

# 英語学術表現リストの階層的構築

## 言語資源としての機関リポジトリの新しい活用

田 中 省 作

### Hierarchically Creating English Academic Expression Lists: New Application of an Institutional Repository as a Language Resource

Shosaku Tanaka

#### abstract

The most significant academic expressions differ depending on the discipline, institution, or faculty. Therefore, it is necessary to build a large corpus that includes a wide range of academic domains in order to create a more appropriate expression list for researchers in different academic fields. The purpose of this study is to identify discipline-specific expressions, especially multi-word expressions, using institutional repositories and natural language processing techniques. An institutional repository is an online collection of the intellectual output of a research institution. This research regards an institutional repository as one of the language resources that reflects the structure of disciplines in its institution, and enables us to create academic expressions for each faculty approximately corresponding to their disciplines. The proposed method is applied to the Kyushu University institutional repository to demonstrate the effectiveness of creating a multi-word list. Further, to the best of our knowledge, this is the first attempt of application of an institutional repository for an unintended purpose.

#### 1. はじめに

科学論文の作成や読解に求められる英語 (English for Academic Purposes: EAP) には、English for General Purposes とよばれる一般目的の英語とは異なる表現や語法が数多くある。さらに、EAP は分野によって、そのような表現や語法が大きく異なることが知られており、分野に依拠した語彙・表現リストの作成は重要な課題の一つである (田地野・水光, 2005)。本研究は、近年、多くの研究機関が構築し、自組織の研究者らが執筆した論文・記事などの著作物を電子的に蓄積・公開しているデータベース「機関リポジトリ」を、そのような英語学術表現リストの作成に活用する。このような機関リポジトリの本来的ではない活用は、著者の知り得る限りなされておらず、本研究は機関リポジトリの新たな有用性を示すものでもある。

## 2. 英語学術表現と機関リポジトリ

### 2.1 英語学術表現と分野依存性

学術英語 (EAP) における表現は、一般目的の英語 (EGP) とは大きく異なることが知られている (Hutchinson&Waters, 1987)。そのような表現は、一般的な資料では十分にカバーすることは難しい。したがって、それらを網羅的に収集し、リスト化することは EAP における重要な課題の一つである。

学術表現は分野にも強く依拠する (田地野・水光, 2005)。たとえば、“Let  $\alpha$  and  $\beta$  be  $\gamma$  and  $\delta$ , respectively” のような表現は、数式を扱うような分野では頻繁に用いられる定型的なものであるが、そうでないような分野では有用とは言い難い。また、分野の基本的概念にかかわる表現も、同様の振る舞いをすることは明らかであろう。このように学術表現は、分野やそれに対応する分類の下で整備する必要がある。それに加え、機関によってその分野構成が大きく異なることが、問題を複雑化させる。

近年、語彙リストをはじめとした言語資料については、大規模電子化用例集 (コーパス) を用いた編纂も頻繁に行われるようになった (大学英語教育学会基本語改訂委員会, 2003; 京都大学英語学術語彙研究グループ, 2009; 田地野他, 2008; 東京工業大学, 2011)。このようなコーパス・アプローチによる編纂では、適当なコーパスを設定・構築した上で、計量的指標を駆使し、語彙・表現間に優先順位を付すことになる。EAP における語彙・表現リストを考えた場合、コーパスの選定・構築だけでなく、「分野」「領域」といった分類単位、そしてその粒度を規定することも容易ではない。

### 2.2 機関リポジトリ

機関リポジトリ (Institutional Repository) は、自組織の研究者らが執筆した論文・記事などの著作物を電子的に蓄積・公開している、オープンアクセスを指向したデータベースである。学術雑誌の高騰や、大学・公的研究機関の情報保全・公開・発信という社会的流れがその発端となり、今時、ほとんどの研究機関が機関リポジトリを構築している。国内では 2013 年 12 月の時点で、少なくとも 285 の機関リポジトリが稼働している<sup>1)</sup>。なお、機関リポジトリの歴史的経緯については、Crow (2002)、逸村・竹内 (2005) や根本 (2013) を参照されたい。

機関リポジトリは、当該機関が取り扱う分野とその組織構造を強く反映した言語資源の一つとみなすことができる (田中, 2013a)。機関リポジトリに蓄積されている著作物は、当該機関から発信されたものなので、関連分野のなかでも特に当該機関が推進している分野・テーマに関するものに集中する。したがって、このような言語資源に基づいた語彙・表現リストは、当該機関の関係者に関連が深いものが列挙される可能性が高い。さらに、機関リポジトリの著作物だけではなく、そこで参照されているような文献を集積することで、当該機関の取り組んでいるテーマに周縁的な言語資料の構築も期待できる。

機関リポジトリは、当該機関の組織構造を軸に著作物を管理していることが多い。代表的な機関リポジトリシステムである DSpace<sup>2)</sup> では、“community” という概念によって著作物を束ねており、それがちょうど「学部・研究科」や「学科」といった組織単位に相当している。したがって、組織構造を勘案した言語資料の作成に、機関リポジトリは大きな助けとなる。

一方、機関リポジトリの現状には問題もある。機関リポジトリはまだ歴史が浅く、機関リポジト

りに対する研究者らの認識は高くない。機関リポジトリに著作物を積極的に登録する研究者も少なく、その結果、多くの機関リポジトリでその蓄積量は十分とは言い難い。比較的整備が進んでいるといわれる九州大学の機関リポジトリ（九州大学学術情報リポジトリ：QIR）<sup>3)</sup>でさえ、直接蓄積している英語著作物は2012年7月時点で5,838点であった<sup>4)</sup>。教員の研究者情報データベース<sup>5)</sup>の登録情報と機関リポジトリの蓄積状況を対比すると、著作権との兼ね合いで必然的に登録されていないものもあるとはいえ、その差は極めて大きい<sup>6)</sup>。

このような状況下、著作物の登録促進は、現在の機関リポジトリの最も重要な課題となっている。逸村・竹内（2005）は、登録促進のために、機関リポジトリの次のようなメリットを関係者に強調したり、登録負荷を低減化したりすべきであると述べている。

1. 無料でアクセスできるオンライン論文は引用されやすい。
2. 自らの研究成果の認知度を高めることができる。
3. 研究成果の長期保存・利用が保証される。

実際、九州大学では、研究者情報データベースの更新と機関リポジトリを同期させる試みや、著作権処理も含めた代理登録サービスを行っている。しかし、登録増加の大きな契機とはなっていないようである。同様の状況は、多くの機関リポジトリで起こっており、登録促進にむけた新しい動機付けの発意は、機関リポジトリの実質化に向けた至要な課題である。

### 2.3 機関リポジトリを活用した英語学術表現リストの構築

本研究では、前節で述べた機関リポジトリを活用し、2.1節で述べたような英語学術表現を研究機関ごと、そしてそれぞれの組織構造で細分化させた形で構築する方法を提案する。このようなアプローチの利点は、主に次の3点である。

1. ある程度充足した機関リポジトリであれば、それを活用することで、当該機関の扱う分野にかかわる学術表現を確実にとらえることができる。
2. 機関リポジトリが有する組織情報を参照することで、分野との対応がとり易い学部・研究科等に学術表現を分類しつつ、収集することができる。当該機関のEAP教育担当者や学習者にとって、学術表現と学部・研究科との関係情報は有益である。
3. 機関リポジトリから収集した学術表現は、当該機関の研究者にとって、身近で有用なものとなり易い。機関リポジトリの登録者たる研究者自身に、その成果が直接的に還元されれば、著作物登録の新たなインセンティブとなる可能性がある。

前節で述べたように、現在の機関リポジトリにはいくつもの問題があり、上述の条件部を十分に満たしているような状況ではない。そのような点で議論の余地があるものの、本研究では、機関リポジトリの新しい活用のパイロットスタディとして、まず現状の機関リポジトリを素直に活用することとする。

### 3. 学術表現リストの作成法

#### 3.1 方針

##### 英語学術表現

本研究で指向する学術表現リストは、松原他（2010）が目指すものと同様で、次節で述べるようなスコアなどで優先順位付けされた表現の集合である。松原他（2010）は、有用な学術表現の特徴として次のような6項目を挙げ、その抽出法を提案している。

1. 高頻度で出現する。
2. 論文に特有の語彙を含む。
3. 短すぎない。
4. 意味的まとまりの列である。
5. 省略表示を含む。
6. 様々な種類の表現と接続する。

本研究では、抽出した表現リストを最終的に関連分野の英語識者が確認・編纂することを念頭に置き、松原他（2010）の抽出法をベースに、文中の依存構造や組織の階層関係を勘案しつつ、英語学術表現リストを作成する。

##### 表現の階層的整備

機関リポジトリから得られる組織情報を参照し、その階層性を表現リストに反映させる。たとえば、「A 大学 B 学部 C 学科」の場合、次のような3つのリストを作成する。

- ・ A 大学の表現リスト
- ・ B 学部の表現リスト
- ・ C 学科の表現リスト

ここで、「A 大学の表現リスト」とは、A 大学の機関リポジトリにある著作物全体から生成されるような表現リストで、上位の表現は比較的どの学部・学科でも使われるようなものである。「B 学部の表現リスト」は、機関リポジトリ内の B 学部の著作物全体から生成されるような表現リストで、「C 学科の表現リスト」も同様である。「A 大学⇒B 学部⇒C 学科」の方向性は、「A 大学内における EGAP（一般学術目的の英語）から ESAP（特定学術目的の英語）」（田地野他, 2005）におおむね対応する。さらに、この階層の最上部に、主に日本人英語学習者を想定し、日本の中高英語を置く。それに対応する著作物は、中高の英語教科書や参考書などが考えられる。つまり、表現リストを、

日本の中高⇒「A 大学⇒B 学部⇒C 学科」

といった具合に階層的に整備する。ちょうど、括弧内で括られた部分が、機関リポジトリに基づき機関ごとに特化される部分である。

このような表現リストの間で、組織の階層関係を考慮し、次のような調整を加える。上部組織の表現リストである一定の上位部分に列挙される表現は、それよりも下部の組織の表現リストでは含めないよう、後処理を行う。このようにすることで、一つ一つの表現リストがコンパクトで、その意味付けもより明確化される。その結果、識者がそれらを編纂する際も判断が下しやすく、最終的な表現リストも使いやすくなる。

### 3.2 手順

機関リポジトリに含まれる英語著作物を事前に組織階層別に分け、それぞれで次のように英語学術表現リストを生成する。

#### (1) チャンク構造と依存関係の同定

各文を構文解析し、チャンク構造と依存関係を同定する。ここで注目するチャンク構造は松原他(2010)に倣い、補文標識(LC)と内部に句構造を含まない基本名詞句(NC)である。各語は動詞の分詞形を除き原形表記に統一した後に、名詞・動詞といった浅い品詞レベルで細分化する。冠詞・数字・記号はそれぞれ<D>・<C>・<S>に置換する。

たとえば、“This paper shows a new method to solve it.”のチャンク構造は、

$$[_{NC} <D> \text{paper}_N] \text{show}_V [_{NC} <D> \text{new}_J \text{method}_N] [_{LC} \text{to}_T] \text{solve}_V [_{NC} \text{it}_P] <S>$$

となる。ここで、 $x_p$  は原形が  $x$  で品詞が  $p$  の語、 $[_y \mathbf{y}]$  は語列  $\mathbf{y}$  が  $Y$  句で、 $N \cdot V \cdot J \cdot P \cdot T$  はそれぞれ名詞・動詞・形容詞・代名詞・TOを表す。なお、文構造を成していないものは分析対象から除く。

例文中に含まれる依存関係は、次の通りである。

1 : <D>	→	paper <sub>N</sub>
2 : paper <sub>N</sub>	→	show <sub>V</sub>
4 : <D>	→	method <sub>N</sub>
5 : new <sub>J</sub>	→	method <sub>N</sub>
6 : method <sub>N</sub>	→	show <sub>V</sub>
7 : to <sub>T</sub>	→	method <sub>N</sub>
8 : solve <sub>V</sub>	→	to <sub>T</sub>
9 : it <sub>P</sub>	→	solve <sub>V</sub>
10 : <S>	→	show <sub>V</sub>

ここで、“ $x \rightarrow y$ ”は  $x$  が  $y$  に係っていること、最左の番号は  $x$  の文中での出現位置を表している。

#### (2) チャンク構造を考慮し、 $n$ -gram を生成

文の前後に文頭・文末を表す特殊記号@を  $n-1$  個付加し、 $n$ -gram を生成する。その際、NC と LC を跨ぐ場合には、それらの語列を一旦 <NC>, <LC> という 1 記号に置換した列も別途考え、そ

れぞれで  $n$ -gram を生成する。

さきほどの例で  $n=3$  の場合、

@ <D> paper<sub>N</sub>  
 <D> paper<sub>N</sub> show<sub>V</sub>  
 paper<sub>N</sub> show<sub>V</sub> <D>

に加え、

@ <NC> show<sub>V</sub>  
 <NC> show<sub>V</sub> <NC>  
 show<sub>V</sub> <NC> <LC>

なども生成される。 $n$  も 2 ~ 10 という具合にある一定の範囲で動かし、これらを累積的に計数する。

計数は、次の2つの観点で行う。一つは、全ての  $n$ -gram をそれぞれ素直に計数するもので、次項の接続確率の算出に用いる頻度（単純頻度）である。もう一つは、依存関係を考慮して計数するもので、次のような条件を満たす  $n$ -gram  $\mathbf{x}$  のみ計数対象とする。

- ・  $\mathbf{x}$  に内容語が存在する。
- ・  $\mathbf{x}$  内の全ての非内容語の係り先が、 $\mathbf{x}$  内の内容語である。

たとえば、上記の “<D> paper<sub>N</sub> show<sub>V</sub>” や “<NC> show<sub>V</sub> <NC>” は上記条件を満たし、計数対象とする。その一方で、“paper<sub>N</sub> show<sub>V</sub> <D>” は <D> が  $n$ -gram 内で依存関係を結んでいないため、計数対象とはならない。また、上記の例ではないが、“<NC> of<sub>P</sub> <NC>” は内容語を含んでいない、“show<sub>V</sub> <NC> of<sub>P</sub>” で of<sub>P</sub> が show<sub>V</sub> ではなく <NC> に係る場合は、やはり計数対象とはならない。このようにして得られる頻度を、依存関係を考慮した頻度とよぶ。

### (3) スコアリング

生成された各  $n$ -gram  $\mathbf{x}$  に対して、次のようにスコアを与える。

$$\text{score}(\mathbf{x}) = f_d(\mathbf{x})l(\mathbf{x})H_L(\mathbf{x})H_R(\mathbf{x})$$

$f_d(\mathbf{x})$  は  $\mathbf{x}$  の依存関係を考慮した頻度、 $l(\mathbf{x})$  は  $\mathbf{x}$  に内含される語・基本句数である。 $H_L(\mathbf{x}), H_R(\mathbf{x})$  は前後に接続する語のエントロピーで、次のように与える。

$$H_a(\mathbf{x}) = - \sum_y P_a(y|\mathbf{x}) \log P_a(y|\mathbf{x})$$

$a \in \{L, R\}$  で、 $P_L(y|\mathbf{x})$  は  $y$  が  $n$ -gram  $\mathbf{x}$  に左接続する確率、 $P_R(y|\mathbf{x})$  は  $y$  が  $\mathbf{x}$  に右接続する確率で、

それぞれ次のように与える。

$$P_L(y|\mathbf{x}) \triangleq f(y\mathbf{x})/f(\mathbf{x})$$

$$P_R(y|\mathbf{x}) \triangleq f(\mathbf{x}y)/f(\mathbf{x})$$

なお、 $f(\mathbf{x})$  は  $\mathbf{x}$  の単純頻度である。

#### (4) フィルタリング

$\text{score}(\mathbf{x})$  に基づき  $n$ -gram  $\mathbf{x}$  を学術表現の候補を抽出していく。その際、次のいずれかの条件が成立する  $\mathbf{x}$  は、抽出の対象外とする。

- ・  $f(\mathbf{x}) < T$  である。
- ・ 自組織よりも上部組織  $i$  の表現リストで上位  $\beta_i\%$  までに抽出済である。
- ・  $\text{score}(\mathbf{x}') > \text{score}(\mathbf{x})$  かつ  $\mathbf{x}$  を内容語に関して完全に包含するような  $\mathbf{x}'$  が抽出済である。

## 4. 実験

### 4.1 データと条件

2012年7月時点の九州大学機関リポジトリ QIR に含まれる英語著作物 5,838 点のうち、形態素数が 2,000 ~ 10,000 の論文 2,965 点を対象とした。延べ形態素数は 15,146,153 である。これらを学部・研究科レベルに相当する 27 部局に細分化し、九州大学全体に加え、それぞれの部局の表現リストを作成する。中高英語の著作物には、平成 14 年度版検定済中高英語教科書（中学 7 シリーズ、高校 28 シリーズ）の本文部分を活用した。延べ形態素数は 736,933 である。

$n=3 \sim 7$  で、最低頻度  $T$  を 5、 $\beta_{\text{中高英語}} = 10$ 、 $\beta_{\text{九州大学}} = 1$  とした。

形態素解析には TreeTagger<sup>7)</sup>、構文解析には Charniak parser<sup>8)</sup> を利用した。

### 4.2 抽出結果

中高英語教科書では 529 組の表現が得られた。上位 20 組の表現を表 1 に示す。なお、以降の結果では品詞情報は省略する。上位 20 組では基本動詞  $V$  が主語・目的語の関係を結ぶ “<NC>  $V$  <NC>” のパターンが多くを占めている。21 ~ 50 位には “more than <NC>”, “<NC> in the world” や “give <NC> <NC>”, “out of <NC>”, “take <NC> to <NC>” など、中高英語で頻繁に紹介される熟語や文型パターンがみられるようになる。<NC> や前置詞を伴った形で重要語彙が抽出されるものと、典型的な文型パターンを示すようなもの、あるいはその両方の性格を併せもつものが抽出されている。

九州大学全体では 7,744 組の表現が得られた。表 2 はその上位 20 組で、\* 印を付した表現は中高英語教科書で抽出された表現の上位  $\beta_{\text{中高英語}}\%$  に含まれていたものである。表 3 は、中高英語教科書の上位表現を除いた上位 20 組である。ここには特定の分野に強く依拠しないような、論理関係・論拠を指示する表現、図表等を指示する表現、論文や研究の導入に使われる表現が得られていることが分かる。

27の学部・研究科については、紙面の都合上、情報科学系の独立研究科であるシステム情報科学研究院と農学部・研究院の結果を示す<sup>9)</sup>。システム情報科学研究院は、229点の著作物から966組の表現が得られた(表4)。“Information Science and Electrical Engineering”は、システム情報科学研究院の英語名の一部で、著作物の著者情報で頻出する。抽出の本来の意図からずれたものではあ

表1：中高英語教科書の抽出結果（上位20組）

$x$	score( $x$ )	$f(x)$	$x$	score( $x$ )	$f(x)$
<NC> don't	47025	762	<NC> tell <NC>	10715	257
there be <NC>	33635	729	<NC> go to	9688	204
when <NC> be	22378	356	<NC> say <NC>	9185	201
<NC> want to	19587	344	<NC> begin to	7610	160
<NC> think <NC>	14483	332	to make <NC>	6892	148
<D> lot of <NC>	14478	207	<NC> find <NC>	6742	148
<NC> take <NC>	14368	294	look at <NC>	5925	193
<NC> know <NC>	13649	283	<D> kind of <NC>	5686	106
<NC> make <NC>	13420	253	be able to	5566	124
<NC> be going to	10843	175	<NC> come to	5324	124

表2：九州大学全体の抽出結果（上位20組、\*は中高英語教科書の上位表現）

$x$	score( $x$ )	$f(x)$	$x$	score( $x$ )	$f(x)$
*<D> number of <NC>	806805	6226	shown in <NC>	226665	5559
based on <NC>	566340	6281	as follow <S>	200795	2106
*there be <NC>	505452	6497	<NC> be shown in <NC>	195657	2390
*when <NC> be	426625	4120	with respect to <NC>	168394	1697
*where <NC> be	400443	4371	on the other hand	121365	2508
*in order to	349408	3620	used in <NC>	118555	2189
such as <NC>	348493	4569	consist of <NC>	103511	1837
due to <NC>	272857	3603	we use	92952	1679
according to <NC>	252775	3345	given by <NC>	91268	1882
by using <NC>	239129	2762	*more than <C>	88643	1181

表3：中高英語教科書の上位表現をのぞいた九州大学全体の抽出結果（上位20組）

$x$	score( $x$ )	$f(x)$	$x$	score( $x$ )	$f(x)$
based on <NC>	566340	6281	used in <NC>	118555	2189
such as <NC>	348493	4569	consist of <NC>	103511	1837
due to <NC>	272857	3603	we use	92952	1679
according to <NC>	252775	3345	given by <NC>	91268	1882
by using <NC>	239129	2762	<S> corresponding author	84428	1762
shown in <NC>	226665	5559	as shown in <NC>	76972	1267
as follow <S>	200795	2106	similar to <NC>	74675	1517
<NC> be shown in <NC>	195657	2390	related to <NC>	73487	1415
with respect to <NC>	168394	1697	associated with <NC>	72277	1410
on the other hand	121365	2508	there exist <NC>	67924	1191

るが、当組織の特徴的な連語表現をとらえた典型的な例といえる。一方、蓄積されている英語著作物数が少ない上に、それらが特定の講座に集中していることもあり、“for low power”や“user’s <NC> test application time”といったものが上位に含まれている。システム情報科学研究院は情報科学・電気電子工学・通信工学を主な専門分野としており、取り扱っている分野の広さを鑑みれば、これらがこの位置で抽出されるのには違和感がある。なお、一般動詞を中心とする表現では、20位以降“implement”や“execute”, “maintain”, “calculate”など、情報系あるいはその周縁分野で必須のものがあげられ、システム情報科学研究院のどの分野・講座、あるいは共通に使用されるものかという直感が働くようなものである。

農学部・研究院は、808点の著作物から2,738組の表現が得られた(表5)。表4のシステム情報科学研究院とは大きく異なることが分かる。農学部・研究院の結果については、今後、当該機関の識者と協働し、その妥当性について論じていく。

直感的には、中高英語教科書・九州大学全体では、それぞれEAP上の位置づけ、もしくは内容・語法的な相違が反映されている、と考え得る。ただ、現状の機関リポジトリで蓄積された著作物数が大幅に減る学部・研究科レベルでは、特定の研究事例にかかわる過剰に具体的で狭小な表現が散見される。実際、システム情報科学研究院では著作物数が少ない上に、登録がある特定の講座や個人に集中していることがその原因の一つである。これらの問題は蓄積数の増加で自然と解消される可能性もあるが、講座間のみならず下位分野の間で著作物の産出スピードが異なるような場合には、

表4：中高英語教科書・九州大学全体の上位表現をのぞいたシステム情報科学研究院の抽出結果(上位20組)

$x$	score( $x$ )	$f(x)$
for low power	1409	54
International Conference on	1033	37
send <NC> to <NC>	1029	53
Information Science and Electrical Engineering	767	69
Table <C><S>	713	28
our approach be	707	40
Figure <C><S> <NC> of <NC>	696	27
user’s <NC>	676	29
test application time	664	36
extract <NC> from	548	40
shift operation insertion technique	548	23
captured by <NC>	514	87
if <D> number of <NC>	501	19
power consumption of <NC>	484	19
cork board system	468	25
referring to <NC>	417	36
than or equal to <NC>	417	22
participant’s <NC>	409	25
propose for <NC>	399	30
take part in <NC>	394	26

表5：中高英語教科書・九州大学全体の上位表現をのぞいた農学部・研究院の抽出結果（上位20組）

$x$	score( $x$ )	$f(x)$
more or less	2873	82
length of <NC>	2504	130
material and <NC>	2480	135
posterior margin of <NC>	2461	61
grown in <NC>	2419	105
hair on <NC>	2249	166
relationship between <NC>	2217	116
No. of <NC>	2048	75
Kyushu University for <NC>	2004	109
exponential function with <NC>	2003	59
after <C> day	1879	68
description of <NC>	1851	107
wish to express <NC> to <NC>	1828	100
separated from <NC>	1642	120
common water hyacinth	1613	74
slightly longer than	1599	60
amount of <NC>	1574	52
distributed in <NC>	1573	119
relative length of <NC>	1554	69
rate of <NC>	1552	56

本質的な問題となり得る。そこで、組織内の著者情報を参照し、「著者」という観点で論文を正規化し、特定の講座や研究者による登録の偏りを是正するような方策を現在、検討している。

## 5. まとめ

本研究では、機関リポジトリを活用した英語学術表現リストの階層的な構築法と、実験結果の一部を示した。今後、抽出結果に対する英語教育担当者・各分野の専門家による評価、機関リポジトリの未充足状況の改善、そして著者に関する情報を織り込んだ論文・表現の重み付けを行った抽出法を検討していく予定である。

また、本研究は機関リポジトリを当該機関の発信型言語資源として捉え、機関リポジトリの本来的ではない新しい活用法を示した。その活用の方向性は、必ずしも学術英語などの言語的知識に限るものではない。特に機関内のさまざまな知識・資源の再発見にも有効であると考えており、現在、機関リポジトリに基づいた研究機関内の潜在的な研究連携関係の発見とクラスタ創成に関する研究も推進している<sup>10)</sup>。

謝辞

九州大学 富浦洋一教授、静岡大学 宮崎佳典准教授、立命館大学 安東正玄氏には、各所属大学の

機関リポジトリのデータ供与にご協力頂いた。ここで記して、深く感謝する。

## 注

- 1) <http://www.nii.ac.jp/irp/list/> (最終アクセス: 2013年12月5日)
- 2) <http://www.dspace.org/> (最終アクセス: 2013年12月5日)
- 3) <https://www.lib.kyushu-u.ac.jp/ja/collections/qir> (最終アクセス: 2013年12月5日)
- 4) OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) などによって、他機関の機関リポジトリも含めた論文データベース間の相互運用性が高まっており、ある機関リポジトリを起点に他機関に蓄積されている著作物でも直接アクセスできる環境が整いつつある。このように他機関に蓄積しているものも含めれば、九州大学関係者の英語著作物は、QIR 経由で約 10.7 万点にアクセス可能となっている (2013年12月現在)。
- 5) <http://hyoka.ofc.kyushu-u.ac.jp/search/index.html> (最終アクセス: 2013年12月5日)
- 6) 本研究プロジェクトの一環として、機関リポジトリ立ち上げ支援・データ補完を目的に、教員研究情報データベース上の書誌情報と検索エンジンを利用し、Web 上に公開されている (ただし、論文データベースは除く) 論文の自動収集を試みたことがある。その際は、収集対象とした論文の約 5% しか得られなかった。著作権上の問題もあり、現在、このようなアプローチは採っていない。
- 7) <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/> (最終アクセス: 2013年12月5日)
- 8) <ftp://ftp.cs.brown.edu/pub/nlparser/> (最終アクセス: 2013年12月5日)
- 9) 九州大学では、2000年4月より研究院・学府制度が導入され、研究科は研究組織たる「研究院」と教育組織たる「学府」に分割されている。
- 10) 科学研究費補助金・挑戦的萌芽研究「機関リポジトリを活用した潜在的な研究クラスタの創出」(代表者: 田中省作, 2013-2015年度)。

## 参考文献

- Crow, R. (2002) The Case for Institutional Repositories: A SPARC Position Paper, ARL Bimonthly Report 223.
- 大学英語教育学会基本語改訂委員会 (2003) 大学英語教育学会基本語リスト JACET List of 8000 Basic Words, 大学英語教育学会英語語彙研究会.
- Hutchinson, T., Waters, A. (1987) English for the Specific Purposes, Cambridge University Press.
- 逸村 裕・竹内比呂也 [編] (2005) 変わりゆく大学図書館, 頸草書房.
- 京都大学英語学術語彙研究グループ (2009) 京大・学術語彙データベース 基本英単語 1110, 研究社.
- Lynch, C. A. (2003) Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age, ARL Bimonthly Report 226.
- 松原茂樹・酒井祐太・小澤俊介・杉木健二 (2010) 学術論文からの英語表現集の自動生成, 第7回情報プロフェッショナルシンポジウム, pp.41-44.
- 根本 彰 [編] (2013) 図書館情報学基礎, 東京大学出版会.
- 田地野彰・水光雅則 (2005) 大学英語教育への提言: カリキュラム開発へのシステムアプローチ, これからの大学英語教育 (竹蓋幸生・水光雅則 [編]), 岩波書店, pp.1-46.
- 田地野彰・寺内 一・金丸敏幸・マスワナ紗矢子・山田 浩 (2008) 英語学術論文執筆のための教材開発に向けて: 論文コーパスの構築と応用, 京都大学高等教育研究開発推進センター, Vol.14, pp.111-121.
- 田中省作・富浦洋一 (2013) 機関リポジトリを活用した英語学術表現リストの階層的構築, 言語処理学会第19回年次大会, pp.318-321.
- 田中省作 (2013a) 言語資源としての機関リポジトリ, 「計量的言語研究の諸相」第2回講演会.
- 田中省作 (2013b) 基本句を考慮した  $n$ -gram の計数, 英語コーパス学会第39回年次大会.
- 東京工業大学 (2011) 東工大英単, 研究社.

(本学文学部教授)