

BSH/NDLSH と NDC における階層構造はどの程度一致するのか 件名と分類の統合的活用法に向けて

谷口 祥一 (慶應義塾大学文学部)
taniguchi@z2.keio.jp

[抄録] BSH 第4版と NDLSH (2013年8月時点データ) の個々の件名に付与されている NDC 新訂9版の分類記号を手がかりに、BSH と NDC の組み合わせ、NDLSH と NDC の組み合わせにおいて、どの程度上位下位関係の階層構造が一致するのか定量的に調査した。BSH と NDLSH をベースにした照合、NDC をベースにした照合をそれぞれ実施し、階層構造が一致しない箇所をプログラムで数えた。NDC の縮約・不均衡項目の正規化、関連項目の活用、上位分類項目からの件名の限定的継承などを組み入れた場合についても実験した。

1. はじめに

分類法は体系性と網羅性を重視し、関連主題項目が並置され集中される特質をもつ。一方、件名法(統制語彙による主題表現法)は、自然言語をベースにして、その直接性や柔軟さに力点を置くため、上位語を複数有する構造もみられる。こうした異なる特性を有する主題表現方式およびその表現結果であるが、統合的に活用する方策を検討する意義はあると考える。

そこで、統合的な活用法の検討の一環として、分類法、件名法の身近な例である BSH (基本件名標目表) 第4版、NDLSH (国立国会図書館件名標目表)、NDC (日本十進分類法) 新訂9版を取り上げ、それぞれの階層構造がどの程度照応するのか確認する。NDC は十進分類法として分類項目の上位下位関係が明示されており、全体として木構造をなす。BSH/NDLSH も上位語・下位語の関係による階層構造をもつが、複数の異なる上位語をもつ場合などがあり、全体としてはグラフ構造である。BSH と NDLSH の個々の件名に付与されている NDC 分類記号を手がかりに、BSH と NDC の組み合わせ、NDLSH と NDC の組み合わせにおいて、どの程度上位下位関係の階層構造が一致するのか定量的に調査した。

本研究と同様なアプローチを採用した研究はないが、和田らは個々の書誌レコードに付与されている分類記号と件名標目の組み合わせが、件名標目表において指示されている対応づけとどの程度一致するのか調査している¹⁾。

2. 階層構造の照合処理

それぞれの機械可読形式データを入手 (NDLSH は Web NDL Authorities から 2013年8月7日時点のデータを取得) し²⁾⁴⁾、BSH と NDLSH については、参照や細目のみのものを除いた。その結果、BSH は 8,034 件名、

NDLSH は 19,184 件名であった。なお、これらすべての件名が、NDC9 版の対応する分類記号をもっている (BSH では平均 1.34 個、標準偏差 0.64、NDLSH では 1.37 個、標準偏差 0.62)。また、NDLSH はこのうち 4,615 件名 (24.1%) は、上位語も下位語ももたない孤立した件名であった (BSH では 433 件、5.4%)。

NDC については、本表データから二者択一項目や不使用項目などを除いた分類項目 9,432 を採用した。加えて、BSH 件名には付与されているが、NDC 本表にはない 435 項目を追加した (計 9,867 項目)。同様に、NDLSH 件名には付与されているが、NDC 本表に設定されていない 1,255 項目を追加した (計 10,687 項目)。

これら前処理済みデータを用いて、① BSH/NDLSH をベースにして、件名の上位語・下位語関係に対して、それぞれの件名に対応づけられた NDC 分類項目間が同じく上位下位関係にあるのかプログラムで照合し、不一致の箇所を検出した。同様に、② NDC 分類項目間の上位下位関係に対して、それぞれ対応づけられた BSH/NDLSH の件名間で上位語・下位語関係にあるのか照合した。

照合処理においては、以下のような点に考慮した。

1) BSH/NDLSH 件名と NDC 分類項目の間には多対多の対応関係がある。そのため、対応づけられた複数ある件名 (または分類記号) のうち、いずれか一つが上位下位関係にあれば合致したものとする。

2) すべての NDC 分類項目が BSH/NDLSH 件名に対応づけられているわけではない。これは NDC をベースにした照合を行う際に問題となるが、a) 件名をもたない箇所では不一致は発生しない (いかなる件名とも合致する) とみなす、あるいは b) 上位 NDC 項目に対応づけられてい

る件名を継承させて照合することとした。

3)NDCには分類記号が体系をそのまま表現できていない縮約項目(215項目)と不均衡項目(350項目)がある。縮約項目は、有力な下位の区分枝を昇格させ、分類記号の短縮を行った項目、不均衡項目は逆に分類記号上、下位に位置づけられているが、本来は同位にある項目を指す。これらは機械可読データでは印が付されているため、照合実験では、NDCの分類記号をそのまま使用する方式とは別に、上記の縮約項目・不均衡項目について本来の階層構造を表現できる形式に機械的に変換したものを使用した(以下、正規化NDCと呼ぶ)。具体的には、階層の各レベルを3桁数字で表す方式を採用し、これに機械的に変換した。

4)NDCには、分類記号上で階層が飛躍する(漸近的ではない)箇所がある。例えば、007.6(データ処理、情報処理)に続けて007.609(データ管理)が出現し、中間階層の007.60がない。階層構造の照合においては問題となるため、機械的に中間階層の項目を生成した。前処理段階において、BSH/NDLSHには出現するが、NDC本表には出現しない分類項目をそれぞれ追加しているため、生成された中間階層項目の数もそれぞれ異なる。BSHとの照合用では336項目、NDLSHでは409項目(正規化NDCの場合419項目)が生成され追加された。

また、照合に以下の処理を追加した場合についても実験した。

1)NDCにおける関連項目の活用:NDC本表において各分類項目の下に関連する分類項目への参照が指示されている。このうち、「をも見よ参照」(→:)を抽出し、実験では当該分類項目と擬似的に等価の項目として照合に用いる方式をも試みた。1,170の項目において、延べ1,564の関連項目が追加された。

2)上位分類項目からの件名の限定的継承:対応するBSH/NDLSH件名をもたないNDC分類項目について、上位分類項目がもつ件名を継承して擬似的に付与する方式をも試みた。

3. 照合実験結果

3.1 BSHとNDCの照合

BSHとNDCの階層構造の照合実験結果を表1に掲載する。

まず、BSHをベースとした照合(表1のA)では、最上位語(248件名)ごとに照合を行った結果、複数の上位語をもつケースがあるため総出現件数名10,421となる。それぞれの件名の上位語・下位語関係に対して、個々の件名に

対応づけられたNDC分類項目間が同じく上位下位関係(または同一項目への対応づけ)にあるのか照合し、上位下位関係にない不一致の箇所を検出した。その結果、不一致の箇所は2,902、全体の27.9%であった(表1のA.①)。その不一致をNDC分類記号の桁数で分けて集計した結果を、表2の「照合A.①」列に示した。NDCの先頭1桁すなわち10区分の段階で既に異なるケースが542(件名出現総数10,421に対して5.2%)、2桁すなわち100区分の段階で異なるケースが620(同5.9%)あった。なお、こうした集計は、件名に複数の分類記号が対応づけられている場合には、その最も大きな桁数での不一致箇所(分類階層上での最も近い箇所)を採用し行っている。ちなみに、この実験においては個々の最上位語から順に階層を降りていく照合法をとっているため、上位語・下位語関係をもたない孤立した件名は結果的に除外されている。

照合の範囲を直近の上位下位という2階層から、3階層に広げ照合したときには、不一致は2,084(20.0%)に大幅に減少した(表1のA.②)。これは、上位下位関係にある件名間では分類記号が上位下位関係(または同一項目)にないが、そのさらに上位語に付与されている分類記号とは上位下位関係(または同一項目)にある場合を合致するとみなした照合である。不一致をNDC分類記号の桁数で分けた集計結果を、表2の「照合A.②」列に示した。

次に、不均衡・縮約項目を本来の体系を表すよう変換した正規化NDCを用いたとき、階層関係が一致しない箇所は2,896となった(表1のA.③)。基本的な照合(表1のA.①)と比べると全体で6件不一致が減少しているが、実際には18件減少し、12件新たな不一致が発生している。不均衡・縮約項目への対処が階層構造の照合において必ずしも効果的とはいえない結果となった。正規化NDCを用い、かつ照合範囲を3階層に拡張したときには、不一致箇所は2,085となった(表1のA.④)。

同様に、BSHが対応づけているNDC分類項目に加えて、NDCによる関連項目を元の項目と擬似的に等価な項目として加えた照合実験を行い、得られた結果を表1のBに示した。不一致の箇所が2,672(表1のB.①)となり、関連項目を加えないときに比べて、230の不一致が減少している。照合の範囲を3階層に拡張した場合、あるいは正規化NDCを用いた場合、表1のAと同じような減少傾向を示した。

次に、NDCをベースにして、NDC分類項

表 1. BSH と NDC の階層構造の照合結果

BSH をベースにした照合: BSH → NDC				
件名の出現総数: 10,421		(件名の異なり数: 7,601; 階層関係をもたない孤立件名は除外)		
	①基本的照合	②3 階層の範囲 での照合	③正規化 NDC を 用いた照合	④正規化 NDC+3 階層の範囲 での照合
A: 最上位語(248 語)ごとに照合				
上位下位関係の不一致	2,902 27.9%	2,084 20.0%	2,896 27.8%	2,085 20.0%
		A.①との相違 -818 (内訳: -818, +0)	A.①との相違 -6 (内訳: -18, +12)	A.②との相違 +1 (内訳: -2, +3)
				A.③との相違 -811 (内訳: -811, +0)
B: 最上位語(248 語)ごとに照合+NDC の関連項目を追加した照合				
上位下位関係の不一致	2,672 25.6%	1,866 17.9%	2,659 25.5%	1,863 17.9%
	A.①との相違 -230	A.②との相違 -218	A.③との相違 -237	A.④との相違 -222
NDC をベースにした照合: NDC → BSH				
分類項目数: 10,203		(9,867+機械的に生成した中間項目 336)		
そのうち、対応する件名をもつ項目数: 5,632				
	①基本的照合	②3 階層の範囲 での照合	③正規化 NDC を 用いた照合	④正規化 NDC+3 階層の範囲 での照合
C: 網目表(100 区分)の単位で照合				
上位下位関係の不一致	2,258 22.1%	2,085 20.4%	2,198 21.5%	2,069 20.3%
		C.①との相違 -173 (内訳: -662, +489)	C.①との相違 -60 (内訳: -111, +51)	C.②との相違 +16 (内訳: -44, +28)
				C.③との相違 -129 (内訳: -614, +485)
D: 網目表(100 区分)の単位で照合+上位分類項目からの件名の限定的継承				
上位下位関係の不一致	2,861 28.0%	2,141 21.0%	2,808 27.5%	2,129 20.9%
	C.①との相違 +603 (内訳: -0, +603)	C.②との相違 +56 (内訳: -58, +114)	C.③との相違 +610 (内訳: -0, +610)	C.④との相違 +60 (内訳: -65, +125)

表 2. 不一致の発生レベル (BSH をベースとした照合結果)

不一致のレベル	照合 A.①	照合 A.②	照合 A.③	照合 A.④
NDC 先頭 1 桁	542	489	542	489
先頭 2 桁	620	586	620	586
先頭 3 桁	768	516	782	525
先頭 4 桁	757	408	746	405
先頭 5 桁	209	83	196	79
先頭 6 桁以降	6	2	10	1
不一致総数	2,902	2,084	2,896	2,085

目間の上位下位関係に対して、それぞれ対応づけられた BSH の件名間で上位語・下位語関係にあるのか照合した。(表 1 の C.①)。

全分類項目数 10,203 のうち、2,258 項目 (22.1%) において不一致となっている。なお、前述の通り、件名をもたない箇所では不一致は発生しない (いかなる件名とも合致する) とみなしている。併せて、対応する件名をもたないときには、上位 NDC 項目に対応づけられている件名を継承させる方式を試みた。その結果が表 1 の D であり、すべての NDC 項目で対応する件名をもつことになり、そのため不一致の箇所が大幅に増加している (2,861 項目、28.0%) (表 1 の D.①)。

表 3. NDLSH と NDC の階層構造の照合結果

NDLSH をベースにした照合:NDLSH → NDC				
件名の出現総数:19,643 (件名の異なり数:14,569; 階層関係をもたない孤立件名は除外)				
	①基本的照合	②3階層の範囲での照合	③正規化NDCを用いた照合	④正規化NDC+3階層の範囲での照合
A: 最上位語(1,533語)ごとに照合				
上位下位関係の不一致	6,343 32.3%	5,559 28.3%	6,303 32.1%	5,530 28.2%
B: 最上位語(1,533語)ごとに照合+NDCの関連項目を追加した照合				
上位下位関係の不一致	6,054 30.8%	5,219 26.6%	5,984 30.5%	5,156 26.2%
NDC をベースにした照合:NDC → NDLSH				
分類項目数:11,096 (10,687+機械的に生成した中間項目 409)				
分類項目数(正規化NDC):11,100 (10,687+機械的に生成した中間項目 413)				
そのうち、対応する件名をもつ項目数:7,717				
	①基本的照合	②3階層の範囲での照合	③正規化NDCを用いた照合	④正規化NDC+3階層の範囲での照合
C: 網目表(100区分)の単位で照合				
上位下位関係の不一致	4,300 38.8%	4,730 42.6%	4,252 38.3%	4,702 42.4%
孤立件名を除いた場合	980 8.8%	1,410 12.7%	932 8.4%	1,382 12.5%
D: 網目表(100区分)の単位で照合+上位分類項目からの件名の限定的継承				
上位下位関係の不一致	5,360 48.3%	4,891 44.1%	5,310 47.8%	4,866 43.8%

3. 2 NDLSH と NDC の照合

NDLSH と NDC の階層構造の照合実験、すなわち NDLSH をベースとした照合、NDC をベースに NDLSH を用いた照合のそれぞれを実行し結果を得た。その結果は、表 3 に掲載してある。ここで採用した照合方式は、すべて BSH と NDC の照合に用いたものと同じである。

NDLSH をベースとした照合では、BSH をベースとした前述の照合と類似する結果となった。件名の総出現数が 19,643 と BSH に比べてほぼ倍の数であるが、NDC への対応づけによる上位下位関係の照合では、不一致の箇所 6,343 と同じく倍の数となった。全体の件名数に対する不一致の割合は 32.3% である。この結果を基点にして、3 階層に照合範囲を拡張したとき、正規化 NDC を用いたとき、さらには NDC の関連項目を追加したときなど、それぞれ不一致が減少する傾向は、BSH の場合と相似する。

他方、NDC をベースにした照合では、BSH の場合に比べて、不一致の割合が増加する。

BSH の場合、全体の 22.1% (表 1 の C.①) が不一致であったのに対し、NDLSH では全体の 38.8% (表 3 の C.①) が不一致となっている。さらに、対応する件名をもたない項目について、上位項目から件名を継承させて照合したとき、不一致が大幅に増加している。これらの大半は、孤立した件名 (4,615 件) が引き起こした不一致である。孤立した件名を除いて集計した場合を、表 3 の C の下に併せて示してある。例えば、C.①基本的照合では、不一致数 980 (全体の 8.8%) であった。

注

- 1) 和田匡路, 川向直樹「件名標目表の代表分類記号と書誌レコードの一致する割合」第 58 回日本図書館情報学会研究大会発表要綱. 2010, p.61-64.
- 2) 日本図書館協会. BSH4-Computer File (基本件名標目表第 4 版機械可読データファイル). 2002.
- 3) 国立国会図書館. Web NDL Authorities.一括ダウンロード用ファイル. <http://iss.ndl.go.jp/ndla/download/> (参照 2013-08-06)
- 4) 日本図書館協会. NDC・MRDF9 (日本十進分類法新訂 9 版) 機械可読ファイル, 1996.