法律文中における単語出現頻度の変化
――法令テキストマイニングの一例――

小川　泰弘　　中村　誠　　外山　勝彦

1. はじめに

近年、コンピュータとコンピュータ・ネットワークの発展に伴い、それらを法学研究や法実務に活用することが提案され[1][2]、実践されてきている。特に、様々なデータベースやインターネット上から必要なデータを収集する検索技術の効果は顕著である。特定のキーワードを元に検索すれば、関連する法令や判決を容易に入手することが可能であり、これまで人手で時間を掛けて行ってきた情報収集を大幅に効率化することができる。

検索技術の問題点は、キーワードなどの手掛かりがはっきり分かっていなければ、求める情報を得ることができないということである。それを補うものとして、近年注目を浴びているのがデータマイニングである。データマイニングとは、収集した多くのデータを分析することにより、データの中に隠れている情報を抽出する手法である。つまり、データマイニングは、手掛かりがはっきりしないときに、それを見つける手助けをしてくれる技術といえる。

本稿では、そうしたデータマイニングの一例として、法律文中における単語の出現頻度の推移から、どのような情報が得られるか考察してみた。今回の法律文のようなテキストデータを対象とするデータマイニングは、特にテキストマイニング[3][4][5]とも呼ばれる。なお、筆者は法学者ではないため、適切な考察を提供できる訳ではないが、これまで見えてこなかった情報を検討する契機となれば幸いである。

2. データマイニング

データマイニングとは、「マイニング」という言葉が示すように、埋もれた情報を掘り出す技術であり、見えなかった宝を見えるようにする点に意義がある。例えば、現在オンラインショッピングにおいては、ユーザが他にどのような商品を購入しているか、購入するまでどのページを経由したか、どの時刻に購入したかなど、考えられるあらゆる情報を収集しておき、それらを分析することにより、利益向上に繋げる技術として利用されている。特に最近では、技術の進歩によって巨大なデータ（ビッグデータ）を扱うことができるようになったことから、それに対するデータマイニングが話題となっている。

データマイニングには、回帰分析やクラスタリングなど、より高度な解析手法も存在するが、今回は単純に法律文中における各単語の出現頻度に着目する。

データマイニングにおいて重要なことは、得られたデータがそのまま答えになる訳ではないということである。今回の実験では、法律文中に出現する単語の出現頻度を求めているが、出現頻度の高い単語が単純に重要語となる訳ではない。実際には、得られたデータを更に人間が分析し、答えを得る必要がある。その意味で、データマイニングは答えに繋がる可能性のある手掛かりを提供してくれる技術であるといえる。

3. 実験

今回の実験では、法律文に対するテキストマイニングの一例として、法律文中の名詞の出現頻度から、興味ある情報を導き出すことを試みる。それにより、法令情報に対するデータマイニングの適用可能性を示すことが目的である。

3.1 実験対象

今回は日本国憲法施行後から、現在までに制定された法律を対象とし

た。法律文は、国立印刷局が運営する「官報情報検索サービス[1)]」で提供されているテキスト化データを使用した。ただし、「官報情報検索サービス」のデータのうち古い時期のものは、官報をスキャンした画像に対して、OCR で処理して得られたデータが含まれており、OCR の際の認識誤りが含まれている。今回このような誤りを残したまま使用したが、将来的にはデータの修正が必要となる。

使用した法律は、昭和 22 年法律第 89 号から平成 24 年法律第 102 号までであり、その総数は 9,915 本である。年ごとの制定本数を表 1 に示す。なお、ここで注意すべき点は、今回の対象となった法律には、一部改正法・整理法・整備法・廃止法が含まれているという点である。それらにはいわゆる改め文が含まれるが、改め文には以下の例のように、同じ単語が複数回出現する場合がある。

次に掲げる法律の規定中「廃疾年金」を「障害年金」に改める。
（障害に関する用語の整理に関する法律（昭和 57 年法律第 66 号第 80 条））

この場合、「年金」が 2 回出現していることになり、注意が必要である。なお、単語をどのように定義するかで出現回数が異なってくる点にも注意が必要である。「廃疾年金」および「障害年金」をそれぞれ 1 単語と数えると、上記の問題は生じない。今回は、後述する形態素解析システム茶筌が分割する単位を 1 単語とした。茶筌は比較的短い単位で単語を区切るシステムである。

法律の題名に対して正規表現[2)]を利用したマッチングを用いることにより、一部改正法・整理法・整備法・廃止法に該当するものを「一部改正法等」として分類した。正規表現とは、マッチングを行う際に、文字

1）　https://search.npb.go.jp/
2）　プログラミング言語 Ruby を用いた具体的な正規表現は以下のとおりである.
一部改正法:/一部を改正する(等の)?法律¥z/
整理法:/(の|伴う|及び)整理(等|(、|及び)合理化(等|並びに臨時特例等)?)?に関する法律¥z/
整備法:/法(律|令)の整備(等)?に関する法律¥z/
廃止法:/廃止(等)?(に関)?する(等の)?法律¥z|廃止法¥z/

表 1　制定法律数

年	S22†	S23	S24	S25	S26	S27	S28	S29	S30	S31	S32	S33	S34	S35
全法律	159	282	286	303	318	358	292	229	196	180	187	193	204	174
一部改正法等	64	125	160	176	216	198	159	151	143	123	123	139	128	128

年	S36	S37	S38	S39	S40	S41	S42	S43	S44	S45	S46	S47	S48	S49
全法律	238	164	183	185	157	151	149	111	97	145	132	132	123	118
一部改正法等	163	128	134	145	114	116	96	87	72	96	94	99	87	96

年	S50	S51	S52	S53	S54	S55	S56	S57	S58	S59	S60	S61	S62	S63
全法律	96	88	97	107	76	111	102	94	83	88	109	109	115	112
一部改正法等	79	73	74	81	64	92	83	80	67	69	95	75	80	88

年	H1‡	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14
全法律	97	85	112	110	98	119	137	120	132	152	226	149	158	192
一部改正法等	76	68	84	84	79	93	99	96	100	108	107	104	124	115

年	H15	H16	H17	H18	H19	H20	H21	H22	H23	H24	合計
全法律	147	167	124	123	136	98	100	72	126	102	9,915
一部改正法等	105	120	91	91	96	73	72	56	77	76	6,854

† S22 は憲法施行後の昭和 22 年法律第 89 号以降のみ
‡ H1 は昭和 64 年に公布された法律 1 本を含む

の並び方のパターンを表現する方法である。その結果、6,854 本が該当した（表 1）。年ごとの一部改正法等の割合を見ると、当初の昭和 22 年と昭和 23 年は一部改正法等が半分以下であるが、それ以降は一部改正法等の方が多く制定されている。平成 11 年のみ例外であるが、これは平成 12 年の中央省庁再編に向けて各種の関連法律が新規に制定されたためと思われる。

なお、今回使用した法律 9,915 本に含まれる文字数は約 9,300 万文字であり、ファイルサイズにして 252 メガバイトである。一般にビッグデータと呼ばれるデータはテラバイト（=100 万メガバイト）を超えるような巨大なデータであり、それと比較すると法律テキストデータのサイズは大きくない。

3.2　実験手順

まず、実験対象の法律文に対する解析方法と、着目するキーワードの選定方法について述べる。

3.2.1　法律文の解析

今回の実験では、まず法律文中の旧字体を新字体に変換し、「登録があつた」の「つ」などの促音の大書きも小書きに修正した。その後、形態素解析により各法律文を処理した。形態素解析とは文を単語に分割し、それぞれに品詞を付与する作業である。今回は、形態素解析システム茶筌（ChaSen）[3] を使用した。ここで、形態素解析システムの精度は、システムが使用する単語辞書に依存する。茶筌に組込まれている単語辞書をそのまま使用した場合、一部の法令用語が正しく解析できなかった。そこで、700 語余りの単語を茶筌の単語辞書に追加して解析誤りを減らした。

その後、法律文から名詞だけを抽出した。なお、茶筌の解析結果では数詞や助数詞も名詞の一種となるが、情報抽出という目的からは不要と判断し、それらは除いた。

次に、年ごとに各名詞の出現頻度を数え上げ、それがどのように変化したかを調べた。1 年ごとでは変動が大きく、変化の傾向を捉えにくかったため、5 年ごとにまとめて一つの期間とした。ただし、昭和 22 年から平成 24 年までの 66 年間は 5 等分できない。また、最初の昭和 22 年は約半年分しかないことから、昭和 22 年分はそれ以後と一緒にして、昭和 22 年から昭和 27 年までを一つの期間とした。なお、単純な出現頻度では、表 1 に示したように、制定法律数ひいては出現単語数が年ごとに異なるため比較しづらい。そこで、期間ごとの出現頻度の順位を求め、その変化を調べた。

3.2.2　キーワードの選定

次に、着目するキーワードをどのように選定するかを考える。もちろ

3）　http://chasen-legacy.sourceforge.jp/

表２　出現順位の変化が大きな単語

吉田、茂、池田、勇人、佐藤、栄作、三木、武夫、福田、赳夫、鈴木、善幸、鳩山、一郎、海部、俊樹、田中、角栄、宮沢、喜一、大平、正芳、中曽根、康弘、岸、信介、橋本、法務、大蔵、通商、大蔵省、運輸省、通商産業省、総務、厚生省、農林省、総理府、建設省、農林水産省、12 月、3 月、6 月、9 月、元年、平成、勅令、総裁、麻薬、海面、線、官吏、少年、配給、計量、臣、附加、入場、競走、政党、差押、差押え、口座、独立、酸、競売、軍人、持株、入所、廃疾、公、借地、適格、上場、株券、地震、現物、看護、聴取、織物、従、予約、有限、事後、完全、水害、糸、絹、綿、毛皮、紡織、アセテート、ガラス、合成、切替、活用、検討、訪問、食事、小笠原、琉球、液化、連結、震災、合金、栄、人造、吸収、防災、鳥、形成、育児、親、強化、被災、排出、次号、肉、流動、混合、近代、指数、誘導体、緑地、個別、先物、参照、通勤、電子、取消し、居宅、比例、基盤、子会社、換価、移行、老齢、集積、公害、再生、従量税、拠出、支援、食、ただし書、軽油、原子炉、原子力、センター、最短、農用地、エネルギー、電磁、申立て、貸付け、訴え、ばい煙、申込み、編み、活性、読替え、老人、不況、浄化槽、高齢、買付け、革新、連携、大震災、子ども、東日本

ん、あらかじめ着目するキーワードがあれば、その出現順位の推移を調べることができるが、データマイニングの利点は、着目するキーワードが分からないときに、その候補を与えてくれる点にある。

そこで今回は、出現順位の変化が大きな単語に着目した。制定された法律に出現する単語は、新規制定であれ、一部改正であれ、それに関係する法律を整備する必要があったと考えられる。特に、ある特定の期間だけ頻出した単語には、その期間に関係法律を整備する社会的な要請があったと考え、それらを自動的に抽出してみた。具体的には、ある 5 年間は出現順位が 500 位以内となる高頻度語であったが、別の 5 年間は出現順位が 5,000 位以下の低頻度であった単語を自動的に抽出してみた。その結果を表 2 に示す。

表 2 を見ると「吉田」「茂」「池田」など人名が多く出現するが、これらは内閣総理大臣の署名として出現する氏名であり、ある期間にのみ頻出するのは当然である。また、「大蔵」は、平成 12 年の中央省庁再編により「大蔵省」が「財務省」へと名称が変わったため、それ以降は法律に出現しない。このように、人名や省庁名に関しては、出現順位が変化した原因は明らかである。

一方で、順位の推移が特徴的であり、その原因に興味を引かれる単語をいくつか発見した。その結果を以下に示す。

3.3　実験結果と考察

「水害」「公害」「麻薬」「廃疾」「障害」「子ども」「児童」の出現順位を表 3 に示す。図 1 は表 3 をグラフにしたものである。なお、表 3 において空欄の箇所は、その単語がその期間中に制定された法律に 1 回も出現しなかったため、順位が付けられなかったことを示す。

各単語の出現順位の推移について考察してみる。まず、「水害」は昭和 28 年～昭和 32 年および昭和 33 年～昭和 42 年に出現順位が高いことから、当時、水害に対処するための法整備がなされたことが推測できる。

また、「公害」は、昭和 43 年～昭和 47 年および昭和 48 年～昭和 52 年の間に出現順位が高くなっている。公害に関する裁判を調べると、新潟水俣病と四日市ぜんそくに関する提訴が昭和 42 年、イタイイタイ病が昭和 43 年、水俣病が昭和 44 年であり、この時期に公害に関する法律が整備されたことがうかがえる。

「麻薬」は戦後しばらく出現順位が高いが、その後、低くなっている。しかし昭和 63 年～平成 4 年の間に出現順位が高くなっている。調べてみると、日本は平成 4 年に「麻薬及び向精神薬の不正取引の防止に関する国際連合条約（平成 4 年条約第 6 号）」を批准しており、その批准のために「麻薬取締法等の一部を改正する法律（平成 2 年法律第 33 号）」「麻薬及び向精神薬取締法等の一部を改正する法律（平成 3 年法律第 93 号）」「国際的な協力の下に規制薬物に係る不正行為を助長する行為等の防止を図るための麻薬及び向精神薬取締法等の特例等に関する法律（平成 3 年法律第 94 号）」といった法律を制定していたことが分かった。

単語によっては、あるときから法律で全く使用されなくなる場合もある。「廃疾」は昭和 57 年以前の多くの法律に出現しているが、それ以降、全く出現していない。これは、昭和 56 年を国際障害者年と国連が定めたことを契機とし、障害に関する法令上の不適当用語の改正について関係者の要望が高まったことに由来する。この単語は「本来、重度心身障害（者）を意味するが、程度の軽重にかかわらず包括的に表現される等法令ごとに定義が一様でない。一般的な「障害の存在」を表現する意に使われているのがほとんどであり、実体的な意義はなくなっている。」[6] といった理由から廃止されたことが分かった。その際には、まず「障害

表３　「水害」「公害」「麻薬」「廃疾」「障害」「子ども」「児童」の年代別出現順位

年代	S22 -S27	S28 -S32	S33 -S37	S38 -S42	S43 -S47	S48 -S52	S53 -S57	S58 -S62	S63 -H04	H05 -H09	H10 -H14	H15 -H19	H20 -H24
水害	2884	440	711	2142	5801		1250	6509	3237	2658	4158	2669	3639
公害	l3617	8287		844	239	366	859	742	857	988	1237	1654	1509
麻薬	544	225	3634	676	2304	5377	2524	1416	332	2387	1744	2578	4200
廃疾	1957	482	196	278	405	174	115						
障害	309	279	158	106	172	88	75	42	173	130	178	108	66
子ども											2261	2123	177
児童	428	505	443	366	348	436	523	387	494	652	409	435	134

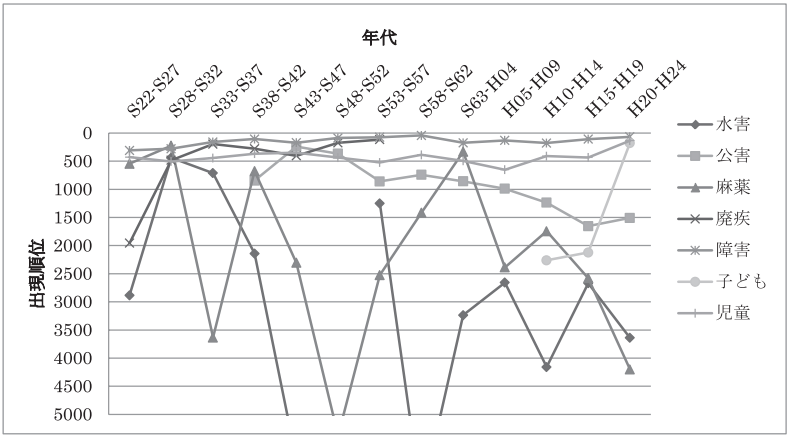

図１　「水害」「公害」「麻薬」「廃疾」「障害」「子ども」「児童」の出現順位の推移

に関する用語の整理のための医師法等の一部を改正する法律」（昭和 56 年法律第 51 号）によって「おし」「つんぼ」「盲」という三つの用語について関係法律の改正が行われた。これに続いて、「不具」「廃疾」「白痴者」という用語を廃止するために、「障害に関する用語の整理に関する法律」（昭和 57 年法律第 66 号）が制定された[7]。このように、単語の出現頻度の変化を手掛かりとして法律の変遷を追うことが可能となる。

一方、従来は使用されてこなかった単語が、ある時点から使用されるようになった場合もある。特に「子ども」は顕著であり、近年は良く使

用される単語であるが、法律に最初に出現したのは「特定非営利活動促進法」（平成 10 年法律第 7 号）であり、それ以前は使用されていなかった[4]。この法律は NPO 法とも呼ばれ、市民が立法過程に大きく関わってできた法律であり、法律の文言や中身についても市民の意見が反映されるよう働きかけたようである[8]。この法律を契機として「子ども」という単語が利用されてきたようであるが、類似語の「児童」という単語は常に高い出現順位であり、なぜ「子ども」が法律で使用されるようになったかは興味深い。

また、複数の単語が同じ傾向を示したものもあった。その例として「糸」「絹」「綿」「毛皮」「紡織」「アセテート」「ガラス」の年代別出現順位を表 4 に、それをグラフにしたものを図 2 に示す。これを見ると、昭和 58 年～昭和 62 年にいずれの単語も出現順位が高くなっている。繊維に関わる単語が多いが、「ガラス」も同じ傾向を示した。調べてみると、例えば「ガラス」が出現した法律は 51 本、「毛皮」が出現した法律は 40 本であったが、そのうち 25 本において両者が同時に出現しており、その多くは関税に関する法律の一部改正であった。つまり、これらの単語が法律に出現するときは、関税に関する場合が多いということである。特に昭和 58 年～昭和 62 年の間に出現順位が高い理由は、「関税定率法及び関税暫定措置法の一部を改正する法律」（昭和 58 年法律第 12 号）、「関税暫定措置法の一部を改正する法律」（昭和 60 年法律第 10 号）、「商品の名称及び分類についての統一システムに関する国際条約の実施のための関係法律の整備に関する法律」（昭和 62 年法律第 80 号）などにおいて、関税に関する法律の一部改正が行われたためである。

一方で、直接の関連は不明であるが、良く似た推移を示した単語もあった。その一例として、表 5 に「勅令」と「官吏」の年代別出現順位を、図 3 にそれをグラフ化したものを示す。調べてみると、「勅令」が出現した法律は 429 本、「官吏」の場合は 316 本であり、両者が共に出現した法律は 79 本であった。「勅令」が出現する場合は、例えば「旧中等学校令（昭和 18 年勅令第 36 号）」のような参照表現であった。一方、「官吏」の出現を調べると、それを「職員」に改める一部改正が多かったが、

4）　ただし，例外的に「繊維製品品質表示法」（昭和 30 年法律第 166 号）に 1 回だけ「子供」が出現している．

表４　「糸」「絹」「綿」「毛皮」「紡織」「アセテート」「ガラス」の年代別出現順位

年代	S22 -S27	S28 -S32	S33 -S37	S38 -S42	S43 -S47	S48 -S52	S53 -S57	S58 -S62	S63 -H04	H05 -H09	H10 -H14	H15 -H19	H20 -H24
糸	3629	1421	964	784	1437	2710	686	317	2995	664	4174	3696	7192
絹	2888	2875	2254	1799	2241	5179	1184	361		690	5116	3072	5384
綿	3098	3745	3484	2955	1518	2859	1438	424	6005	969	4178	4364	4635
毛皮	3287	4090	2017	1673	2090	3457	1861	468	5015	998	5078	9774	5324
紡織				1093	1666	5170	912	279	2467	608	4844	3637	4962
ｱｾﾃｰﾄ			2185	2444	5382	2430	1199	433	6292	566	4502	2742	6357
ガラス	2828	4554	1317	1271	1329	2068	716	466	1655	1147	4687	4169	

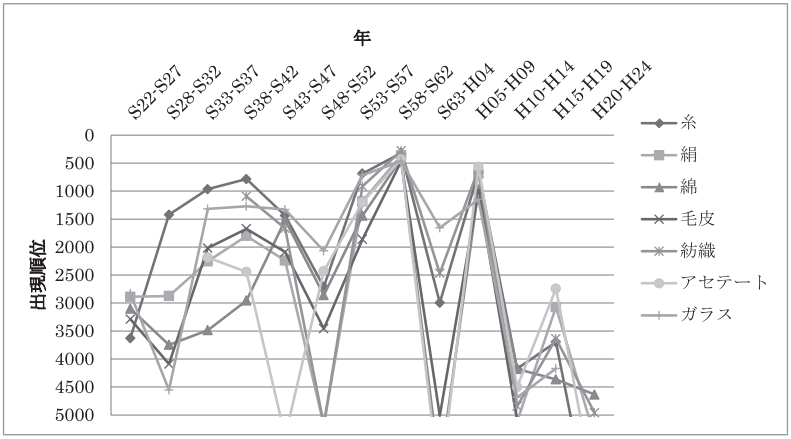

図２　「糸」「絹」「綿」「毛皮」「紡織」「アセテート」「ガラス」の出現順位の推移

「収税官吏」「出納官吏」のように現在も使用されている例もあった。「勅令」と「官吏」に直接の関連は無いが、ともに古い用語であり、それらの出現順位の傾向が似ているのは興味深い。

4. 関連研究

本稿と同じように法令情報における単語の出現頻度に着目した研究としては、Katz らの研究[9]がある。ここでは、判例における単語の出現頻度から、法令言語の「進化」が捉えられるのではないかと提案されて

表５ 「勅令」「官吏」の年代別出現順位

年代	S22 -S27	S28 -S32	S33 -S37	S38 -S42	S43 -S47	S48 -S52	S53 -S57	S58 -S62	S63 -H04	H05 -H09	H10 -H14	H15 -H19	H20 -H24
勅令	368	744	2093	1544	1574	3129	4042	4119	3362	3777	2984	2582	5169
官吏	403	1091	1792	1998	2036	2967	7804	3989	3629	3788	2992	2748	3604

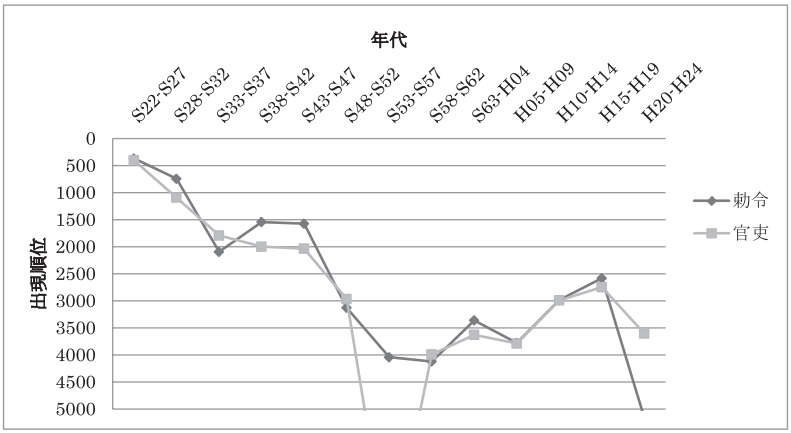

図３ 「勅令」「官吏」の出現順位の推移

おり、その手法を実装した Legal Language Explorer[5] というサイトが運営されている。このサイトでは、入力したキーワードのアメリカ連邦最高裁判例における出現頻度の推移を見ることができる。また、法令ではなく、出版された本の中での出現頻度の推移は Google books Ngram Viewer[6] で調べることが可能である。

また、本稿で示した法律文中での単語の出現頻度を、新聞における出現頻度と比較をすることにより、社会で問題になった事柄がどのように法律に反映されていくかの様子を捉えることが可能になるかもしれない。

法令文に対するテキストマイニングはこれから盛んになっていくと考えられる。特に法令関連のデータベースで知られる LexiNexis は、関連

5） http://legallanguageexplorer.com/
6） http://books.google.com/ngrams/

会社がビッグデータを扱うオープンソースの技術 HPCC Systems[7)] を提供しており、法令を含む各種分野でのデータマイニングに乗り出している。

5. おわりに

本稿では、法令文に対するテキストマイニングの一例として、法律に出現する単語の頻度から、いくつかの興味深い視点を取り上げてみた。このようなテキストマイニングでは、法律文を組織的・網羅的に調べることにより、客観的なデータを得ることができる。得られたデータをどのように解釈するかは、研究者の力量を問われるところである。隠れた情報を見えるようにするデータマイニングは、法情報学の分野において新たな視点と展開をもたらすものと考えられる。そのような点から、多くの法情報の研究者が気軽にデータマイニングを利用できる環境整備ができればよいと考えている。

謝辞

本稿の執筆にあたりまして、名古屋大学大学院法学研究科附属法情報研究センター特任助教の島亜紀先生に、「子ども」の使用例に関する有益なご助言をいただきました。また同大学院法学研究科博士課程後期課程の佐野智也氏に、関連研究に関する有益なご助言をいただきました。ここに感謝の意を表します。

参考文献

［1］加賀山茂，松浦好治（編）: 法情報学―ネットワーク時代の法学入門 第 2 版補訂版，有斐閣（2006）.

［2］指宿信 : 法情報学の世界，第一法規株式会社（2010）.

［3］那須川哲哉，諸橋正幸，長野徹 : テキストマイニング―膨大な文書データの自動分析による知識発見―，情報処理，Vol. 40，No. 4，pp.358-364（1999）.

［4］金明哲 : テキストデータの統計科学入門，岩波書店（2009）.

［5］ローネン・フェルドマン，ジェイムズ・サンガー，（辻井潤一監訳）: テキストマ

7） http://hpccsystems.com/

イニングハンドブック，東京電機大学出版局（2010）.

［6］東京都中野区 : 不快･差別用語の考え方（昭和 57 年 5 月 12 日 57 中総総第 115 号）各課（所）長あて総務課長通知，http://www.city.tokyo-nakano.lg.jp/reiki/reiki_honbun/aq60006851.html（1982）.

［7］国立国会図書館 : 障害に関する用語の整理に関する法律案会議録, 日本法令索引, http://hourei.ndl.go.jp/SearchSys/viewShingi.do?i=109601075（1982）.

［8］秋山訓子：市民の手で法律を作る―日弁連と衆院法制局の試み，「世界」，岩波書店，Vol.4，pp.281-287（2012）.

［9］Katz, D.M., Bommarito II, M.J., Seaman, J., Candeub, A., Agichtein, E.: Legal N-Grams? A Simple Approach to Track the ‘Evolution’ of Legal Language, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pp.167-168（2011）.