専門語彙を手がかりとした知識構成の展開：生命科学分野を例に

浅石卓真 † 影浦峡 †

† 東京大学大学院教育学研究科
asaishi@p.u-tokyo.ac.jp

中学・高校・大学の教科書における生命科学分野の知識の構成を、専門語彙のネットワーク構造として分析・比較することで、学校段階に応じた語彙体系の特徴を明らかにした。さらに、個々の専門用語に対して５つの概念カテゴリーを導入し、各概念カテゴリーに対応した専門用語のネットワーク上での位置付けを分析することで、どのような専門用語を中心に語彙体系が構成されているかについての基礎的な知見を得た。

1　はじめに

現在、科学技術の急速な進展とその日常生活への浸透に伴い、専門知識を基盤とするコミュニケーションの必要性が社会的に増している。この中で、個人の知識レベルや学習段階に応じた語彙・辞書資源の整備は、専門知識を基盤とするコミュニケーションの円滑化を支援するために有効な方策の一つである。そのためにはまず、知識レベルや学習段階に応じた形で現実に存在する語彙の特徴を把握しておく必要がある。

そこで本研究では、中学・高校・大学の教科書における知識の構成を専門語彙のネットワーク構造として分析・比較することで、学校段階に応じた語彙体系の特徴を明らかにすることを目的とする。なお、体系化された語彙が表す知識の構成には、単語間の paradigmatic な関係を基にした構成と、syntagmatic な関係を基にした構成の２種類がある。前者はシソーラスやオントロジーのように上位・下位関係を中心とした「概念体系」としての知識構成であり、後者はテキスト一般に見られる因果関係などを中心とした「論述構造」としての知識構成である。本研究ではこれらのうち、概念体系としての知識構成に焦点をあわせる。

分析の枠組みを図１に示す。本研究では、知識(概念体系)の構成要素である概念の表象として「専門用語」を位置づけ、概念体系を近似するように専門用語の集合を構成した「専門語彙の体系」を、直接の分析対象とする。

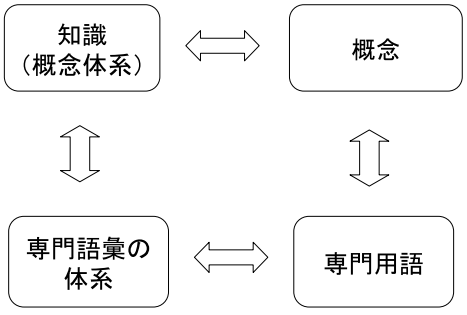

図１　分析の枠組み

2　データ

生命科学分野を分析の対象とした。中学校の「理科」、高校の「生物」で基本的な内容が教えられており、学校段階に応じた知識構成の展開を追うためのモデルとなる事例と考えたためである。各学校段階の教科書として、中学では三浦登ほか「理科（2分野）上下」（東京書籍）を、高校では石川統ほか「生物 I」（東京書籍）を、大学では浅島誠ほか「生命科学」（羊土社）を選択した。

教科書中の索引語を専門用語とみなし、その語構成を利用することで概念体系を近似する語彙体系を作成する。専門用語の語構成は分野の概念体系に準拠してなされており、語構成要素の共有は概念上のつながりを一定程度反映しているとみなすことができる [3]。そこで、各専門用語を学術用語語基表 [4] に従い語基分割した後、専門用語を頂点、語基の共有関係を辺とする語彙ネットワークを作成する[1]。表１に、専門用語の語構成上の統計量を示す。表１の T は専門用語数、N と V は延べ語基数と異なり語基数、N_c と V_c は削除語基を除いた後の延べ語基数と異なり語基数を、S は単一語基から構成されている専門用語の数を表す。

さらに、語彙体系がどのような専門用語を中心に構成されているかを分析するため、概念の種類に応じて専門用語を分類した上で、語彙ネットワーク上での各専門用語の位置付けを観察する。ここでは Sager (1990) と Kageura (2002) を参考に、以下の５つのカテゴリーに分類した [3][1]。表２に概念カテゴリー別の専門用語の内訳を示す。表２から、学校段階が上がるごとに「ME」の比率が低下し、その他の比率が増加していることが分かる。これは、知識の重点が「何があるか」から「いかにあるか」に移行するためと考えられる。

物質的実体（ME）（例）「被子植物」「肝臓」
抽象的実体（AE）（例）「遺伝子暗号」
動作（AC）（例）「受動輸送」「光合成」
性質（QL）（例）「二重らせん構造」「恒常性」
関係（RL）（例）「誘導の連鎖」「濃度差」

[1] ただし、概念に直結しない語基（「の」「的」「化」、数字など）は削除語基として共有関係には含めない。

表 1　語構成上の統計量

	T	N	V	N/T	N/V	N_c	V_c	N_c/T	N_c/V_c	S（%）
理科（２分野）	113	180	120	1.593	1.500	169	114	1.496	1.482	50（44 %）
生物 I	627	1,146	628	1.828	1.825	1,037	600	1.654	1.728	240（38 %）
生命科学	480	1,023	489	2.131	2.092	886	464	1.846	1.909	143（30 %）

表 2　概念カテゴリー別の専門用語数

	ME	AE	AC	QL	RL	合計
理科（２分野）	93（84 %）	0（0 %）	15（13 %）	4（3 %）	1（1 %）	113（100 %）
生物 I	450（72 %）	17（3 %）	86（14 %）	62（10 %）	12（2 %）	627（100 %）
生命科学	321（67 %）	13（3 %）	100（21 %）	27（6 %）	19（4 %）	480（100 %）

3　分析指標

3.1　語彙体系全体の分析指標

はじめに、語彙体系全体の分析指標を紹介する。

コンポーネントの分布

コンポーネントの分布からは、語彙体系の概要を観察する。例えば、語彙ネットワークが３つのコンポーネントから構成されていれば、３つの独立した概念体系にそれぞれ対応する部分的な語彙体系から全体が構成されていると捉える。

次数分布

次数分布からは、語彙体系上に存在する関連語集合の規模別分布を観察する。語構成要素として頻度 f_i の語基が存在することで、語彙ネットワーク上では規模が f_i で各頂点の次数が $f_i - 1$ のクリークが形成されるが、それらは相互に概念上のつながりを持つ関連語集合を表す。また、次数の非常に高い専門用語は、大規模な関連語集合同士の接点に位置していると考えられる。

密度

密度は、語彙体系全体としての概念上のつながりの強さを表しており、専門用語集合の「同質性」の強さと捉えることができる。密度が低ければ、語彙体系は多様な概念を表す専門用語の集合として構成されていると考えられる。

クラスター係数

クラスター係数の高さは、専門用語集合において関連語集合が多く含まれていることを示している。これは、概念体系上において一つの概念に対し関連概念が複数存在する場合が多いことに対応している。

平均頂点間距離

平均頂点間距離は、語彙体系全体の結束性の指標として用いる。ただし、孤立頂点や小規模なコンポーネントは語彙体系上では周縁部分と考えられるので、平均頂点間距離は最大コンポーネント部分に対して適用する。

3.2　中心性指標

次に、語彙ネットワーク上での各専門用語の位置付けを分析するための中心性指標を紹介する。以下、$|G|$ はネットワークの頂点数を表す。

近接中心性

近接中心性 $C_c(i)$ は、以下の式で定義される。

$$C_c(i) = \frac{|G| - 1}{\sum_{i \neq j} d(v_i, v_j)}$$

$d(v_i, v_j)$ は頂点 i と j の距離を表す。近接中心性は、語彙ネットワーク上で地理的・空間的な中心部に位置する専門用語を語彙体系の中心とみなす。

固有ベクトル中心性

固有ベクトル中心性 $C_{ev}(i)$ は、以下の式で定義される。

$$C_{ev}(i) = \frac{1}{\lambda} \sum_{j=1}^{|G|} a_{ij} C_{ev}(j)$$

a_{ij} は隣接行列 A の成分を、λ は A の最大固有値を表す。固有ベクトル中心性は、大規模な関連語集合部分を語彙体系の中心とみなす。

媒介中心性

媒介中心性 $C_b(i)$ は、以下の式で定義される。

$$C_b(i) = \frac{2C'_b(i)}{(|G| - 1)(|G| - 2)}$$

ここで $C'_b(i)$ は以下のように定義される

$$C'_b(i) = \sum_{j=1}^{|G|} \frac{g_{jk}(i)}{g_{jk}}$$

g_{jk} は、頂点 j と k の最短経路数を、$g_{jk}(i)$ はそれらのうちで頂点 i を通る道の数を表す。媒介中心性は、部分的な語彙体系同士をつなぐ役割を果たす専門用語を語彙体系の中心とみなす。

4　分析結果と考察

4.1　語彙体系全体の構造的特徴

表 3 に語彙ネットワークの特徴量を示す。表 3 の $|G|$ と $\|G\|$ は頂点と辺の数を、$\#C$、$|max|$、I はコンポーネントの数、最大コンポーネントの規模、孤立頂点の数を、$\Delta(G)$、Z は最大次数と平均次数を、D、C、L はそれぞれ密度、クラスター係数、平均頂点間距離を表す。

はじめに、語彙ネットワークの概要を観察すると、いずれの学校段階でも約 10〜20 のコンポーネントと多数の孤立頂点から構成されているが、学校段階が上がるにつれて最大コンポーネントの比率が増加し孤立頂点の比率が低下する。これは、語彙体系上で断片的に存在していた多数の専門用語が次第に一つの主要部分に統合されることを示している。これは、学校段階が上がるにつれて、生命科学分野の一部として位置付けの明確な概念が増えるためと考えられる。

次に、次数分布（図 2）からは、学校段階にかかわらず語彙体系上には様々な規模の関連語集合が部分的に重なり合いながら存在していることが窺える。ただし、高校段階では突出して次数の高い専門用語が存在することから、複数の大規模な関連語集合が固まって存在していると考えられる。これらは、専門分野内での位置付けに応じて概念に対する認識の解像度が異なること、また、高校段階では概念体系上の特定の箇所とその周辺で集中的に解像度が高まるが、大学段階では複数の箇所でそれぞれ解像度が高まるためと考えられる。

また、密度の推移から、専門用語集合の同質性は中学から高校にかけては大きく下がるが、高校から大学にかけてはやや高くなることが分かる。これは、高校段階になると専門分野全体をカバーするために多様な概念が含まれるようになるが、大学段階では多様な概念の一つ一つに対して認識の解像度が上がり、関連概念が生じるためと考えられる。

さらに、いずれの語彙ネットワークも、高いクラスター係数と小さい平均頂点間距離を持つ（cf. Newman（2003）[2]）。これは、語彙体系上には多数の関連語集合が存在し、それらが高い結束性を保ちながら全体が構成されていることを示している。これは、一つの基本概念に対して分野特有の派生概念が存在する場合が多く、また概念体系の主要部分では概念の量にかかわらず専門分野としての結束性が保たれているためと考えられる。

4.2　概念カテゴリーに応じた専門用語の位置付け

本節では、概念カテゴリーに応じた専門用語の語彙体系上での位置付けを観察する。まず表 4 に最大コンポーネント（LC）と、その他のコンポーネント及び孤立頂点（$SC+I$）を構成する専門用語集合の概念カテゴリー別の内訳を示す。LC 内で

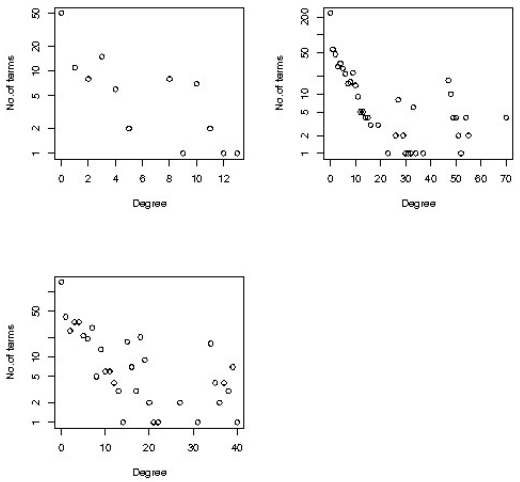

図 2　次数分布（上段左：「理科（２分野）」、上段右：「生物 I」、下段左：「生命科学」）

は「ME」と「AC」の比率が高く、語彙体系の主要部分はこれらの概念を表す専門用語から構成されていることが分かる。

次に、最大コンポーネントに含まれる専門用語の中心性を概念カテゴリー間で比較する。表 5 に近接中心性、表 6 に固有ベクトル中心性、表 7 に媒介中心性の分布の要約統計量を示す。概念カテゴリー間で平均値・最大値を比較すると、いずれの指標でみても殆どの学校段階で「ME」が最も高く、各段階の語彙体系は物質的実体概念を表す専門用語を中心に構成されていることが分かる。次に中心性が高いのは「AE」または「AC」であり、「QL」と「RL」はいずれの中心性も低い。以下、学校段階に応じた中心性の推移を述べる。

近接中心性については、どの概念カテゴリーでも中学から高校にかけて大幅に低くなり、高校から大学にかけて高くなる。これは、高校段階では専門用語集合が専門分野全体をカバーしつつ分散して存在するが、大学段階ではそれらが全体的に語彙体系の中心部に集まることを示している。

固有ベクトル中心性については、平均値の推移は学校段階間及び概念カテゴリーごとに異なるが、どの段階でも「ME」「AE」「AC」の一部には中心性の非常に高いものがあり、「QL」「RL」も一部の段階で中心性が高いものがある。これは、語彙体系上の主要な関連語集合には様々な概念を表す専門用語が含まれることを示している。

媒介中心性については、学校段階が上がるごとに多くの場合で平均値・最大値は低くなる。これは、学校段階が上がり語彙体系に含まれる専門用語数が増加しても、専門分野としての結束性を保つために関連語集合同士が重なり合いながら存在するため、少数の専門用語の媒介機能への依存度が小さくなることを示している。

表 3　ネットワーク特徴量

	$\|G\|$	$\|\|G\|\|$	$\#C$	$\|max\|$（%）	I（%）	Z	$\Delta(G)$	D	C	L
理科（２分野）	113	148	9	21（19 %）	51（45 %）	2.619	13	0.0234	0.721	2.081
生物 I	627	2401	22	321（51 %）	235（37 %）	7.659	70	0.0122	0.678	5.116
生命科学	480	1788	14	308（64 %）	139（29 %）	7.450	40	0.0156	0.701	4.398

表 4　語彙ネットワークを構成する専門用語集合の内訳

		ME（%）	AE（%）	AC（%）	QL（%）	RL（%）	合計
	中学	15（71 %）	0（0 %）	6（29 %）	0（0 %）	0（0 %）	21（100 %）
LC	高校	236（74 %）	12（4 %）	43（13 %）	28（9 %）	2（1 %）	321（100 %）
	大学	191（62 %）	10（3 %）	77（25 %）	17（6 %）	13（4 %）	308（100 %）
	中学	78（85 %）	0（0 %）	9（10 %）	4（4 %）	1（1 %）	92（100 %）
$SC+I$	高校	214（70 %）	5（2 %）	43（14 %）	34（11 %）	10（3 %）	306（100 %）
	大学	130（76 %）	3（2 %）	23（13 %）	10（6 %）	6（3 %）	172（100 %）

5　まとめ

本研究では、生命科学分野の知識構成を、概念体系を近似する専門語彙のネットワーク構造として捉え、中学・高校・大学の各学校段階に応じた語彙体系の特徴を明らかにした。また、専門用語を５つの概念カテゴリーに分類し、各カテゴリーに対応した専門用語の語彙ネットワーク上での位置付けを分析することで、語彙体系がどのような専門用語を中心に構成されているかについての基礎的な知見を得た。

今後は、語彙ネットワーク上でそれぞれの中心性が特に高い専門用語集合の属性を詳しく分析すると共に、専門用語間における具体的な関係の種類も考慮して分析を行うことで、各学校段階における語彙体系の特徴をより詳細に明らかにしていきたい。また、本研究は生命科学分野を対象とした事例分析に留まるが、知識の構成やその学校段階に応じた展開過程は、専門分野ごとに異なることが予想される。そこで、本研究と同様の分析を複数の分野に適用して分野間比較を行うことで、専門分野の特性に応じた語彙体系の特徴を明らかにしたいと考えている。

謝辞

本研究は国立情報学研究所企画型共同研究「異種情報源の特性を考慮した実用的な専門用語対訳辞書の構築と活用」の支援を受けました。ここに謝意を表します。

参考文献

[1] K. Kageura. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth.* John Benjamins, Amsterdam, 2002.

[2] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, pp. 167–256, 2003.

[3] J. C. Sager. *Practical Course in Terminology Processing.* John Benjamins, 1991.

[4] 野村雅昭, 石井正彦. 学術用語語基表. 国立国語研究所, 1989.

表 5　近接中心性の分布の要約統計量

		平均	最大値	最小値	分散
	中学	0.54547	0.66667	0.31746	0.01504
ME	高校	0.11596	0.14050	0.00309	0.00039
	大学	0.24114	0.34035	0.12025	0.00193
	中学	-	-	-	-
AE	高校	0.11239	0.13665	0.09435	0.00026
	大学	0.23187	0.30187	0.14447	0.00197
	中学	0.43357	0.62500	0.30769	0.01088
AC	高校	0.10908	0.13990	0.07067	0.00035
	大学	0.23121	0.31913	0.14881	0.00178
	中学	-	-	-	-
QL	高校	0.08601	0.13591	0.00309	0.00076
	大学	0.21890	0.27833	0.18674	0.00084
	中学	-	-	-	-
RL	高校	0.10878	0.10880	0.10876	0.00000
	大学	0.21059	0.28426	0.16141	0.00191

表 6　固有ベクトル中心性の分布の要約統計量

		平均	最大値	最小値	分散
	中学	0.67107	1.00000	0.01091	0.19578
ME	高校	0.18326	1.00000	4.59322E-15	0.11890
	大学	0.12548	0.99401	1.58472E-09	0.10375
	中学	-	-	-	-
AE	高校	0.16033	0.92172	1.99055E-07	0.12657
	大学	0.20690	0.99200	1.58472E-09	0.16932
	中学	0.24241	0.97519	0.01090	0.13078
AC	高校	0.09890	0.93095	1.04007E-13	0.07470
	大学	0.13115	1.00000	9.23788E-09	0.10320
	中学	-	-	-	-
QL	高校	0.03704	0.92129	1.04007E-13	0.03256
	大学	0.00967	0.05764	2.66346E-05	0.00028
	中学	-	-	-	-
RL	高校	0.00087	0.00087	0.00087	0.00000
	大学	0.06798	0.98245	1.35432E-08	0.05628

表 7　媒介中心性の分布の要約統計量

		平均	最大値	最小値	分散
	中学	0.07298	0.26842	0.00000	0.01301
ME	高校	0.01130	0.17045	0.00000	0.00057
	大学	0.01298	0.16525	0.00000	0.00072
	中学	-	-	-	-
AE	高校	0.03059	0.15500	0.00000	0.00263
	大学	0.01273	0.04780	0.00000	0.00047
	中学	0.01667	0.10000	0.00000	0.00167
AC	高校	0.01490	0.17553	0.00000	0.00148
	大学	0.00815	0.10536	0.00000	0.00023
	中学	-	-	-	-
QL	高校	0.01430	0.10650	0.00000	0.00089
	大学	0.00787	0.04291	0.00000	0.00016
	中学	-	-	-	-
RL	高校	0.00208	0.00625	0.00000	0.00001
	大学	0.00513	0.02580	0.00000	0.00008