様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ４ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５２０６２５

研究課題名（和文）機関リポジトリを活用した大学別発信型語彙リストのオーダメイド作成法

研究課題名（英文）A made-to-order method to make a communicative English vocabulary list for individual universities by utilizing an institutional repository

研究代表者
徳見　道夫（TOKUMI, MICHIO）

九州大学・言語文化研究科（研究院）・教授

研究者番号：９００９９７５５

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：科学論文の執筆や論文の読解において求められる重要な英語語彙は，分野や組織によって異なるため，分野や部局等の組織別に選定されることが望ましい．本研究は，近年，大学などの主要な研究機関で整備されている機関リポジトリ（自機関の著作物を電子アーカイブし公開するオンラインデータベース）を活用し，大学・部局別の重要語彙リストを効率的に作成する方法を提案した．実際に，九州大学を対象として，提案手法による部局別重要語彙リストを作成し，その有用性を確認した．

研究成果の概要（英文）：It is desirable that an important English vocabulary required in writing and reading scientific papers should be chosen on the basis of individual organizations such as disciplines and departments because it differs according to disciplines and organizations. Our research proposes an efficient method to make important vocabulary lists separately for universities and departments by making the most of an institutional repository (an online archive system for publications written by members of the institution) which has been developed at leading research institutions such as universities in recent years. Taking up Kyushu University as a subject of investigation, we have made important vocabulary lists for various departments by using the method and affirmed their value.

研究分野：エリザベス朝文学および英語教育

キーワード：機関リポジトリ　図書館情報学　ESP/EAP　語彙リスト　個人の語彙分布

１．研究開始当初の背景

学習者の環境に応じて重要語を選定した「語彙リスト」は，英語教育において欠かすことのできない教育資料の一つである．学習者にとっての大学英語教育前後を考えてみると，既に日本の中高英語教育や学習環境を考慮した「ALC SVL 12000」や「JACET 8000」といった教養的・一般目的の語彙リストの整備は充実しつつあった．さらに，その発展的な教材として，大学で独自の語彙リストを作成しているところもあり，北海道大学の「北大語彙表」，東京工業大学の「東工大英語学術語彙データベース（東工大英単）」，京都大学の「京大学術語彙データベース」等がある．

このように，一般的語彙リストだけではなく，その先に大学別の語彙リストがあれば，学習者の語彙学習指針ともなり，日本の大学英語教育に資するところは大きい．しかし，このような大学別語彙リストの作成には，次のような問題があった．

(1) 大学によって教学組織や学問領域が異なり，求められる語彙が大きく変わること

教学組織としての学部構成や学問領域が変われば，求められる語彙は当然大きく変化する．同規模の大学でも所属する研究者や伝統的に注力している学問領域が異なるので，求められる語彙が大学間で完全に一致することはない．したがって，他大学で作成された特定大学の語彙リストを，他の大学に転用することはできない．

(2) 全学英語担当教員であっても学内のあらゆる学問領域を把握し，重要語を選定することは極めて難しいこと

大学ごとに編纂された語彙リストは，全学英語教育を担当している教員らの努力による場合が多い．しかし，そのような教員らであっても，学内のあらゆる学問領域を把握し，そこで求められる語彙を網羅することは，大学がかかえる各分野の多様性や日々の進展を鑑みれば，事実上不可能である．

また，大学で編纂される語彙リストについては，次のような構成上の問題がある．

(3) 語彙に教学組織や学問領域等の情報が付与されていないこと

ある学部で重要な語であっても，他学部ではそうでないものは非常に多い．ある大学の語彙リストといっても，その大学の学習者の立場からすれば，自身が所属する学部・学科や，自身の学問領域にかかわりがでる語なのかどうか，整理されている方が望ましい．しかし，既存の大学別の語彙リストには，そのような情報が付与されていない．

このような問題を鑑み，本研究は大学別の語彙リストの作成に，次のような「機関リポジトリ」に着目し，これを活用した作成法を提案した．機関リポジトリとは，研究機関などが自組織の研究者らが執筆した論文や記事・講義ノートなどの著作物を電子アーカイブし，公開しているデータベースである．本課題の申請時，日本では，150 弱の機関リポジトリが稼働していた．機関リポジトリに蓄積された英語著作物は，その大学に極めて密着していると推測され，加えて著者らの所属学部といった情報も付与されており，上述した問題解決に有効であると考えた．

２．研究の目的

本研究の主目的は，各研究機関が備えている機関リポジトリを活用し，前述した問題を解決することである．具体的には，次の 2 点である．

(1) 機関リポジトリを活用し，大学・部局別の語彙リストを効率的に作成する方法（オーダメイド作成法）を確立する．

(2) 提案法を活用した大学別の語彙リストを作成し，一連のプログラム等とあわせて公開を目指す．

３．研究の方法

(1) 学術語彙と機関リポジトリ

①学術語彙の依存性

学術英語(EAP)における語彙は，一般目的の英語とは大きく異なることが知られている(Hutchison&Waters, 1987)．学術語彙は分野にも強く依拠し(田地野&水光, 2005)，それらを考慮しつつ整備しなければならない．また，大学等の研究機関向けの語彙リストの編纂では，その分野構成が機関によって異なることにも留意が必要となる．

近年，語彙リストの編纂に，コーパスが用いられることも多い(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; 東京工業大学, 2011)．このようなコーパスを用いた語彙リストの編纂では，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・構築に加え，「分野」「領域」といった単位，そしてその粒度を規定することは容易いことではなかった．

②言語資源としての機関リポジトリ

機関リポジトリは，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．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． 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ことになる．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 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，研究者らの認識は必ずしも高くない．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．教員の研究者情報データベース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．

(2) 頻度に基づいたナイーヴな方法

近年，重要語彙の選定には，コーパスを活用した方法がよく用いられている．基本的には，コーパス中で頻出する語を重要語の候補とするものである．前節で述べたように，機関リポジトリへの論文の登録状況は十分ではなく，このような方法を素直に適用すると，登録の偏り等が重要語彙の選定にも影響を与える可能性がある．

ここでは，後述する実験でも活用した，2014 年 5 月時点での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）を用いた，単純な頻度ベースの重要語彙選定の結果を確認しておく．後述する実験の条件同様，QIR に含まれている英語論文を形態素解析し，形態素数が 2,000～10,000 の論文は 3,986 編を対象とする．QIR 全体の論文中の語を原形に直し，品詞で細分化した後に計数する．そのうち，名詞・動詞・形容詞・副詞のなかの頻度上位 50 語をみてみると，英文を構成するために欠かせない“be”, not”のような基本語彙（品詞は略）から，学術論文ではどの分野でも使われるであろう”fig”, “(et.) al”等の他に，”cell”や”water”, “temperature”, “soil”など分野に強く依拠したものが含まれていることが分かる．3,986 編の部局の内訳を調べてみると，最も多くを占めるのが農学部・研究院・学府で 1,650 編にもなる．その他，生物系・医薬系部局の論文が 596 編もあり，単純に頻度順で語を列挙すると，これらの部局の影響が強く出てしまうことが分かった．部局のみを計数対象とした場合も，同様の偏りがみられる．関連講座の研究発信ペースの高さや，講座間での機関リポジトリへの登録状況の偏りによるところが大きいと考えられる．

(2) 個人の語彙分布に基づいた重要語彙の選定

①選定方針

前述したように，部局の論文をひとまとめに計数してしまうと，講座の論文産出ペースや機関リポジトリへの登録意識の相違に強く影響される場合がある．そこで，本研究では，部局の語彙分布をその部局に属する研究者個人の語彙分布を平等に合算することで与えた．個人の語彙分布については，機関リポジトリ内に登録されている当該研究者が執筆した論文から，共著状況を勘案しつつ，推定する．このように，部局の語彙分布を構成する単位を「個人（著者）」とすることで，経年によるスタッフの変容にも柔軟に対応することができる利点もある．

②重要語彙の選定法

機関リポジトリから得られる論文 p は，少なくとも次のような情報から形成されるものとする．

$$p = \langle t, \langle\langle a_1, f_1\rangle, \langle a_2, f_2\rangle, \cdots, \langle a_K, f_K\rangle\rangle\rangle$$

t は p の本文テキスト情報，$\langle a_i, f_i\rangle$ は i 番目の著者情報で，a_i は著者名，f_i は p 上の a_i の所属部局名を表す．つまり，上記 p は K 名の共著論文である．機関リポジトリからは，このような論文情報の集合 $P = \{p_1, p_2, \ldots, p_N\}$ が得られる．

p に対する著者情報の集合を，

$$C(p) = \{\langle a_1, f_1\rangle, \langle a_2, f_2\rangle, \cdots, \langle a_K, f_K\rangle\}$$

と表し，部局名 f が著者名 a の著者 $\langle a, f\rangle$ が執筆した論文の集合を $Q(a,f)$ と表すこととする．

$$Q(a, f) = \{p : C(p) \notin \langle a, f\rangle\}$$

前節で述べたように，部局名 f の語彙分布 $\boldsymbol{v}_f$ を，f に所属する複数の著者の語彙分布 $\boldsymbol{v}_{\langle a,f\rangle}$ から合成する．その $\boldsymbol{v}_{\langle a,f\rangle}$ も $\langle a, f\rangle$ が執筆した一般には複数の論文 p の語彙分布 $\boldsymbol{v}_p$ の合成として算出する．ただし，その際，共著状況を考慮する．なお，語彙分布 $\boldsymbol{v}_\alpha$ は確率分布であるので，次のような性質を充たす．

$$\text{i. } \forall w[\boldsymbol{v}_a(w) \geq 0]$$
$$\text{ii. } \sum_w \boldsymbol{v}_a(w) = 1$$

まず，論文 p の語彙分布 $\boldsymbol{v}_p$ は，次のように最尤推定する．

$$\boldsymbol{v}_p(w) \approx \frac{freq(w;p)}{\sum_{w'} freq(w';p)}$$

ここで，$freq(w;p)$ は p の本文 t 中の w の頻度である．

著者 $\langle a, f\rangle$ の語彙分布 $\boldsymbol{v}_{\langle a,f\rangle}$ は，$\langle a, f\rangle$ が執筆した全ての論文の語彙分布から算出する．論文には共著の場合もある．そこで，$\langle a, f\rangle$ が執筆した各論文の語彙分布を，次のように与える．

$$\boldsymbol{v}_{\langle a,f\rangle}(w) \approx \frac{1}{U_{\langle a,f\rangle}} \sum_{p\in Q(a,f)} u(o(a,f;p),|C(p)|)\boldsymbol{v}_p(w)$$

ただし，$o(a,f;p)$ は p における $\langle a, f\rangle$ の著者順を返し，$u(i,K)$ は重みで，次のような性質を充たす．

$$\text{iii. } \forall K \forall i[u(i,K) \geq u(i+1,K)]$$
$$\text{iv. } \sum_{i=1}^{K} u(i,K) = 1$$

つまり，論文が共著の場合，論文の語の使用傾向には，著者順が前の著者の語彙分布がより強く，あるいは直前の著者同等に影響することを仮定し，著者の語彙分布を推定している．$U_{\langle a,f\rangle}$ は正規化項で次のように与えられる．

$$U_{\langle a,f\rangle} = \sum_{p\in Q(a,f)} u(o(a,f;p),|C(p)|)$$

部局名 f の語彙分布 $\boldsymbol{v}_f$ は，P から得られるその部局に所属する著者の語彙分布 $\boldsymbol{v}_{\langle a,f\rangle}$ を，対等に平均化することで算出する．

$$\boldsymbol{v}_f \approx \frac{1}{|A(f)|} \sum_{a\in A(f)} \boldsymbol{v}_{\langle a,f\rangle}(w)$$

ただし，$A(f)$ は f に所属する個人名を列挙した集合で，次のような性質を充たすものとする．

$$\text{v. } \forall a \forall f \exists p[a \in A(f) \supset \langle a, f\rangle \in C(p)]$$

つまり，$A(f)$ は，$a \in A(f)$ となる $\langle a, f\rangle$ が書いた論文が少なくとも 1 編は P に含むよう構成する．

③実験

(1) データと方法

2014 年 5 月時点の九州大学機関リポジトリ QIR の英語科学論文について，メタ情報から著者名や所属部局，対応する PDF ファイルからテキスト化の処理を通して本文を抽出した．それら論文の本文に対し，TreeTagger で形態素解析を行い，形態素数が 2,000～10,000 の論文 3,986 編を実験データとした．形態素解析後，各形態素は全て原形表記に統一した上で，名詞・動詞・形容詞・副詞という浅い品詞レベルで細分化し，計数した．

論文 p における $u(i,|C(p)|)$ は，$1 \leq i \leq |C(p)|$ で一律 $1/|C(p)|$ とした．

また，$A(f)$ には 2013 年 5 月時点の九州大学研究者情報で，所属が確認できた教員名を含めた．P から得られる九州大学所属の個人は延べ 3,729 名に及ぶが，$\sum A(f)$ は 674 である．部局別の語彙分布については，$|A(f)| \geq 10$ となる 12 部局を推定対象とした．部局と著者数の内訳を表 1 に示す．

部局	著者数
医学部・研究院・学府	161
農学部・研究院・学府	112
工学部・研究院・学府	84
システム情報科学研究院・学府	48
理学部・研究院・学府	47
総合理工学府	46
生物資源環境科学府	37
数理学研究院・学府	26
薬学部・研究院・学府	17
歯学部・研究院・学府	15
経済学部・研究院・学府	10
比較社会文化学府・研究院	10

表 1

(2) 結果

QIR 全体すなわち九州大学の語彙分布の上位 50 語を，図 1 に示す．

be, have, use, cell, al, show, figure, fig, not, study, et, result, japan, high, patient, system, also, method, time, university, analysis, effect, value, case, number, as, table, original, service, group, other, model, level, rate, temperature, low, datum, solution, however, follow, obtain, expression, such, water, sample, kyushu, then, condition, function, increase

図 1

九州大学全体では頻度ベースの方法の結果から，大きく変化はしないものの，当然，部局に所属する教員の規模の影響が，やや強く現れることとなった．

同様の条件で，システム情報科学研究院・学府の語彙分布の上位 50 語の図 2 に示す．システム情報科学研究院・学府では，上位 50 位でさえ，劇的に変化し，電気・電子・情報系がよりバランス良く収録されるようになる．

algorithm, system, sensor, nanoparticles, current, solution, size, function, model, fig, problem, sample, density, distribution, plasma, magnetic, query, information, am, pad, field, frequency, datum, level, particle, table, experimental, region, growth, paper, propose, example, loss, node, power, base, string, taste, signal, small, input, apply, measurement, marker, phys, optimization, chip, test, state, machine

図 2

(3) 公開と活用

本研究で作成した 13 種の語彙リストと次に述べる例文については，関係機関・部署との調整がつき次第，公開を予定している．

各語には中高英語教科書での頻度順位や，文書数などを考慮したリランキングのためのスコアに加え，機関リポジトリ中の論文の実例も提示している．これらの語を学んだり，使用したりする際には，実例は極めて有用である．たとえば，図 3 はシステム情報科学研究院・学府における“system”の 5 つの実例，図 4 は医学部・研究院・学府における“system”の実例の一部である．例文には出典 ID を振っており，実例同士が同じ論文，著者らに起因するものかどうかも区別できる．

このように，機関リポジトリを活用し，例文そのものも部局に強く関連するものを容易に提示でき，従来にはない粒度の細かい分野対応が可能となる．

1. The application-specific nature of embedded **system** creates new opportunities to customize processor architecture for a particular application.
2. Furthermore, we analyze security and privacy of our e-voting **system** and RFID **system**
3. For conventional analog video **systems** there are well-established performance standards.
4. For estimating RHP, we apply it to VEIDL, which is a virtual classroom **system**.
5. In Section 4, we demonstrate its potential usefulness by showing some possible extensions of this **system**.

図 3

Hutchinson, T. and Waters, A. (1987) English for the Specific Purposes, Cambridge University Press.
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, 大学英語教育学会語彙研究会.
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.
東京工業大学 (2011) 東工大英単, 研究社.
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.

1. This **system** is helpful especially to map the inside of the diseased, structurally complicated or anomalous cardiac chambers.
2. These slave **systems** represent a visuospatial sketchpad for visual images and a phonological loop for speech-based information.
3. We also used this **system** for virtual tours of remote institutions.
4. However, even the modified grading **system** incompletely predicts the prognosis of the individual patient with GIST.
5. We could complete the vesico-urethral anastomosis using the ZEUS **system** for 100 min without any intraoperative complications.

図 4

4．研究成果

(1) 機関リポジトリを活用した大学等機関・部局別語彙リストの作成法

3 で述べたような機関リポジトリの諸問題を鑑み，個人の語彙分布に基づいた語彙リスト作成法を提案した．

(2) 大学・部局別語彙リスト

提案手法に基づき，九州大学および部局別語彙リストを作成した．

(3) 機関リポジトリのデータ操作等にかかわる各種プログラム

機関リポジトリの各種操作，研究者情報に関するデータベースと同期し，論文等を補完するプログラムを作成した．

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）

① 田中省作: 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-, 立命館文學, 第 636 巻（尾田政臣先生退職記念号）, pp.1090-1100 (2014) [査読無]

② 小林雄一郎, 田中省作, 阿部真理子: 情報量基準に基づく習熟度尺度の再検討, 統計数理研究所共同研究リポート, 第 321 巻, pp.29-43 (2013) [査読無]

③ 田中省作: ジニ係数に基づいたランダムフォレストにおける部分木の重要度, 統計数理研究所共同研究リポート, 第 321 巻, pp.15-27 (2013) [査読無]

④ Koyama, Y., Tanaka, S., Miyazaki, Y.,

Fujieda, M.: Development of a Corpus-assisted Writing System for Research Papers by Science and Technology Students, ILAC Selections -Autonomy in a Networked World, pp.65-67 (2013) [査読有]

⑤ 田中省作，冨浦洋一，徳見道夫: 機関リポジトリから得られる著者の語彙分布に基づいた部局別重要語彙の選定，じんもんこん 2014 論文集，第 3 巻，pp.207-212 (2014) [査読有]

⑥ 宮崎佳典，田中省作，才茂真輝: 論文英語要旨に基づいた機関別学術語彙リスト生成プログラムの開発，電子情報通信学会技術研究報告，第 114 巻第 228 号，pp.11-16 (2014) [査読無]

〔学会発表〕（計 8 件）

① Kobayashi, Y., Tanaka, S., Tomiura, Y., Miyazaki, Y., Tokumi M.: Identifying Discipline-specific Expressions Based on Institutional Repository, Digital Humanities Australasia 2014 (2014 年 3 月 19 日, The University Club of Western Australia・Perth, Autstralia)

② 田中省作，宮崎佳典，小山由紀江，藤枝美穂: 分野依存性を考慮した用例提示型英文書作成支援ツールの開発，教育システム情報学会第 2 回研究会 (2013 年 7 月 13 日，千歳科学技術大学・北海道千歳市)

③ 小林雄一郎: 情報量基準に基づく習熟度尺度の再検討，言語研究と統計 2014 (2014 年 3 月 29 日，統計数理研究所・東京都立川市)

④ 田中省作: 分類型ランダムフォレストにおける部分木の重要度，言語研究と統計 2014 (2014 年 3 月 29 日，統計数理研究所・東京都立川市)

⑤ 土田航平，宮崎佳典: 学術機関毎の専門語彙生成を目的とした論文英語要旨取得プログラムの開発，情報処理学会第 76 回全国大会 (2014 年 3 月 12 日，東京電機大学・東京都足立区)

⑥ 田中省作，冨浦洋一，宮崎佳典，徳見道夫: 機関リポジトリの言語資源としての活用−大学毎の部局別英語重要語彙の選定−，第 62 回日本図書館情報学会研究大会 (2014 年 11 月 30 日，梅花女子大学・大阪府茨木市)

⑦ 戸沢信晴，宮崎佳典，田中省作: チャンク情報を考慮した例示型英文書作成支援ツール，外国語教育メディア学会中部支部第 84 回支部研究大会 (2014 年 11 月 22 日，静岡大学・静岡県浜松市)

⑧ 田中省作: タスク駆動型のコーパス構築と情報処理技術，英語コーパス学会第 40 回大会 (2014 年 10 月 5 日，熊本学園大学・熊本県熊本市) [招待講演]

〔図書〕（計 1 件）

① 徳見道夫（監修）田中俊也，江口 巧，大津隆広，鈴木右文，Stephen Laker（編集）: 九大英単，研究社，194 ページ (2014)

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

ホームページ等

６．研究組織

(1) 研究代表者

徳見　道夫（TOKUMI MICHIO）
九州大学・大学院言語文化研究院・教授
研究者番号：90099755

(2) 研究分担者

冨浦　洋一（TOMIURA YOICHI）
九州大学・大学院システム情報科学研究院・教授
研究者番号：10217523

田中省作（TANAKA SHOSAKU）
立命館大学・文学部・教授
研究者番号：00325549

宮崎佳典（MIYAZAKI YOSHINORI）
静岡大学・大学院情報学研究科・准教授
研究者番号：00308701

(3) 連携研究者

小林雄一郎（KOBAYASHI YUICHIRO）
立命館大学・研究員
研究者番号：00725666