# 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

田中 省作
立命館大学 文学部
冨浦 洋一
九州大学大学院 システム情報科学研究院
徳見 道夫
九州大学大学院 言語文化研究院

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．そこで，本発表は機関リポジトリを活用し，著者個人の語彙の確率分布（語彙分布）に基づいた部局別の英語重要語彙の選定法を提案する．まず，機関リポジトリから組織情報を反映した形で英語科学論文を抽出し， 各部局に所属する著者個人の語の確率分布（語彙分布）を，自身の各論文の語彙分布から共著状況を考慮しつつ，推定する．さらに，部局の語彙分布を所属する個人の語彙分布の平均と考え，高確率の語を部局の重要語の候補とする．

# Identification of Important Words for each Faculty based on Author Vocabulary Distributions in Institutional Repository

Tanaka, Shosaku
College of Letters
Ritsumeikan University
Tomiura, Yoichi
Faculty of Information Science and Electrical Engineering
Kyushu University
Tokumi, Michio
Faculty of Language and Cultures
Kyushu University

Since important words in the researchers' vocabulary differ depending on their discipline or organization, these words should be identified under each such divisions. This research proposes a new method to identify important words from the common vocabulary of each faculty based on papers in an institutional repository, which is an online archive system for publications written by members of the institution. The proposed method estimates the vocabulary distribution for each faculty; this vocabulary distribution is a probability distribution estimated based on author vocabulary distributions of each faculty obtained from the published papers stored in the institutional repository. Based on the vocabulary distribution, words that have high probability are extracted as candidates for important words of the faculty-specific vocabulary.

## 1 はじめに

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大きく異なるため，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，大学のような研究・教育機関を念頭に，機関リポジトリから組織情報を反映した形で得られる英語論文に基づいた，当該機関の部局別英語重要語彙の選定法を提案する．

外国語教育における近年の重要語彙選定法の主流は，コーパスなどの言語資料を用いる頻度をベースとしたものである．3.1 節で示すように，当該機関の機関リポジトリから部局別に英語論文を抽出し，部局毎の論文をまとめて計数し，高頻度の語を重要語の候補とすると，各部局の研究分野を十分に反映できていない場合がある．この主因は，機関リポジトリの論文登録の偏りによるところが大きい．

そこで本研究では，バランス良く部局の重要語彙を選定するために，まず，各部局に所属する研究者個人の語の確率分布（語彙分布）を，自身の各論文の語彙分布から共著状況を考慮しつつ推定する．その後，部局の語彙分布を所属する研究者の語彙分布の平均とし，高確率の語を当該部局の重要語の候補とする．

重要語彙を一般の研究者が一から選定することは難しいものの，「これは当該分野で必須となる語

か」といった形で提示される語に対して判断することは，その分野である程度の英語著作物に関わったことがある者であれば難しくはない．本研究は，そのような判断を求める語彙の抽出を効率的に行うものである．

なお，現在，我々は本研究の他にも，機関リポジトリを当該機関の重要な情報資源の一つとして捉え，機関内の研究者間のつながりの発見等への活用も進めている [6]．このような機関リポジトリの本来的でない活用を示し，機関リポジトリに新たな付加価値，関係者に著作物登録のインセンティブを与えることも本研究の副次的目的である．

## 2 学術語彙と機関リポジトリ

### 2.1 学術語彙の依存性

学術英語（English for Academic Purposes: 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 [1]．学術語彙は分野にも強く依拠し [7]，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 [2, 3, 4]．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い．

### 2.2 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 [8]．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．

### 2.3 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．

本節では，4 節の実験でも活用する，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．4 節の実験条件同様，QIR に含まれている英語論文を形態素解析し，形態素数が $2,000 \sim 10,000$ 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

be, have, use, not, show, al, fig, as, time, then, figure, follow, system, result, number, also, table, high, study, et, japan, function, value, cell, case, model, other, water, method, give, such, area, university, condition, analysis, datum, effect, equation, obtain, kyushu, temperature, rate, do, low, solution, group, however, soil, type, plant

図 1: 九州大学（QIR 全体）の頻度上位 50 語

上位 50 語を図 1 に示す（品詞は省略）．そこには，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る．

同様に，部局のみを計数対象とした場合も確認しておく．電気・電子・情報系の研究科であるシステム情報科学研究院・学府は，403 編の英語論文が QIR に含まれている．上記と同様の基準で計数し，全部局の論文 3,986 編における文書頻度が全体の 70%以上，つまり 2,790 以上となる語を除いた後の頻度上位 50 語を図 3 に示す．なお，3,986 編 70%以上の文書で出現する語には，“be”, “have” や前置詞など英文を構成する基本的な語，どの分野でも使われるような “figure” などがそれに該当する．電気・電子・情報系を研究対象としているシステム情報科学研究院・学府であるが，上位 50 語にはシステム LSI や計算機アーキテクチャに関連するものが目立つ．これは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．

## 3 個人の語彙分布に基づいた重要語彙の選定

### 3.1 選定方針

2.3 節で示したように，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

water, fig, et, plant, table, soil, area, forest, effect, rice, temperature, growth, low, rate, cell, production, specie, day, female, activity, system, analysis, concentration, increase, length, leg, level, small, type, condition, content, group, total, surface, acid, different, field, find, species, long, pp, sample, treatment, leaf, segment, change, protein, material, control, distribution

図 2: 農学部・研究院・学府の頻度上位 50 位

system, algorithm, cache, memory, datum, instruction, model, power, pattern, function, set, information, design, problem, input, energy, processor, example, paper, table, size, section, propose, application, program, output, node, performance, let, string, base, bit, approach, fig, consider, reduce, execution, technique, variable, image, call, length, generate, consumption, assume, order, computer, architecture, language, operation

図 3: システム情報科学研究院・学府の頻度上位 50 語

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る．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．

### 3.2 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 $P = \{p_1, p_2, \cdots, p_N\}$ 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．

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

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．

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

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．

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

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．

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

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．

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

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．

## 4 実験

### 4.1 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[9] で形態素解析を行い，形態素数が $2,000 \sim 10,000$ 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，$\sum_f A(f)$ は 674 である [1]．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．

### 4.2 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 4 に示す．

九州大学全体では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が分かる．著者数が最も多かった医学部・研究院・学府の語彙分布のなかで，2.3 節の条件同様，全文書 70%以上の文書で出現しない上位

[1] 674 名に含まれない個人は，学生や研究員，過去，在任していた教員である．

| 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: 著者数（$|A(f)|$）が 10 以上の部局（$f$）

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 4: 九州大学（QIR 全体）の語彙分布の上位 50 位

cell, patient, et, cancer, tumor, expression, group, level, original, fig, factor, analysis, control, gene, treatment, blood, effect, protein, hospital, system, disease, low, table, region, clinical, image, include, increase, role, type, perform, risk, report, mouse, significant, datum, rate, significantly, associate, department, suggest, carcinoma, finding, response, surgery, stroke, year, small, lesion, receptor

図 5: 医学部・研究院・学府の語彙分布の上位 50 語

et, water, fig, cell, plant, effect, table, rice, soil, protein, analysis, rate, temperature, gene, growth, production, area, treatment, acid, low, strain, activity, concentration, level, sample, forest, day, group, different, condition, ml, increase, system, control, weight, food, leaf, solution, content, total, model, indicate, difference, specie, datum, min, reaction, length, type, ph

図 6: 農学部・研究院・学府の語彙分布の上位 50 語

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 7: システム情報科学研究院・学府の語彙分布の上位 50 語

50 語を図 5 に，論文数では最も多かった農学部・研究院・学府の同条件上位 50 語を図 6 に示す．

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 7 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．

紙面の都合上，その他，結果の詳細については，発表時に述べる．

### 4.3 公開と活用

本研究で作成した 12 部局上位 1,000 語と，九州大学全体の上位 2,000 語を，次の URL で公開している．

`http://www.cl.ritsumei.ac.jp/IR/FV-QIR/`

各語には中高英語教科書での頻度順位や，別途報告予定の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 8 はシステム情報科学研究院・学府に

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 8: システム情報科学研究院・学府における “system” の例文

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 9: 医学部・研究院・学府における “system” の例文

おける “system” の 5 つの実例，図 9 は医学部・研究院・学府における “system” の実例である．ただし，ここでは紙面の都合上，一文 20 語以下で，それぞれ異なる論文から一文ずつ示した．上記の公開データでは，例文に出典 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．

## 5 おわりに

本研究は，機関リポジトリから得られる部局別の論文を活用し，著者の語彙分布の合成として部局の語彙分布を算出し，部局別の重要語彙の選定を試みた．このようにすることで，比較的，部局の研究分野をバランス良くカバーすることができる．ただし，その前提となるのは，部局の構成員たる個人が少なくとも一編は，当該の機関リポジトリに登録していることである．よって，機関リポジトリの未充足状況は，大きな問題となるため，他の機関リポジトリとの連携，英語抄録を活用したスムージングのような手立ても検討している [5]．

また，機関リポジトリから得られる語彙の固有の性質など，通常の学術論文コーパスをベースとした語彙 [3] と比較し，明らかとする予定である．さらに，部局の語彙分布における語の発生確率だけではなく，文書数などの別の尺度を導入し，リランキングも試みており，別途報告する．

## 参考文献


1) Hutchinson, T. and Waters, A.: English for the Specific Purposes, Cambridge University Press (1987).
2) 大学英語教育学会基本語改訂委員会: 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 (2003).
3) 京都大学英語学術語彙研究グループ: 京大・学術語彙データベース 基本英単語 1110, 研究社 (2009).
4) 東京工業大学: 東工大英単, 研究社 (2011).
5) 宮崎佳典, 田中省作, 才茂真暉: 論文英語要旨に基づいた期間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, 114(228), pp.11-16 (2014).
6) 小野龍太郎, 冨浦洋一, 田中省作, 上瀧恵里子: オーサトピックモデルを用いた論文分析による潜在的研究グループの発掘, 言語処理学会第 20 回大会年次大会, pp.628-631 (2014).
7) 田地野彰, 水光雅則: 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 (2005).
8) 田中省作: 言語資源としての機関リポジトリ, 第 2 回「計量的言語研究の諸相」講演会 (2013).
9) Schmid, H.: TreeTagger – a language independent part-of-speech tagger (online), available from ⟨http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/⟩ (accessed 2014-11-02).