情報検索支援のための語彙拡張：シソーラス統合の事例

大村玲子（慶應義塾大学大学院）reiko.omura@keio.jp

１．はじめに

インターネット時代における統制語彙を「知識組織化体系」（Knowledge Organization System: KOS）と捉えると、その種類や概念の表現要素、構造の複雑性をはじめ、カバーする分野や言語に至るまで多種多様なものが乱立する現況である。また、この時代にはエンドユーザーが主流な検索者になったことから、利用者の立場からも、ワンストップで横断的検索を行い効率的な結果を得るような利便性が求められている。すなわち、KOS 間の「互換性（Compatibility）」や「相互運用性（Interoperability）」を高めることが重要となり、実際に海外では大規模プロジェクトを中心にした事例が数多く報告され運用に至っている。

本研究は、既存の統制語彙の統合を行い、その結果拡張された語彙を「検索シソーラス」[1]という手段で利用者に提供することによって検索を支援することを目的とする。実際には、事例としてスポーツ分野の日本語版検索シソーラスを構築することを想定し、既存のスポーツ領域シソーラス（英語）、教育領域シソーラス（英語）、件名標目表（日本語）の 3 種類の統制語彙の統合を試みた。統合方法には、Zeng と Chan（2004 年）[2]が展開した翻訳・翻案、サテライト、直接写像の 3 種類を利用した。本発表では、統制語彙の統合を行った結果を報告するとともに、シソーラスの特徴である階層が拡張に及ぼす影響を分析する。併せて統合過程における翻訳の問題も取り扱う。

２．統制語彙の統合

２．１　統制語彙の 3 種類

統制語彙には、①スポーツ分野の領域シソーラス *SIRCThesaurus 6*（以下 SIRC）[3]、②広域領域をカバーする『基本件名標目表[第 4 版]』（以下 BSH）[4]、③教育分野の領域シソーラス *ERIC Thesaurus*（以下 ERIC）[5]の 3 種類を利用した。表 1 にその概要を示す。その中でも特に本研究の目的に深く関与する項目は、領域（特定か全領域）、統制語彙の種類（シソーラスか件名標目表）、優先語彙数、言語（英語か日本語）であり、「領域」「統制語彙の種類」「言語」の異なる統制語彙を扱うことになる。

表 1. 統合に利用する統制語彙の概要

	SIRC	BSH	ERIC
領　域	特定領域（スポーツ）	全領域	特定領域（教育）
統制語彙の種類	シソーラス	標準件名標目表	シソーラス
優先語彙数（一般のみ）	6,949	7,847	4,520
言　語	英語	日本語	英語
製作国	カナダ	日本	米国
製作者	スポーツ情報資料センター（スポーツ局認可の NPO）	日本図書館協会（公益社団法人）	教育科学研究所（教育庁配下）
創設年、最新改訂年と版	1981、2002 (第 6 版)	1956、1999(第 4 版)	1964、2015/10
調査に利用した媒体	pdf 版	冊子体	オンライン版

２．２　統合方法の 3 種類

2.1 で説明した統合を可能とする方法として、2004 年に Zeng と Chan が展開した統合方法論のうち、①翻訳・翻案、②サテライト、③直接写像、を選択した。これらは排他的ではなく、相互に補填しあう効果が期待される。ここで、①翻訳・翻案は、既存の統制語彙を翻訳あるいは翻案し、新たな統制語彙を構築すること、②サテライトは、スーパーストラクチャー（上層）構造に対し、サテライトの関係を持つ部分的な統制語彙を構築すること、③直接写像は、異なる統制語彙間の用語の等価関係などに基づいて統制語彙を構築すること、と簡潔に定義づけられる。

２．３　翻訳ツール

翻訳のツールには Weblio 英和辞典・和英辞典 [6]を利用した。Weblio は、研究社新英和・和英中辞典などの辞書や日本語 WordNet を含む語彙を数多く収録し、これらを一気にインターネット上で横断検索できることが強みである。Weblio にない語彙は、Google 検索の結果、使用頻度や信頼性に基づき総合的に判断した。

２．４　統合の手順

3 種の統制語彙は、BSH をスーパーストラクチャーに、SIRC と ERIC をサテライトに位置づけて統合を行った。BSH の優先語「スポーツ」（第 1 階層）は、その下位層である第 2 階層に下位語（NT）45 語と連想語（RT）1 語を持つ。本研究では階層関係の中でも NT の語彙拡張に注目し、上位語（BT）と RT は拡張の可能性を示す箇所にのみ補足的に扱うこととした。つまり、BSH 単体では「スポーツ」は第 2 階層で 45 語に拡張される。

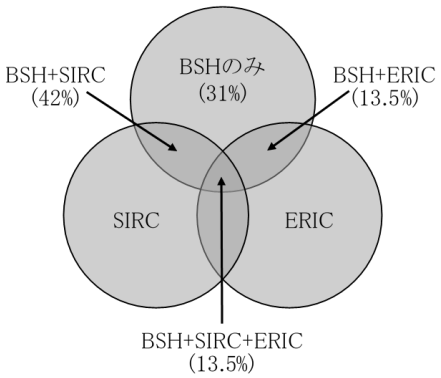
次に、BSH「スポーツ」の翻訳を行い、SIRC の”Sports”と ERIC の”Athletics”に対応づける。そしてそれらに関係づけられる語彙と BSH の第 2 階層 45 語との対応づけへと進む。つまり、BSH45 語の翻訳英語が、SIRC と ERIC の「スポーツ」に関わる語彙と対応づけられる場合に統合を進めるというのが基本的な手順である。そして統合作業は、BSH、SIRC、ERIC それぞれの最下位層まで行うものとした。

３．統合の結果として拡張された語彙

３．１　統合の概観

3 種の統制語彙の第 2 階層に限定し、BSH「スポーツ」の NT45 語を 100%とした場合の、SIRC と ERIC の NT への対応づけの割合を図 1 に表わす。その結果、BSH と SIRC の対応づけが 55.5%（42+13.5%）となり、ERIC との対応づけより多く、従って BSH と SIRC の組み合わせが語彙拡張に一番貢献する可能性があると考えられる。

図 1.「スポーツ」の第 2 階層における対応づけ

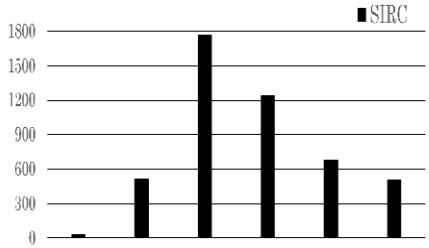


３．２　階層と語彙拡張

語彙拡張のために階層がどのように関わっているのであろうか。表 2 は統制語彙の階層ごとの拡張語彙数を示している。3 種とも階層構造を持つ統制語彙ではあるが、階層数には大きな違いがある。すなわち、BSH や ERIC に比べ、SIRC の階層数は 7 階層と圧倒的に多い。その中でも、階層 4 の語彙数が最も多く（1,771 語）、また階層 2 から 3 への語彙数が 19 倍増と最高に増加している。

統合の結果、BSH の階層 2 の 45 語に対して、単純計算で 100 倍以上の語彙（4,839 語）の拡張がなされたことになる。また BSH 単体での語彙数（108 語）に対しては 45 倍程度の拡張となった。

表 2.「スポーツ」に見る階層と語彙拡張（単位：語彙数）



	階層 1	階層 2	階層 3	階層 4	階層 5	階層 6	階層 7	合計
BSH	1	45	63					108
SIRC	1	27	511	1,771	1,237	680	501	4,727
ERIC	1	4						4
合計		76	574	1,771	1,237	680	501	4,839

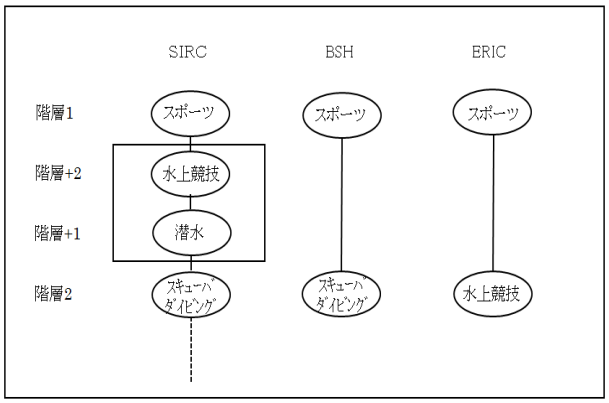
このように、SIRC の階層数の多さが語彙拡張に最も貢献していることは明らかであり、一方において階層数を増やすことでより特定性の高い語彙を追加できるという長所もある。

３．３　階層構造の相違を利用する拡張

SIRC の階層の多さは、統合後に階層 1 と 2 の間に中間層を作る結果ともなり、これも語彙拡張の一役を果たしている。その一例が図 2 に示す「スキューバダイビング」であり、階層 1 の「スポーツ」と階層 2 の「スキューバダイビング」との間に「水上競技」と「潜水」の 2 語と 2 階層が加わることになった。この種の他の用語には「チームスポーツ」「冬季スポーツ」「格闘技」などがあった。

第 2 階層で SIRC に対応づけられた語彙（25/45 語）のうち中間層が伴う（18 語）割合は 72%となり、その内訳は 1 階層増しが 75%、2 階層増しが 21%、そして最大の 3 階層増しが 4%であった。

図 2. SIRC の中間層介在による拡張例

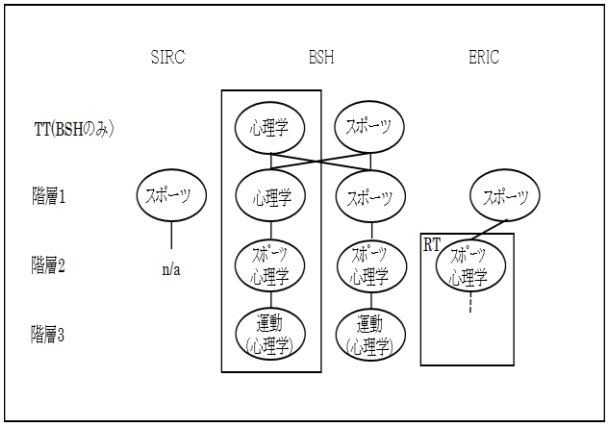


３．４　NT による語彙拡張の限界

ここまでは主に SIRC の階層構造を利用した NT のみによる語彙拡張を説明してきた。シソーラスには NT のほかに BT や RT の用語関係もあり、これらを利用した語彙拡張の可能性もある。NT のみによる語彙拡張の限界に、多分野にまたがる概念を持つ語彙への対応づけがあり、その例が「スポーツ心理学」や「スポーツ医学」である。図 3 に示すように、「スポーツ心理学」は、BSH では「スポーツ」の NT に位置づけられ、異分野の「心理学」とは BSH 独自の TT（最上位標目）で関連付けられる。他方 ERIC は「スポーツ」の RT として「心理学」の NT「スポーツ心理学」に結び付けられている。この例のように、ERIC の構造は階層が少なく（＝NT と BT が少ない）、RT の設定が頗る多いという特徴を持つ。

RT は、”階層的にではなく概念的に密接に関連し、かつ等価集合に入らない用語間の関係”[7] と定義され、異なるカテゴリーに属する用語などを拡張するために有効であることは間違いない。しかし実際は、その定義づけの難しさからも察するように、統合の際どこまで連想関係を拡張するかの判断が困難である。

図 3. 多分野にまたがる語彙の拡張例



４．翻　訳

４．１　翻訳の概観

スポーツ分野の語彙はカタカナ表記による翻字が多くまた新語も多い。その理由として、スポーツがオリンピックや国際競技大会の繁栄に伴い国際的な分野になったことが挙げられる。一方、その語彙には文化的・地域的影響を多大に引き継ぐものも多い。それにも拘わらず専門用語の英和・和英辞書が十分整備されていないこの分野の語彙の和英訳・翻字作業はとりわけ慎重を極める。

このような場合の翻訳において、シソーラスの特徴である Scope Note や限定語、階層構造は大いに役立った。とりわけ Scope Note が充実している SIRC においては、翻訳における語彙の曖昧さを最小限に止め、最適な翻訳語を選択する手助けになった。また、用語を翻訳する際に必要となる文脈（語と語の意味的関係）を理解するために、シソーラスの階層構造を参考にすることができた。

４．２　翻訳の具体的問題

日本語による検索シソーラスを用意するため、統合の手段としての翻訳、特に英語から日本語への翻訳時における主要な注意事項と問題事項を列記する。（例は全て「球技」より）

【注意事項】

（１）表記のゆれを極力網羅する。

例）“Shoot”(ﾊﾞｽｹｯﾄﾎﾞｰﾙ)シュート、ショット

（２）同義語を極力網羅する。

例）”Pitching”(野球) 投球、ピッチング

（３）和製英語を採用する。

例）”Walk”(野球) フォアボール

【問題事項】

（１）翻字できない名称？

例）”Aki”(球技) ？（発音も不明）

（２）名詞と動名詞の取扱い？

例）”Guard”と”Guarding”(ﾊﾞｽｹｯﾄﾎﾞｰﾙ)

問題事項（１）は、特に欧米以外の国発祥のスポーツでまだ日本に馴染みの薄い用語であるため、正しい発音も特定できず翻字が難しい例である。そして、（２）の名詞と動名詞の問題に関しては、SIRC の場合、動名詞と名詞の関係が上位下位であったり並列であったりと様々であり、勿論翻訳も一律でないため、とりわけ対応が難しい組み合わせである。半自動翻訳を行ってもこの種の注意・問題事項は機械判定が難しく、最終的には人手による修正が必要となることが予想される。

５．おわりに

インターネット時代における KOS 間の相互運用性は、エンドユーザーの情報検索を支援する上で重要である。本研究では、Zeng と Chan の統合方法 3 種を利用し、「領域」「統制語彙の種類」「言語」の異なる 3 種の統制語彙を統合してスポーツ分野に特化したシソーラスの語彙拡張を試みた。その結果、それぞれの統制語彙の階層構造や用語関係に見られる特徴を最大限活かして語彙拡張を可能にする結果を導くことができた。また、翻訳においては、翻字できない用語や取扱いに注意が必要な名詞と動名詞の組合せなど具体的な問題が明らかとなった。従来のシソーラスの存続性が危ぶまれる昨今、既存の統制語彙を容易に安価に有効利用して、検索シソーラスという手段で検索者に情報を提供できるよう研究を続けていきたい。

引用文献

1) Shiri, Ali. Powering search: the role of thesauri in new information environments. Information Today, 2012, p.7-8, (ASIST monograph series).

2) Zeng, Marcia Lei; Chan, Lois Mai. Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for Information Science and Technology. 2004, vol.55, no.5, p.377-397.

3) SIRCThesaurus 6. Sport Information Resource Centre, 2002, (入手 2014-08-13).

4) 基本件名標目表 (BSH). 第 4 版, 日本図書館協会, 1999.

5) ERIC website. Institute of Education Sciences, US Department of Education. https://eric.ed.gov, (accessed 2015-01).

6) Weblio website. ウェブリオ. http://ejje.weblio.jp/, (accessed 2014-08).

7) Aitchison, Jean; Gilchrist, Alan. シソーラス構築法. 第 2 版, 内藤衛亮ほか訳. 丸善, 1989, p.49.