BCCWJ 図書館サブコーパスの 文体情報検索ツールによるテキスト分析

柏野 和佳子*（国立国語研究所 言語資源研究系）
中村 壮範（国立国語研究所 コーパス開発センター）

Text Analysis of the Library Subcorpus of the BCCWJ Using a Text Search Tool

Wakako Kashino （Dept. Corpus Studies, NINJAL）
Takenori Nakamura （Center for Corpus Development, NINJAL）

１．はじめに

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」（2009-2012 年度）において，『現代日本語書き言葉均衡コーパス』（BCCWJ）の図書館サブコーパス（10,551 サンプル）に収録されている全ての書籍サンプルを対象に，文体情報を付与した（柏野ほか 2012a，柏野ほか 2012b）。文体情報とは，テキスト構造が単純（例：章節構造）なもの（全体の約 8 割）に付与した「専門度，客観度，硬度，くだけ度，および語りかけ性度」といった評定値や，テキスト構造・紙面形式などの点で上記分類になじまないもの（約 2 割）にその特徴を表すものとして付与した「対談，Q&A 形式，図解，用語解説」等の分類情報である。今回，我々は，それらの付与した文体情報をもとに，特定の文体的特徴をもつテキストを検索し，形態論情報等を一覧できる文体情報検索ツールを構築した。本ツールを用いて概観できる，図書館サブコーパスの文体的特徴について報告する。

２．付与した文体情報

２．１　アノテーション作業の概要

文体情報のアノテーション作業の概要は次のとおりである。

●対象：BCCWJに収録されている図書館サブコーパス（10,551サンプル）の書籍サンプル[1]。

●1サンプルの範囲と長さ：サンプル全体の印象判定を行うため，1サンプル全体を範囲とする。サンプルの一部より一定量の字数を揃えて抽出するようなことはあえてしない。1サンプルの平均はおよそ3,000語。

●作業者：言語データ構築経験有のおおよそ20～50代の女性，延べ9名。

●内容：

①文体判断が可能と判断されるもの，即ち，テキスト構造が単純（例：章節構造）なものを内容・表現の文体的特徴の印象判定により細分類する。（→2.2節）

②①以外，文体判断が単純にいかないと判断されるもの，即ち，テキスト構造・紙面形式に特徴をもつものを選別，分類する。（→2.3節）

●態勢：

①典型例抽出を目的とする作業（Kashino and Okumura 2010）

・判断のゆれを検証するために同一サンプルを作業者3人で判定。最終的には3人の結果を見直

* waka@ninjal.ac.jp

[1] 1986 年から 2005 年までの 20 年間に発行された書籍のうち，東京都内の 13 自治体以上の公共図書館で共通に所蔵されていた書籍を母集団とし，そこから抽出したサンプルから成るサブコーパスである。

し判定値を一つに決定。3,324サンプルまでこの態勢。（全体の31.5%）

②全体付与を目的とする作業（柏野ほか2012b）

・1サンプル作業者1人で判定。残りの①以外を作業し，全10,551サンプルを完了。

２．２　文体的特徴を表す分類指標

BCCWJ に収録されている書籍サンプルには，NDC（日本十進分類法）によるジャンルや，C コード（日本図書コード）による販売対象[2]，また，著者情報，形態論情報などに加え，テキストの内容・表現の文体的特徴を表す指標として，次のものを設定した。そのうえで，印象判定を行い，その結果を付与した。

(a) 専門度　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

(b) 客観度　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

(c) 硬度　1 とても硬い，2 どちらかといえば硬い，3 どちらかといえば軟らかい，4 とても軟らかい

(d) くだけ度　1 とてもくだけている，2 どちらかといえばくだけている，3 くだけていない

(e) 語りかけ性度　1 とても語りかけ性がある，2 どちらかといえば語りかけ性がある，3 特に語りかけ性はない

２．３　文体判断が単純にいかない特徴をもつものの分類指標

文体判断が単純にいかないと判断されるもの，即ち，テキスト構造・紙面形式に特定の特徴をもつものがある。そこで，次のような分類指標を設け，該当するサンプルの選別，分類を行った。理由は，それらは，必要に応じて，その特徴による類型で整理分類をし，別途，その文体を吟味すべきと考えたからである。なお，いずれの分類指標も該当サンプルすべてに付与したため，複数の分類指標が付与されたサンプルも存在する。

［テキスト構造・紙面形式上の特徴］

(a) 対話系（対話，対談・座談，インタビュー，往復書簡，シナリオ，その他対話形式）

(b) 引用系（Q&A 形式，投稿形式，その他引用編集形式）

(c) 視覚表現多用系（コマ割多用，図解，その他写真やイラストの多用）

(d) データベースやリスト系（用語解説，辞書形式，見本・カタログ形式，その他リスト形式）

［内容や表現上の特徴］

(e) 前書きや後書きである

(f) 明治時代より以前の古い言葉が多い

(g) 外国語が多い

(h) 数式やプログラミング言語などが多い

(i) 法律文が多い

(j) 教育現場で使いがたそうである[3]

(k) その他一定量の「本文」が認めがたい

---

[2] C コードの第 1 桁は販売対象を示す。0:一般　1:教養　2:実用　3:専門　5:婦人　6:学参 I（小中）　7:学参 II（高校）　8:児童　9:雑誌扱い

[3] 例えば，暴力的な描写や性的な描写を含むものを区別するための指標である。文体情報付与のための指標というよりは，コーパス活用のためのテキスト整理の指標として設けたものである。

２．４　アノテーション文体情報のリスト

アノテーション文体情報はリスト形式でまとめている。次の表 1 にはじめの 10 件分を例として示す[4]。このリストを用いることにより，特定の文体的特徴をもつテキスト群の，該当するテキスト数や，コーパス収録時に付与されているサンプル ID，書名，NDC（日本十進分類法），C コード（図書分類コード）といった書誌情報を得ることまでは簡単にできるようになっている。しかしながら，このようなリストだけでは，絞り込んでいったテキストの内容分析にすぐにとりかかることができない。テキストを分析したい場合は，ここで得られたサンプル ID をもとに，別途，BCCWJ の収録 DVD より該当テキストを抽出するという大きな手間が必要になる。そこで，アノテーション文体情報をもっと簡便に活用できるように，特定の文体的特徴をもつテキストを検索し，形態論情報等をすぐさま一覧できる文体情報検索ツールを構築することとした。

表 1 アノテーション文体情報のリストの例

番号	書誌情報				文体的特徴					備考	非対象とする該当理由										
	SampleID	NDC	Cコード	タイトル	専門度	客観度	硬度	くだけ度	語りかけ性度	小説の主人公あるいは語り手の人称	対話系か	引用系か	視覚表現多用系か	データベースやリスト系か	前書きや後書きか	明治時代より以前の古い言葉が多いか	外国語が多いか	数式やプログラミング言語などが多いか	法律文が多いか	教育現場で使いたそうであるか	その他，一定量の「本文」が認めがたいか
1	LBa2_00007	288	1323	光明皇后	3 一般向き	2 どちらかといえば客観的	2 どちらかといえば硬い	3 くだけていない	3 特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否
2	LBa2_00027	292		NHK大黄河	3 一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	2 どちらかといえばくだけている	2 どちらかといえば語り性がある	[私+達]	否	否	否	否	否	否	否	否	否	否	否
3	LBa5_00003	596		日本の郷土料理	3 一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	3 くだけていない	3 特に語り性はない	私	否	否	否	否	否	否	否	否	否	否	否
4	LBa7_00003	762		フルトヴェングラー	3 一般向き	3 どちらかといえば主観的	3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否
5	LBa9_00002	933		呪われたブルー・エラー	3 一般向き		3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	わたし	否	否	否	否	否	否	否	否	否	否	否
6	LBa9_00062	913	0193	汚名	3 一般向き		2 どちらかといえば硬い	2 どちらかといえばくだけている	3 特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否
7	LBa9_00068	910	3391	日本文芸史	2 やや専門的な一般向き	3 どちらかといえば主観的	2 どちらかといえば硬い	3 くだけていない	2 どちらかといえば語り性がある	わたくし	否	否	否	否	否	否	否	否	否	否	否
8	LBa9_00107	913		アローン・アゲイン	3 一般向き		3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	わたし	否	否	否	否	否	否	否	否	否	否	否
9	LBan_00008			時間と空間の冒険	5 小学生・幼児向き		3 どちらかといえば軟らかい	2 どちらかといえばくだけている	3 特に語り性はない	わし	否	否	否	否	否	否	否	否	否	否	否
10	LBan_00022		8093	ちょっと気になる転校生	5 小学生・幼児向き		4 とても軟らかい	1 とてもくだけている	3 特に語り性はない	おれ	否	否	否	否	否	否	否	否	否	否	否

３．Web アプリケーション「図書館コーパス検索ツール」の構築

アノテーション文体情報のより簡便な活用を目的に，特定の文体的特徴をもつテキストを検索し，形態論情報等を一覧できる文体情報検索ツールとして，Web アプリケーション「図書館コーパス検索ツール」を構築している。先行事例としては， 1 億語規模の BNC に対し，Lee(2001)が 70 種類（書き言葉 46 種，話し言葉 24 種）の言語使用域や，対象読者の性別や年齢層別，著者の属性等によって分類したものがある。

３．１　「図書館コーパス検索ツール」の仕様

「図書館コーパス検索ツール」は，Internet Explorer などのブラウザから利用することができる Web アプリケーションである。インターネットが利用できる環境と標準的なブラウザがあれば，特別なソフトをインストールすることなく利用することができる。ただし，現在はまだ開発段階であるため，研究所内のネットワークからのみアクセス可能である。

BCCWJ の図書館サブコーパスに付与された文体情報及び，書誌情報に基づく検索機能と，

4 表の 11 列目の「備考：小説の主人公あるいは語り手の人称」とは，アノテーション作業時にわかる範囲で，小説の場合は主人公を，それ以外は筆者自身をどの人称で表現しているかの抽出を試みたものである。

文体情報と形態論情報(短単位)を利用した簡単な集計とグラフ表示の機能を備える。

システムの概要は以下のとおりである。

【ソフトウェア】

・開発ツール	Microsoft Visual Studio 2012
・フレームワーク	asp.net
・サーバ OS	Microsoft Windows Server 2008R2 Standard
・Web サーバソフト	Microsoft Internet Information Services ver7.5
・DBMS	Microsoft SQLServer2012

【ハードウェア】

・サーバ本体	DELL PowerEdge R310
・CPU	Intel Xeon X3430(2.40GHz，8MB キャッシュ，1333MHz)
・メモリ	24GB
・HDD	300GB 15000RPM(6Gbps SAS HDD/3.5 インチ)×3 ※RAID5 構成で実質容量 600GB

【動作確認】

Google Chrome32，Internet Explorer10

３．２　「図書館コーパス検索ツール」の表示画面

次のように 3 つの表示画面から構成される。

⑴検索画面

文体情報及び，書誌情報に基づく検索が可能。

図 1 は，NDC の一桁目が「6」であり，文章の硬軟が「とても硬い」という指定を行っている画面である。左下の「検索」の BOX に，自動的に検索式を生成し，表示する。

⑵検索結果の表示画面

図 2 のように，該当サンプルのリストを表示する。該当件数が多い場合は，複数ページに分割表示する。

⑶データ表示画面

図 1 の左下にある，「データ表示」というリンクテキストをクリックすると，検索結果のテキストデータの各種情報を表示する。現在のところ，(a) 品詞（大分類），(b) 語種，(c) 活用形，(d) 平均文長，(e) 文末の文体，(f) 品詞，(g) 語彙素，以上のデータ表示が可能である。(a)～(e)までは，円グラフとリストを表示し，(f)，(g)はリストのみ表示する。画面例は次の節に示す。

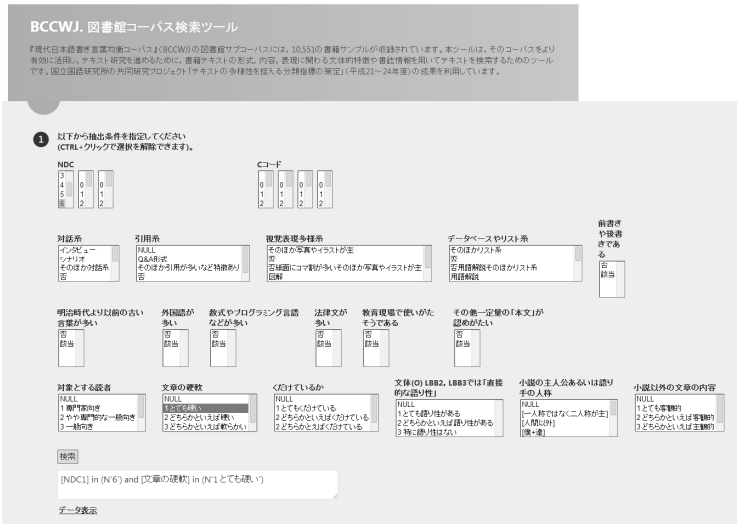


図 1 検索画面

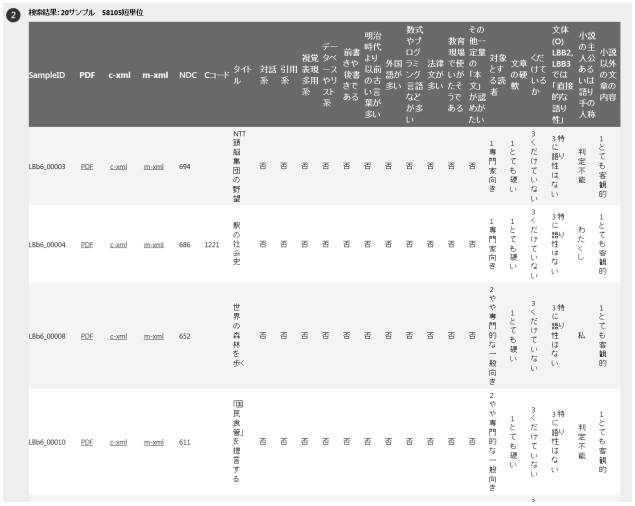


図 2 検索結果の表示画面

３．３　テキスト分析例「とても硬い」と「とても軟らかい」の場合

最後に，「図書館コーパス検索ツール」を用いたテキスト分析例として，「とても硬い」と「とても軟らかい」のテキストを比較する。図は，「データ表示画面」で得られるものである（ただし，印刷の都合上，円グラフは別途作成したものと差し替えている）。

(1) 該当テキストのサンプル数と短単位数

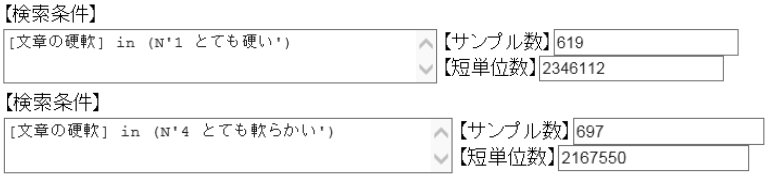
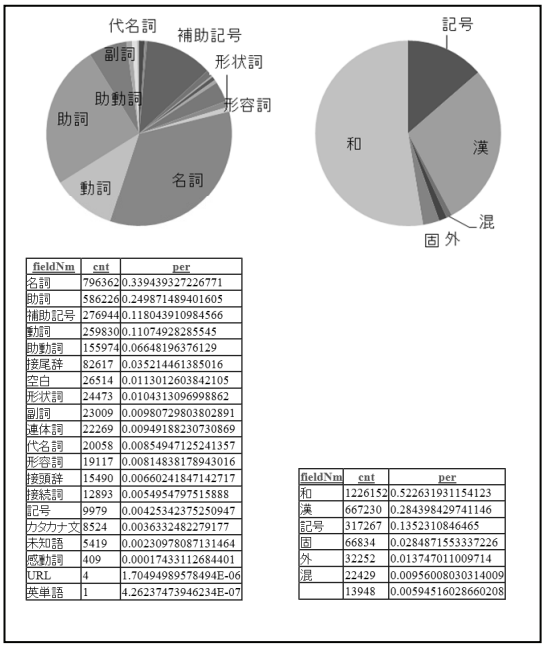


図 3 「とても硬い」「とても軟らかい」のサンプル数と短単位数

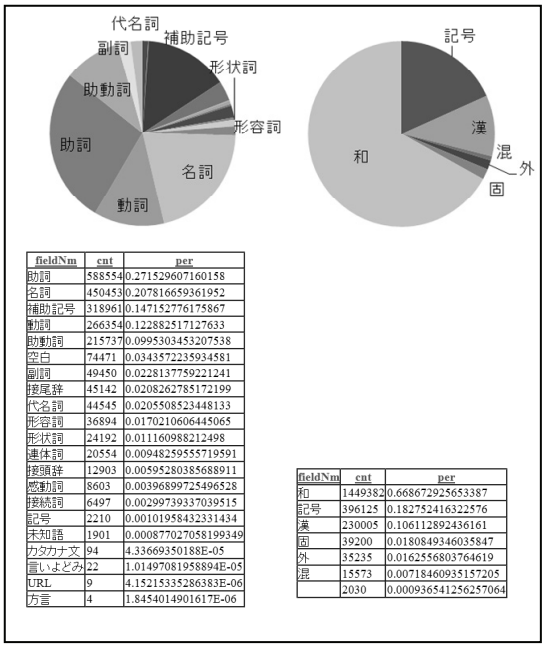
(2) 品詞（大分類），語種



fieldNm	cnt	per
名詞	796362	0.339439327226771
助詞	586226	0.249871489401605
補助記号	276944	0.118043910984566
動詞	259830	0.11074928285545
助動詞	155974	0.06648196376129
接尾辞	82617	0.035214461385016
空白	26514	0.011301260384 2105
形状詞	24473	0.0104313096998862
副詞	23009	0.00980729803802891
連体詞	22269	0.00949188230730869
代名詞	20058	0.00854947125241357
形容詞	19117	0.00814838178943016
接頭辞	15490	0.00660241847142717
接続詞	12893	0.0054954797515888
記号	9979	0.0042534237525094 7
カタカナ文	8524	0.003633248227917 7
未知語	5419	0.00230978087131464
感動詞	409	0.0001743311268440 1
URL	4	1.70494989578494E-06
英単語	1	4.26237473946234E-07

fieldNm	cnt	per
和	1226152	0.522631931154123
漢	667230	0.284398429741146
記号	317267	0.1352310846465
固	66834	0.0284871553337226
外	32252	0.013747011009714
混	22469	0.00956008030314009
	13948	0.00594516028660208

図 4 「とても硬い」品詞（大分類）と語種



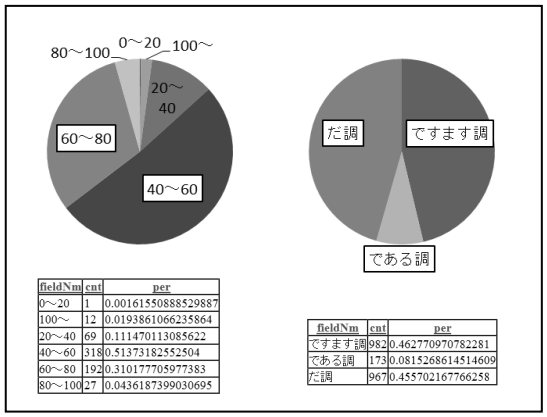
fieldNm	cnt	per
助詞	588554	0.271529607160158
名詞	450453	0.207816659361952
補助記号	318961	0.147152776175867
動詞	266354	0.122882517127633
助動詞	215737	0.0995303453207538
空白	74471	0.0343572235934581
副詞	49450	0.0228137759221241
接尾辞	45142	0.0208262785172199
代名詞	44545	0.0205508523448133
形容詞	36894	0.0170210606445065
形状詞	24192	0.011160988212498
連体詞	20554	0.00948259555719591
接頭辞	12903	0.00595280385688911
感動詞	8603	0.00396899725496528
接続詞	6497	0.00299739337039515
記号	2210	0.00101958432331434
未知語	1901	0.000877027058199349
カタカナ文	94	4.33669350188E-05
言いよどみ	22	1.01497081958894E-05
URL	9	4.15215335286383E-06
方言	4	1.8454014901617E-06

fieldNm	cnt	per
和	1449382	0.668672925653387
記号	396125	0.182752416322576
漢	230005	0.106112892436161
固	39200	0.0180849346035847
外	35235	0.0162556803764619
混	15573	0.00718460935157205
	2030	0.000936541256257064

図 5 「とても軟らかい」品詞（大分類）と語種

図 4，図 5 に示すように，このように，該当テキストに対して，BCCWJ に付与されている形態論情報を利用し，品詞（大分類）や語種の円グラフと一覧とを得ることができる。品詞（大分類）を比べると，「とても硬い」は名詞と助詞の二つが半分以上を占め，一方，「とても軟らかい」はその二つでは半分以下であり，代わりに，補助記号，助動詞の比率が高いことなどがわかる。語種を比べると，「とても硬い」での漢語比率，「とても軟らかい」での和語の比率の高さが顕著にみてとれる。

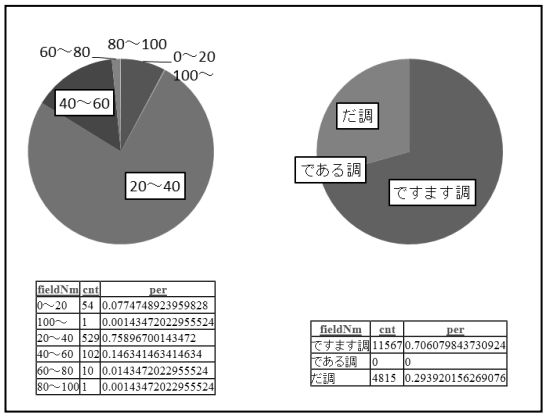
(3) 平均文長，文末の文体



fieldNm	cnt	per
0〜20	1	0.00161550888529887
100〜	12	0.0193861066235864
20〜40	69	0.111470113085622
40〜60	318	0.51373182552504
60〜80	192	0.310177705977383
80〜100	27	0.0436187399030695

fieldNm	cnt	per
ですます調	982	0.462770970782281
である調	173	0.0815268614514609
だ調	967	0.455702167766258

図 6 「とても硬い」平均文長と文末の文体



fieldNm	cnt	per
0〜20	54	0.0774748923959828
100〜	1	0.0014347202295552 4
20〜40	529	0.75896700143472
40〜60	102	0.146341463414634
60〜80	10	0.0143472022955524
80〜100	1	0.0014347202295552 4

fieldNm	cnt	per
ですます調	11567	0.706079843730924
である調	0	0
だ調	4815	0.293920156269076

図 7 「とても軟らかい」平均文長と文末の文体

平均文長は，サンプル全体の文字数を，BCCWJ に付与されている「sentence 開始タグ」の数で割って求めている。その結果，「とても硬い」の平均文長は長く，「とても軟らかい」の平均文長は短いという特徴がはっきりわかる。

BCCWJ では公開ファイル形式のひとつとして XML 形式がある。文体の文末は，会話部分に付与されている XML 形式の「quote タグ」を参照し，その部分を除外し，それ以外の文末に含まれる「です/ます。」「である。」「だ。」の数を合計数で割って求めている。その結果，「とても硬い」は「だ調」と「である調」が多く，「とても軟らかい」は「ですます調」が多い，と，はっきりとした違いがみてとれる。しかしながら，文体の文末の計数機能は現在，改良中である。「です/ます。」「である。」「だ。」の終止形以外の形で終わる場合や，常体「だ・である体」の文末として，動詞，形容詞，名詞などが文末にくる場合も数え，個別の数を出し，さらに詳細に分析できるようにする予定である。

(4) 品詞と語彙素

表 2 品詞一覧（抜粋）

	とても硬い				とても軟らかい		
	fieldNm	cnt	per		fieldNm	cnt	per
1	助詞-格助詞	389312	16.6%	1	助詞-格助詞	306269	14.2%
2	名詞-普通名詞-一般	364536	16.6%	2	名詞-普通名詞-一般	256923	11.9%
3	名詞-普通名詞-サ変可能	172170	7.3%	3	動詞	216737	10.0%
4	動詞	166974	6.6%	4	動詞-一般	142013	6.6%
5	動詞-非自立可能	146186	6.2%	5	補助記号-読点	135973	6.2%
6	補助記号-読点	122366	5.2%	6	動詞-非自立可能	124341	5.7%
7	動詞-一般	113644	4.8%	7	補助記号-句点	103699	4.8%
8	名詞-数詞	91859	3.9%	8	助動詞-接続助詞	99697	4.6%
9	助動詞-係助詞	83696	3.6%	9	助動詞-係助詞	88136	4.1%
10	助動詞-接続助詞	76109	3.3%	10	空白	74471	3.4%
11	補助記号-句点	60661	2.6%	11	名詞-普通名詞-サ変可能	54406	2.5%
12	接尾辞-名詞的-一般	56993	2.4%	12	副詞	49450	2.3%
13	名詞-普通名詞-副詞可能	47680	2.0%	13	名詞-普通名詞-副詞可能	46342	2.1%
14	補助記号-括弧開	37632	1.6%	14	代名詞	44545	2.1%
15	補助記号-括弧閉	37335	1.6%	15	接尾辞-名詞的-一般	36940	1.7%
16	名詞-普通名詞-助数詞可能	36908	1.6%	16	動詞-副詞的	33326	1.6%
17	空白	26614	1.1%	17	補助記号-括弧開	33070	1.6%
18	副詞	23009	1.0%	18	補助記号-括弧閉	33066	1.6%
19	連体詞	22269	0.9%	19	助詞-終助詞	31268	1.4%
20	動詞-副詞的	20963	0.9%	20	動詞-連体詞的	27667	1.3%
21	代名詞	20058	0.9%	21	名詞-数詞	23847	1.1%
46	助詞-終助詞	2965	0.1%	46	接尾辞-形容詞的	920	0.0%

表 2 は，「データ表示画面」から得られる品詞一覧をコピーし，編集しなおした表である。このように，ツールから得たデータを，別途編集し，利用することもできるようになっている。表 2 より，「とても硬い」では，「名詞-普通名詞-サ変可能」と「名詞-数詞」の多いことがわかる。また，「とても軟らかい」では，「副詞」「代名詞」「助詞-終助詞」の多いことがわかる。

「データ表示画面」には，頻度 1 に至るまでの全ての語彙素について，語彙素，語彙素読み，品詞，頻度，割合の一覧が示される。そこから，「名詞-普通名詞-サ変可能」と，「助詞-終助詞」を抽出したものが，次の表 3 である。両者に重なって出現する語には網掛けをして表示する。また，「助詞-終助詞」には頻度を添えて示す。表 3 をみると，「とても硬い」と「とても軟らかい」では重なりが少ないことがわかる。「とても硬い」の方に多く現れる語には，「研究，教育，存在・・・」といった学術的な語が多くあり，「とても軟らかい」の方に多く現れる語には，「料理，結婚，食事・・・」といった生活的な語が多くある。そして，終助詞の一覧をみると，「とても軟らかい」では多くの種類の終助詞が高頻度で出現していることがわかる。

表 3　語彙素一覧より得た「名詞-普通名詞-サ変可能」と「助詞-終助詞」の語

「とても硬い」名詞-普通名詞-サ変可能（上位 20 語）：
物，関係，研究，意味，教育，存在，生活，労働，活動，生産，利用，戦争，運動，規定，著作，裁判，決定，影響，支配，計画

「とても軟らかい」名詞-普通名詞-サ変可能（上位 20 語）：
物，話，仕事，心，一緒，電話，生活，関係，意味，料理，結婚，勉強，食事，説明，涙，存在，びっくり，意識，約束，旅

「とても硬い」助詞-終助詞（頻度 50 以上）：
か 2165，よ 213，な 150，ね 103，わ 63，の 58，ぞ 53

「とても軟らかい」助詞-終助詞（頻度 50 以上）：
か 7391，よ 7026，ね 5442，な 3730，の 2109，わ 1650，さ 1018，ぞ 608，い 514，ぜ 356，もの 356，かしら 323，や 120，ねん 119，け 117，じゃん 96，で 71

４．おわりに

Web アプリケーション「図書館コーパス検索ツール」の構築について報告した。本ツールを用いることで，特定の文体的特徴をもつテキストを検索し，形態論情報等を簡単に一覧することができるようになった。その一例として，「とても硬い」「とても軟らかい」と判定されたテキストが，どのようなテキストの集合となるのか，ツールから得られる，品詞，語種，平均文長，文末文体，品詞一覧，語彙素一覧を用い，比較分析を行った。そして，それらツールから得られる情報によって，両者の文体差を支える言語的特徴の違いをはっきり確認できることを示した。本ツールを活用することで，さまざまに付与した文体的特徴を支える言語的特徴が何であるかの分析を進めていくことができる。

本ツールは，インターネットが利用できる環境と標準的なブラウザがあれば，特別なソフトをインストールすることなく利用することができる。現在は，研究所内のネットワークからのみアクセス可能であるが，広く利用してもらえる方法を検討していきたい。

謝　辞

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」，「コーパスアノテーションの基礎研究」による成果を含みます。また，文部科学省科学研究費補助金基盤研究(C)「コーパス分析に基づく辞書の位相情報の精緻化」（課題番号:23520572）の助成を得たものです。

文　献

Lee, Y. D. (2001) Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle, *Language Learning & Technology*, 5:3, pp.37-72.

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，立花幸子，保田祥，丸山岳彦，奥村学，佐藤理史，徳永健伸，大塚裕子，佐渡島紗織(2012a)「テキストの硬さと軟らかさの考察 ー『現代日本語書き言葉均衡コーパス』の収録書籍を対象にー」『第 1 回コーパス日本語学ワークショップ』予稿集,pp.131-138.

柏野和佳子，立花幸子，保田祥，飯田龍，丸山岳彦，奥村学，佐藤理史，徳永健伸，大塚裕子，佐渡島紗織，椿本弥生，沼田寛(2012b)「書籍テキストへの文体情報付与の試み」『第 2 回コーパス日本語学ワークショップ』予稿集,pp.155-164.

附録

文体的特徴を表す分類指標の付与サンプル例

(1)専門度：1 専門家向き（LBi4_00021『がんと遺伝子』）

> Ｅ２Ｆ以外のＲＢ結合タンパク質としては，転写因子ＲＡＸ，Ｔ細胞が活性化するときに誘導されるＩＬ－２，ＧＭ－ＣＳＦ，ＨＩＶ－２などの転写を活性化する転写因子Ｅｌ Ｆ－１や先に述べた細胞周期を制御するサイクリンＤなどがある．おもしろいことに，Ｅｌ Ｆ－１やサイクリンＤのＲＢ結合ドメインにはｌａｒｇｅＴ抗原やＥ１Ａタンパク質と同じようにＬＸＣＸＥというアミノ酸配列が存在する．また，ＲＢタンパク質は骨格筋分化を支配する重要な遺伝子群ＭｙｏＤファミリー（ＭｙｏＤ，ｍｙｏｇｅｎｉｎ，ＭＲＦ４，ｍｙｆ－５）の産物とも複合体を形成し筋分化にも関与しているらしい．

(2)専門度：4 中高生向き（LBf9_00090『超魔炎獄変』）

> 白く薄い空気のヴェールが、漂うように揺らめいている。
> シャ…ァァン、シャラ…ァァン…。
> 闇を抜け、霧の中を渡る金属の響き。それは魔を覇する浄化の音。
> 響きに道を開けるかのようにすぅ…っと霧が左右に分かれた。
> それは。
> 霧の中にたたずむそれは。
> 闇。
> …いや。闇ではない。

(3)客観度：1 とても客観的（LBo3_00158『行政法要論』）

> たとえば委員会の開催が「急施を要する場合」にあたるかどうかとか、公衆浴場の施設が「公衆衛生上不適切」かどうかは通常人の経験則によって十分判断できる事柄であるから、羈束裁量であって裁判所の終局的な判断に服すべきものとする。これに対し、外国人の在留期間の更新を適当と認めるに足る相当の理由があるかどうかは、出入国管理行政の責任者である法務大臣の政治的判断に委ねらるべきであり、また、原子炉の安全性の認定は高度の科学的専門技術的知見に基づく総合的判断であるから、行政庁の便宜裁量事項であり、その当否は裁判所の審理・判断にはなじまないとする（最判昭和五三年一〇月四日民集三二巻七号一二二三頁、同平成四年一〇月二九日民集四六巻七号一一七四頁）。

(4)客観度：4 とても主観的（LBo3_00132『教師をめざす若者たち』

どんなに上手な言葉を使っても、思っていないことを発すれば、子供に伝わらない。どんなに下手な言葉でも、心から伝えたいという愛情があれば、伝わるものであるということを信じることができました。この実感は日本でも通じる「教育の原則」であると思いました。

二日目、子供たちと綿花摘みを一緒にしました。敦煌の子供たちの手は「仕事をしている手」でした。

(5)硬度：1 とても硬い（LBi3_00033『現代法社会学入門』）

取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。

(6)硬度：4 とても軟らかい（LBa4_00010『恐竜の世界をたずねて』）

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへしたへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三畳紀のはじめにいた、「テコドント」（図８６）という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

(7) くだけ度：1 とてもくだけている（LBf9_00067『男はオイ！女はハイ…』）

最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをば何となく眺めているうちに、どうしても欲しくなった商品があった。

よし、こいつひとついってやれとばかりすぐ電話にとびついた。

「ハイ、こちら―です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。

「商品番号をおっしゃって下さい」

といわれて答える。

さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

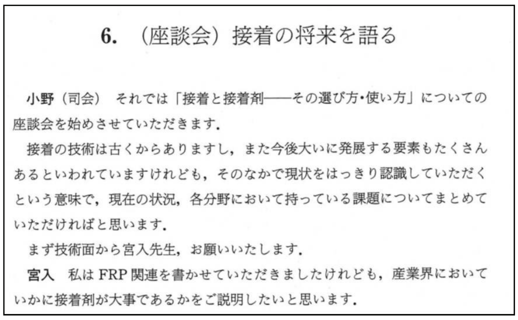
(8) 語りかけ性度：1 とても語りかけ性がある（LBt1_00013『5 分間集中力トレーニング』）

精神的に疲れていると、「ああなったら、どうしよう」「こうなったら、どうしよう」と常に不安だらけになります。

動物病院にいらっしゃる飼い主さんには、過剰な不安を抱えている人や心配性の人がとても多い。実はそれがペットの病気をさらに悪化させることになっていますが、そういう認識をお持ちの飼い主さんは、あまりいません。詳しい説明は避けますが、不安や心配性を放っておくと、動物の具合が悪くなり、当然それが自分にも返ってきます。

それでは、どうすればいいのでしょうか。

文体判断が単純にいかない特徴をもつものの分類指標の付与サンプル例



6. （座談会）接着の将来を語る

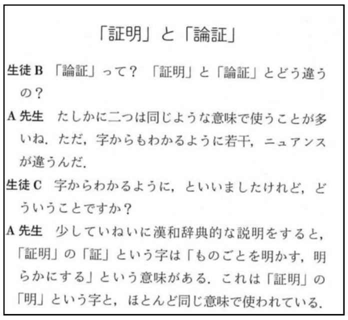
小野（司会）　それでは「接着と接着剤――その選び方・使い方」についての座談会を始めさせていただきます。

接着の技術は古くからありますし，また今後大いに発展する要素もたくさんあるといわれていますけれども，そのなかで現状をはっきり認識していただくという意味で，現在の状況，各分野において持っている課題についてまとめていただければと思います。

まず技術面から宮入先生，お願いいたします。

宮入　私はFRP関連を書かせていただきましたけれども，産業界においていかに接着剤が大事であるかをご説明したいと思います。

図 1 座談：LBd5_00012『接着と接着剤』



「証明」と「論証」

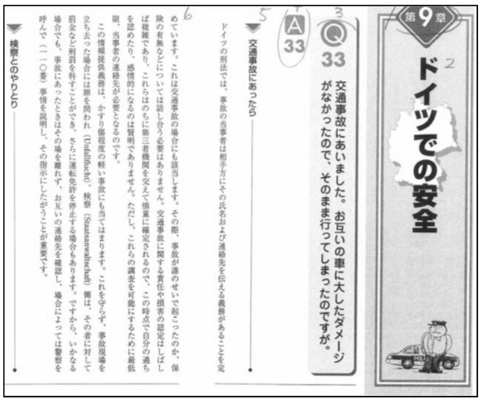
生徒B「論証」って？「証明」と「論証」とどう違うの？

A先生　たしかに二つは同じような意味で使うことが多いね．ただ，字からもわかるように若干，ニュアンスが違うんだ．

生徒C　字からわかるように，といいましたけれど，どういうことですか？

A先生　少していねいに漢和辞典的な説明をすると，「証明」の「証」という字は「ものごとを明かす，明らかにする」という意味がある．これは「証明」の「明」という字と，ほとんど同じ意味で使われている．

図 2 対話：LBpn_00038『なぜ数学を学ぶのか』



第9章 ドイツでの安全

Q 33 交通事故にあいました。お互いの車に大したダメージがなかったので、そのまま行ってしまったのですが。

A 33

交通事故にあったら

警察とのやりとり

図 3 Q&A：LBr3_00037『ドイツ暮らしの法律Q&A』

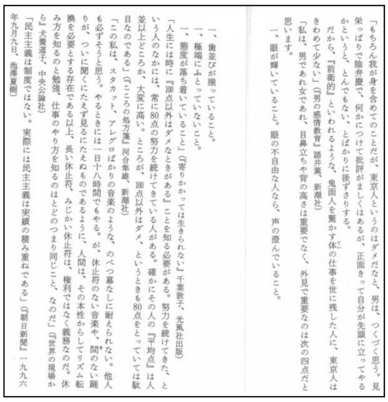


図 4　引用：LBl8_00017『試行錯誤の文章教室』

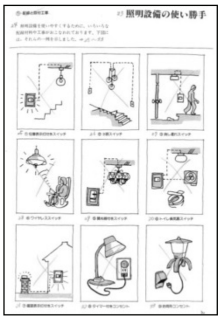


図 5 イラスト：LBb5_00011『絵ときインテリアライティングの技法早わかり』

図 6　コマ割り：LBsn_00022『東京で遊ぼ』

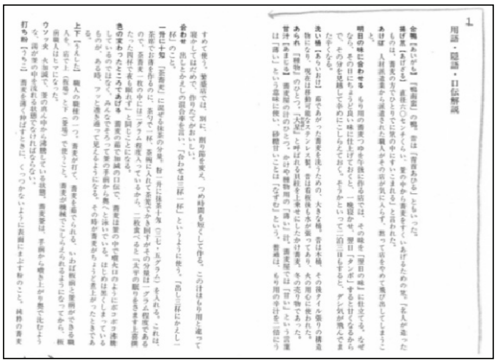


図 7　辞書：LBp6_00009『蕎麦屋のしきたり』

図 8　カタログ：LBj6_00025『熱帯魚・水草カタログ』