.1%	34	1.3%	8	0.3%

表3 「主観的・客観的」の判定結果別テキスト数

	1.とても客観的		2.どちらかといえば客観的		3.どちらかといえば主観的		4.とても主観的	
全員一致	21	1.3%	84	5.1%	70	4.2%	53	3.2%
2人一致	150	9.1%	578	34.9%	271	16.4%	105	6.3%
計	171	10.3%	662	40.0%	341	20.6%	158	9.5%

表4 「硬軟と丁寧さ」の判定結果別テキスト数

	1.とても硬くて丁寧		2.どちらかといえば硬くて丁寧		3-1.どちらかといえば軟らかくて丁寧	
全員一致	2	0.1%	250	9.4%	60	2.2%
2人一致	50	1.9%	707	26.5%	231	8.7%
小計	52	1.9%	957	35.8%	291	10.9%
	3-2.どちらかといえば軟らかくてくだけている		4-1.とても軟らかくて丁寧		4-2.とても軟らかくてくだけている	
全員一致	53	2.0%	4	0.1%	18	0.7%
2人一致	326	12.2%	34	1.3%	35	1.3%
小計	379	14.2%	38	1.4%	53	2.0%

表5 「直接的な語り性」の判定結果別テキスト数

	1.直接的な語り性あり		2.直接的な語り性なし	
全員一致	1	0.3%	311	82.7%
2人一致	20	5.3%	44	11.7%
計	21	5.6%	355	94.4%

表2〜表5をみると，判定の一致したテキスト数が選択肢によって差のあることがわかる。「一般向き」や「直接的な語り性なし」といった，該当テキスト数がそもそも多数あるようなものは全員一致のテキスト数が多い。逆に該当するものが少なそうなもの，特に「専門家向き」，「小学生向き」「とても硬くて丁寧」「とても柔らかく丁寧」「直接的な語り性あり」は，全員一致のテキスト数が非常に少ない。そのようなものこそ典型例の抽出と分析の必要なものであるため，それらは2人一致のテキストからも典型例の候補を抽出した。そのようにすることにより，今回の試行作業において，全分類指標の典型例をそろえることができた。現在，各典型例の言語的な特徴分析を進めている[4]。

5. 分類指標から捉えるNDC別書籍テキストの特徴

最後に，コーパスに収録されているNDC別の書籍テキストの特徴を，本分類指標によって詳細に捉え得ることを報告する。現時点までの分類指標の付与済みテキスト数は3,324である。そのNDC別の内訳は表6のとおりである。表6にあるとおり，総記，言語，null（NDC未付与）のテキスト数は少ないため，以降この3つ以外のNDCのテキストを分析対象とする。

表6 NDC別付与済みテキスト数

番台	類	テキスト数
0番台	総記	8
100番台	哲学	209
200番台	歴史	346
300番台	社会科学	773
400番台	自然科学	196
500番台	技術，工学	178
600番台	産業	97
700番台	芸術，美術	204
800番台	言語	4
900番台	文学	1306
null	null	3
合計		3324

3 人の判断一致，不一致に関わらず，各人の選択結果1つを1点として，各テキストの分類指標の選択肢を点数化した。それを割合になおし，平均との差分を求めた。さらに，各分類指標がどちらに触れているかの尺度を次のとおり求めた。なお，判断のゆれを考慮し，選択肢別に重みづけは行わなかった。

「専門度」（選択肢1〜2の和と3〜5の和との差分）

「客観度」（選択肢1〜2の和と3〜4の和との差分）

「硬度」（選択肢1〜2の和3〜4の和との差分）

「丁寧度」（選択肢1，2，3-1，4-1の和と3-2，4-2の和との差分）

「語り度」（選択肢1と2の差分）

以下，求めた尺度の結果をNDC別に図示する。

4 「硬軟と丁寧さ」については柏野ほか(2012)を，「直接的な語り性の有無」については保田ほか(2012)を参照。

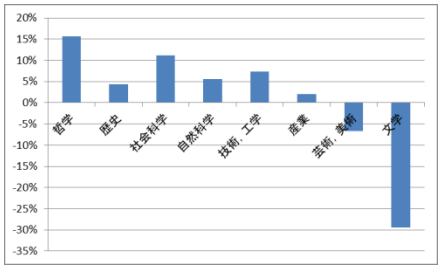

図 1 NDC 別専門度の違い

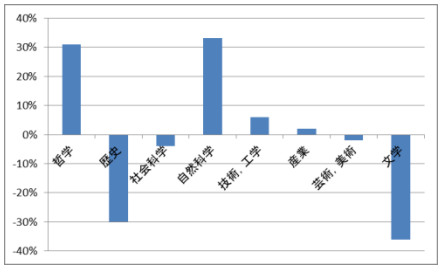

図 2 NDC 別客観度の違い

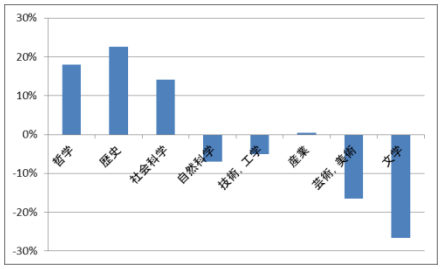

図 3 NDC 別硬度の違い

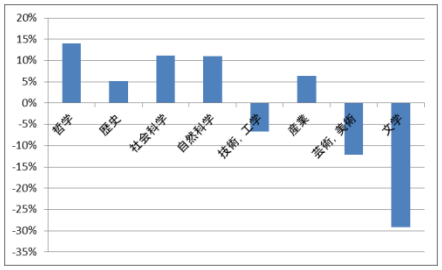

図 4 NDC 別丁寧度の違い

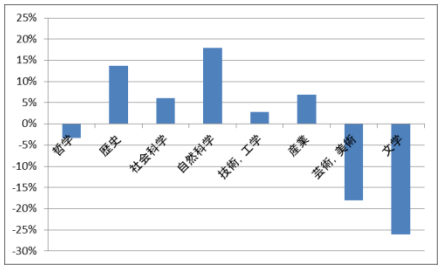

図 5 NDC 別語り度の違い

図 1～図 5 より，BCCWJ 図書館サブコーパスに収録された各 NDC の書籍テキストには，次のような特徴のあることを明らかにすることができた。

●専門度：文学以外の専門度が高いことは予測されたが，中では哲学がもっとも高い。また，芸術，美術の専門度が低いことは予測外。

●客観度：小説類は対象外であるため，「文学」はエッセー類のみ対象とした結果である。それらエッセー類は客観度が低いと判断されるのは予測通り。しかし，歴史もまた低いということが判明。

●硬度と丁寧度：両者は相関する傾向がみえるが，その中で，自然科学は硬度が低く丁寧度が高いという傾向をもつことが目立つ特徴。

●語り度：「丁寧度」と相関する傾向がみえる。自然科学や歴史の語り度が高いことが判明。また，哲学は丁寧度は高いが語り度は低めであり，逆に，技術，工学は丁寧度は低いが語り度は高めであるということが判明。

6. おわりに

BCCWJ に収録する書籍コーパスの有効活用を可能とするための分類指標の人手付与について，その概要，結果，効果について報告した。

抽出できた典型例の分析を進め，人手及び機械処理で付与する分類指標の正確さの向上を目指す。そして，少なくとも BCCWJ の図書館サブコーパスに収録される 10,551 サンプルの全てに分類指標を付与し，コーパスの研究や教育の利用価値を高めることを目指す。

さらに文体的な特徴を支える言語表現の分析を進め，辞書記述への応用を考えている。

[**謝辞**] 研究補助をしてくださった立花幸子さん，保田祥さんと本アノテーション作業の協力者に感謝します。本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものです。また，BCCWJ の構築は，文部科学省科学研究費補助金特定領域研究 「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度，領域代表者：前川喜久雄）による補助を得たものです。

[**参考文献**]

EAGLES(1996), Preliminary recommendations, *Preliminary Recommendation on Text Typology EAG-TCWG-TTYP/P*, Version of Jun 1996.

Wakako Kashino and Manabu Okumura(2010), An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese, *Proc. of PACLIC24*, pp. 433-438.

柏野和佳子ほか(2012)「テキストの硬さと軟らかさの考察―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『コーパス日本語学ワークショップ予稿集』.

保田祥ほか(2012)「「語り性」を有する書きことばの典型例の分析」『コーパス日本語学ワークショップ予稿集』.