

情報科入試問題における教科書用語分布と領域特性

岩崎 翼† 赤池 英夫† 赤澤 紀子† 角田 博保† 中山 泰一†

1. はじめに

高等学校の共通教科「情報」は、2022 年度より学習指導要領の改訂に基づき、「情報 I」が必修教科目、「情報 II」が選択科目として新設された。これに伴い、2025 年度大学入学共通テストにおいて、初めて「情報」が正式な出題科目として実施された。これにより、情報科は高等教育との接続を意識した科目としての位置付けが一層明確となり、出題範囲や指導内容の明確化に対する関心が高まっている。

こうした背景のもと、情報科における知識体系の構築や出題傾向の把握に向けた分析が進められており、その一環として教科書に掲載された用語に注目する研究が行われている。先行研究では、情報科の検定教科書の索引に現れる用語を網羅的に収集・分類し、世代ごとの特徴や用語の領域的偏りを明らかにする試みが報告されている [1]。

「情報 I」の学習内容は、学習指導要領に基づき「情報社会の問題解決」「コミュニケーションと情報デザイン」「コンピュータとプログラミング」「情報通信ネットワークとデータの活用」の四つの大領域に分類されており、教科書や入試問題に現れる用語の分布にもこの構造が反映されていると考えられる。

本研究では、高等学校情報科の検定教科書に掲載された用語と、共通テストを含む大学入試問題において実際に出現した用語とを比較分析し、それぞれの出現傾向や領域的特性を明らかにすることを目的とする。特に、教科書用語の分布と入試問題の出題傾向との関係性を明らかにすることで、今後の情報科教育および大学入試における出題・指導の在り方について示唆を得ることを目指す。

そこで、調査の一端として、任意の文書を対象とし、その中で使われている教科書用語(以下、単に用語)[2]を抽出することで、その文書が注目している「情報 I」の領域がある程度推定できるのではないかと考えた。そのためには、文書中の用語の出現を調べる必要がある。文書に対して各用語を一つ一つ探索すればよいが、用語数が多いため効率的ではない。また、単なる完全一致では目的の用語以外も検出されることになるため、別途後処理が必要となる。しかし見方を変えれば、各用語が文書に含まれているかは、文書を、それを構成する単語あるいは形態素に分割し、それら文書構成要素が用語のリストに含まれるか調べればよさそうである。今回は、対象の文書に自然言語処理の前処理として施される形態素解析を行ない、そこから得られる形態素と用語との一致を調査する方針とする。しかしながら、用語は複数の形態素から構成されることも多いため、隣接する複数の形態素の並びが用語であるかどうかを、人手、あるいは別途用意したプログラムで確認する必要がある。いずれの場合であっても、用語が 1 つの形態素として認識されることが多ければ、作業の手間を減らす(効率化する)ことができよう。本報告では、形態素解析ツールとして広く知られている MeCab[5]を利用し、上述の目的のために教科書用語を MeCab 利用者独自の語彙として持つ辞書(ユーザ辞書)を追加することによる効率化について検討した。

† 電気通信大学

2. 関連研究と分析対象

高等学校情報科における知識体系の構築や大学入試との接続に関する研究は近年注目を集めており、特に教科書に現れる用語に着目した分析が進められている。

赤澤ら [1] は、2003 年度から 2023 年度にかけて発行された情報科全教科書の索引用語を網羅的に収集・分類し、世代ごとの中核的用語、新出用語、消滅用語を整理した。特に、用語を「情報 I」の学習指導要領に基づく四つの大領域(情報社会の問題解決、コミュニケーションと情報デザイン、コンピュータとプログラミング、情報通信ネットワークとデータの活用)に対応付けて分析し、各世代における領域的な偏りや出現傾向の違いを明らかにしている。

2.1 教科書用語

これまで著者らは、情報科の検定教科書すべて(「情報 A」「情報 B」「情報 C」「社会と情報」「情報の科学」「情報 I」「情報 II」)の索引に現れる用語を集計し教科書用語と定義した。用語、コード、総意率、掲載教科書をまとめた情報科全教科書用語(以下、教科書用語と呼ぶ)を作成した[3]。ここで、用語とは教科書の索引に現れた用語、コードとは学習指導要領の領域、総意率とはどれだけの教科書で使われているかを示したものである[2]。

本研究では、教科書用語を用いて分析を行う。教科書用語では、文部科学省が公開している高等学校情報科「情報 I」教員研修用教材の分類を利用し「情報 I」の領域を定めた。本研究においては、この「情報 I」の教科書領域(表 1)の大領域を用語の領域分けに利用することにしている。

表 1: 情報 I の領域

第 1 領域	情報社会の問題解決
(ア)	問題を発見・解決する方法
(イ)	情報社会における個人の果たす役割と責任
(ウ)	情報技術が果たす役割と望ましい情報社会の構築
第 2 領域	コミュニケーションと情報デザイン
(ア)	メディアの特性とコミュニケーション手段
(イ)	情報デザイン
(ウ)	効果的なコミュニケーション
第 3 領域	コンピュータとプログラミング
(ア)	コンピュータの仕組み
(イ)	アルゴリズムとプログラミング
(ウ)	モデル化とシミュレーション
第 4 領域	情報通信ネットワークとデータの活用
(ア)	情報通信ネットワークの仕組みと役割
(イ)	情報システムとデータの管理
(ウ)	データの収集・整理・分析
特別な領域	全領域, 多義性, 一般

2.2 情報科入試

本研究では、大学入試問題における情報科関連語彙の出現傾向を把握するために、実際の問題文を対象に形態素解析を行い、教科書用語との一致を調査した。

入試問題の分析にあたっては、まず2025年度から大学入学共通テストに導入された「情報Ⅰ」に加え、これに先行する形で出題された「情報関係基礎」や、複数大学が実施した個別入試問題を対象とした。対象問題には、大学入試センターが公表した試作問題・参考問題・サンプル問題のほか、電気通信大学、広島市立大学、京都産業大学、慶應義塾大学、日本大学などの2025年度個別入試問題が含まれる。さらに、高等学校卒業程度認定試験の「情報」サンプル問題も分析対象に加えた。

これらの問題文に対し、情報科検定教科書の索引語から構成されたユーザ辞書を MeCab に登録し、語彙の一致判定を行った。ユーザ辞書は、用語のカテゴリ・説明を含む新版と、索引語のみを含む旧版の2種を準備し、それぞれについて IPADIC および NEologd をベース辞書として適用することで、計4通りの解析を実施可能な構成とした。

ただし本稿では、過去のヒューリスティクスに基づく語句抽出手法との比較を重視するため、旧教科書用語リストをベースに作成したユーザ辞書を用いた解析結果に限定して報告する。その他の組合せ（新版辞書・ベース辞書の違い）については、今後の比較分析にて活用を予定している。

3. 手法

3.1 形態素解析について

自然言語処理の基礎的な前処理として形態素解析がある。形態素とは、言語学における概念で、意味をもつ表現要素の最小単位である。よって、それ以上分解すると意味をなさなくなる。形態素解析とは、文章を形態素に分解するとともに、その品詞や意味を解析することである。たとえば、「私の名前は太郎です。」を形態素に分解し、品詞とともに示すと表2のようになる。

表2：形態素解析の例

私	の	名前	は	太郎	です	。
名詞	助詞	名詞	助詞	名詞	助動詞	記号

3.2 MeCab について

MeCab は形態素解析に広く利用されている。その公式ページ[5]には以下の紹介がある。

MeCab は京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンです。言語、辞書、コーパスに依存しない汎用的な設計を基本方針としています。パラメータの推定 Conditional Random Fields (CRF) を用いており、ChaSen が採用している隠れマルコフモデルに比べ性能が向上しています。また、平均的に ChaSen, Juman, KAKASI より高速に動作します。ちなみに和布蕪(めかぶ)は、作者の好物です。

Random Fields (CRF) を用いており、ChaSen が採用している隠れマルコフモデルに比べ性能が向上しています。ま

た、平均的に ChaSen, Juman, KAKASI より高速に動作します。ちなみに和布蕪(めかぶ)は、作者の好物です。

公式ページにおける MeCab の使い方示された、標準入力から与えた「すもももももものうち」の解析例を図1に示す。

```
% mecab
すももももももものうち
すもも 名詞, 一般,*,*,*,*, すもも, スモモ, スモモ
も 助詞, 係助詞,*,*,*,*, も, モ, モ
もも 名詞, 一般,*,*,*,*, もも, モモ, モモ
も 助詞, 係助詞,*,*,*,*, も, モ, モ
もも 名詞, 一般,*,*,*,*, もも, モモ, モモ
の 助詞, 連体化,*,*,*,*, の, ノ, ノ
うち 名詞, 非自立, 副詞可能,*,*,*,*, うち, ウチ, ウチ
EOS
```

図1：公式ページでの例

3.2.1 MeCab が使用する辞書について

MeCab はベースとして使用するシステム辞書と、ユーザが各自のドメイン固有の語句を扱うためのユーザ辞書を併せて利用することができる。システム辞書には以下が用意されており、必要に応じて MeCab 本体とは別にダウンロードし使用する。

- ・IPA 辞書: IPA コーパスに基づき CRF でパラメータ推定した辞書であり、これが推奨されている。もともと形態素解析エンジン ChaSen(茶筌)に同梱されていた ipadic を MeCab 用にしたものであり、正確には mecab-ipadic と呼称される(以下、単に ipadic)。
- ・Juman 辞書: 京都コーパスに基づき CRF でパラメータ推定した辞書。
- ・Unidic 辞書: BCCWJ コーパスに基づき CRF でパラメータ推定した辞書。

使用が推奨されている IPA 辞書は、ベースである IPA コーパスの更新がないため、新語や固有名詞に弱いといった欠点がある。この欠点を補うシステム辞書が mecab-ipadic-NEologd[6] である(以下、単に neologd)。この辞書は Web 上の日本語情報源から抽出された、多数の neologisms(新たに創作された言葉や句)を含んでいる。とりわけ Web 上の文書を解析する際には、このシステム辞書とデフォルトの辞書(ipadic)を併用することが推奨されている。neologd は、デフォルト辞書に含まれていない、新語を逐次追加しており近年の文書の形態素解析に向いている。しかし、以前は週に2度の精力的な更新が行なわれていたものの、その最終更新は2020年8月末である。

なお、辞書の更新が頻繁でなく、解析速度の低下を避けたいのであればシステム辞書への、さもなければユーザ辞書への変更が推奨されている。

3.2.2 ユーザ辞書の作成法

ユーザ辞書は、然るべきエントリからなる行を持つ csv ファイルから作成できる。具体的には各単語に対して、以下の行を用意する(ページに収めるために折り返している)。

```
表層形, 左文脈 ID, 右文脈 ID, コスト, 品詞, 品詞細分類 1, \
品詞細分類 2, 品詞細分類 3, 活用型, 活用形, 原形, 読み, 発音
```

品詞により指定の必要がない項目は*とし、辞書生成ツールに自動的に決めさせる項目は空欄のままよい。また、活用語か否かで用意すべきエントリが異なる。今回は、用語を名詞(活用しない語)として登録するため、用意すべき項目は表層形、コスト、品詞(「名詞」で固定)、品詞細分類 1(「一般」で固定)、原形(表層形と同一)、読み、発音のみである。以下に、用語「最短経路」「双方向メディア」の例を示す(ページに収めるために折り返している)。後者では、読みと発音の違いがみられる。

```
最短経路,,4096,名詞,一般,*,*,*,*最短経路,\
サイタンケイロ,サイタンケイロ
双方向メディア,,4096,名詞,一般,*,*,*,*双方向メディア,\
ソウホウコウメディア,ソーホーコーメディア
```

なお、ここでは明示的な指定をしていない左(右)文脈 ID は、その単語を左(右)から見たときの内部状態の識別子であり、コストと併せて形態素の切り出しの判定に使用される[8]。また、コストはその単語の出現しやすさを示しており、値が小さいほど出現しやすいものとして扱われる。

ユーザ辞書として登録したい単語に関するエントリを定義した csv ファイルを、ユーザ辞書に変換するには以下のようにすればよい。

```
$ mecab-dict-index -d/usr/local/lib/mecab/dic/ipadic \
-u user.dic -f utf-8 -t utf-8 user.csv
```

このコマンドには、ベースとする辞書のあるディレクトリを-dの引数、ユーザ辞書名を-uの引数、csvファイルの文字コードを-fの引数、作成する辞書の文字コードを-tの引数として与える。作成されたユーザ辞書を MeCab から利用するには、使用するシステム辞書が収められたディレクトリにある設定ファイル(dicrc)に次の行を加えればよい。

```
userdic = ユーザ辞書へのパス
```

また、作成したユーザ辞書には情報科全教科書用語リストの情報を付与しており、形態素解析とともに用語のカテゴリや説明を見ることができる。

```
質的データと量的データ。
質的データ 名詞,一般,*,*,*,*質,シツテキデータ,シツテキデータ,4ウ,データの性質,量的データ
でないデータ,,0.92,,0.06,東21東21日21,東11東12東11東12東13東14開11数11数12日12第11,,,,東A1,,,
,,,,
ど フィラー,*,*,*,*,*と,と
量的データ 名詞,一般,*,*,*,*量的,リョウテキデータ,リョウテキデータ,4ウ,データの性質,数値
あらかわされ数字の大小に意味をもつデータ,1.00,,0.06,東21東21日21,東11東12東11東12東13東14開11数
11数12日11日12第11,,,,東A1,,,,,
記号,句点,*,*,*,*,*。
EOS
```

3.3 調査対象

今回対象とした用語は、情報科全教科書用語リスト 2024/5/9 改訂版(情報科全教科書用語リスト改 240509.xlsx、以下単に用語リスト)[3]に示された 5,875 語である。これらには「情報 I」には現れないものも多く含んでいるが、まずは全体的な様子を観察するために全ての用語を用いた。このリストは、高等学校情報科において現在までに出版された検定済教科書すべての索引に現れる

用語を整理した用語集(赤澤らによる論文[1]が刊行された時点のもの)を基に、情報処理学会情報入試委員会により用語のカテゴリ付けや、要約説明文付け等の手を加え、再構成されたものである。

まずは、用語リストにある用語が、MeCab のシステム辞書にどの程度登録されているか調査する。単純な方法だが、ある用語 w を名詞と考え、これを形態素解析する。もちろん、 w によっては、その前後の文字と結びつき予期せぬ結果となる可能性もあるが、今回はおおまかな状況確認を行なっている段階であるのでこれ以上は追求しない。そのうえで、 w がいくつの形態素に分割されるのかを調べ

る。用語リストでは、同義語が | で区切られて示されている。文書によって同義語のいずれが用いられるのか不明であるので、同義語のそれぞれを用語として扱うことにする。つまり、用語リストに A|B|C とあれば、A, B, C のそれぞれを用語として扱うのである。

また、同義語を個々の用語に展開した後、用語 w が p (q) r の形をしている場合、それ自身が形態素であることは考えにくい。そこで、 pr と qr に展開し、それぞれを用語とすることが考えられる。ただしその場合、教科書用語ではなくなる場合がある(もちろん pr , qr が教科書用語として現れていることも多々ある)。また、本来は q の意味するところを考えて扱うのが適切であろう。

以下に例を示す。

- ファイアウォール(防火壁)：和訳されたもの。
- KB(キロバイト)：省略された用語の読み。
- 表(データベース)：文脈を指示。
- CATV(Cable Television)：略語をスペルアウト。
- ISO-2022-JP(JUNETコード)：出現時期の違いはあるが同様の存在。
- 文字(テキスト)型：ふたつの用語をまとめて表記したもの。

以上を踏まえ、次の 3 種の条件を組み合わせた 6 つのケースについて調査した。

1. 条件 1: p (q) r の形をした用語を pr と qr に展開し、それぞれを用語として扱う(YES)か否(YES)か否(NO)か。
2. 条件 2: 条件 1 が YES の場合に pr , qr が 1 文字のとき用語として扱う(YES)か否(NO)か。条件 1 が NO であれば本条件は考慮されない。
3. 条件 3: MeCab のシステム辞書として neologd を用いる(YES)か否(NO)か。

条件の組み合わせと対応するケース、および、その条件で用語リストから個々の用語に展開した際の用語数を表 3 に示す。

表 3: 用語リストの用語の扱いによる条件の組み合わせとそれらの用語数

ケース	条件 1	条件 2	条件 3	用語数
1	NO	-	NO	6,656
2	NO	-	YES	6,656
3	YES	NO	NO	6,550
4	YES	NO	YES	6,550
5	YES	YES	NO	6,545
6	YES	YES	YES	6,545

それぞれのケースに対し MeCab で解析した場合、その全用語が隣接する何個までの形態素で表現されるのかを調べた結果(累積カバー率)を表 4 に示す。

表 4: 各ケースに対する累積カバー率(%)

形態素の個数	ケース 1	ケース 2	ケース 3	ケース 4	ケース 5	ケース 6
1	28.20	45.66	29.04	46.89	28.98	46.84
2	68.61	78.41	70.56	80.56	70.54	80.55
3	86.96	92.01	89.53	94.72	89.52	94.71
4	94.74	96.60	96.87	98.32	96.87	98.32
5	97.69	98.53	99.10	99.57	99.10	99.57
6	99.04	99.37	99.60	99.80	99.60	99.80
7	99.61	99.83	99.85	99.94	99.85	99.94
8	99.79	99.92	99.92	99.98	99.92	99.98
9	99.95	100.00	99.97	100.00	99.97	100.00
10	99.98		100.00		100.00	
11	100.00					

たとえば、辞書に ipadic を使い、p (q) r の形の用語を展開しない場合(ケース 1)では、ひとつの形態素になるのは 28.20% であり、隣接する 2 つの形態素になるのは 68.61% である。総じて、辞書として neologd を用いた場合(ケース 2, 4, 6)、ipadic を用いる場合よりも少ない形態素で用語が表されていることがわかる。neologd を用いる他ケースとは僅差であるが、ケース 4 では隣接する 3 つの形態素で、全用語の 95% ほどがカバーできる。表 4 をグラフで表したものを図 2 に示す。

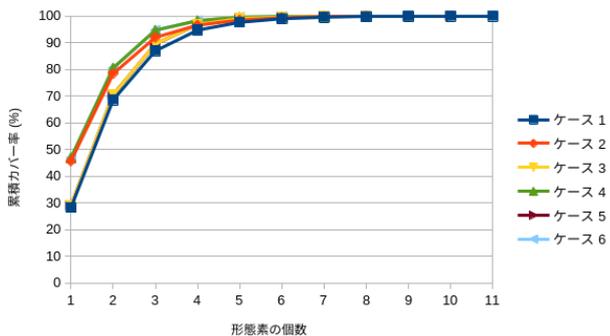


図 2: 各ケースに対する累積カバー率(%)

この結果より、neologd をシステム辞書とし、教科書の索引に p (q) r とある用語を pr, qr に展開して得られる(拡大)用語リスト(ケース 4)から教科書用語用のユーザー辞書を作成することにした。

3.4 用語リストを反映したユーザー辞書の作成

MeCab のシステム辞書として最新の neologd を用いることとして、隣接する 2 つ以上の形態素となる教科書用語をユーザー辞書として用意すれば当初の目的(作業の効率化)を達成できると考えられる。

前節で示したケース 4 では全用語数は 6,550 個である。そのうち、neologd を用いることでひとつの形態素となるものが 3,071 個である。なので、残りの 3,479 個の用語

からユーザー辞書を作れば全用語がひとつの形態素として取り出せると期待される。3.2.2 節で説明した、辞書の元となる csv を用意すればよいが、適切なコストがわからないので 4096 に仮決めした。また、読みと発音であるが、用語に対応する複数の形態素における読みと発音を接続して作成した。たとえば「双方向メディア」は次のように解析される。

双方向 名詞, 一般, *, *, *, *, 双方向, ソウホウコウ, ソーホーコー

メディア 名詞, 一般, *, *, *, *, メディア, メディア, メディア

これから「双方向」の読みと発音は「ソウホウコウ」「ソーホーコー」、「メディア」の読みと発音はともに「メディア」であることがわかる。これらを用いて「双方向メディア」の読みと発音として「ソウホウコウメディア」「ソーホーコーメディア」を得る。他の用語に関しても同様である。

ユーザー辞書を作成し、これが参照されるように設定後、前節と同様の調査を行なった。結果を表 5 およびグラフで表したものを図 3 に示す。

表 5: ユーザー辞書を用いた場合の累積カバー率(%)

形態素の個数	1	2	3	4	5
累積カバー率	97.18	98.79	99.82	99.97	100.00

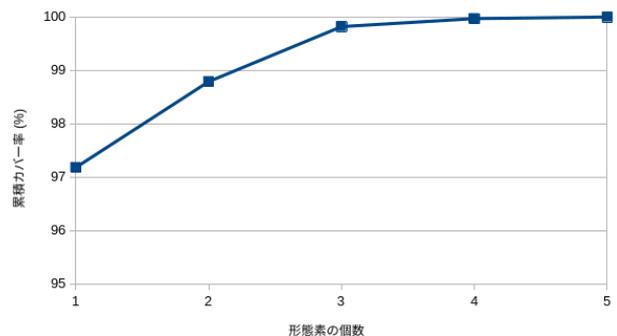


図 3: ユーザー辞書を用いた累積カバー率(%)

ユーザー辞書を用いない場合に比べ、ひとつの形態素として扱われる率は 97% 以上となったものの 100% ではない。たとえば、「デジタルウォーターマーク」はユーザー辞書を追加しても「デジタル」と「ウォーターマーク」のふたつの形態素として扱われる。

デジタル 名詞, 一般, *, *, *, *, デジタル, デジタル, デジタル

ウォーターマーク 名詞, 固有名詞, 一般, *, *, *, ウォーターマーク, ウォーターマーク, ウォーターマーク

これは MeCab のコスト計算の結果を反映していると考えられる。そこで、「デジタルウォーターマーク」のコスト値を 2000 まで下げてから再びユーザー辞書を作成して用い

ると、以下のようにひとつの形態素として扱われるようになる。

デジタルウォーターマーク 名詞, 一般, *, *, *, *, デジタルウォーターマーク, デジタルウォーターマーク, デジタルウォーターマーク

このように、コスト値の調整によりカバー率を高めることはできそうである。しかし、極度のチューニングには、機械学習における過学習のような弊害が現れる懸念がある。

また、今回調査はしていないが、用語の前後にくる文字によって意図しない形態素として取り出される可能性が十分にある。これに対するにはやはり人手、あるいは適当なヒューリスティクスに基づいたプログラムを別途用意して確認する必要がある。これらについては別の課題とする。

4. 結果と考察

著者らの以前の手法では Mecab のシステム辞書である neologd のみを用いて形態素解析を行うものだった(以下、旧手法)。今回はユーザー辞書を新たに作成し、システム辞書と併用することでより詳細な教科書用語の抽出を目指した。

ユーザー辞書におけるコスト値は、デフォルトの値である 4096 を用いていた。しかし、これでは抽出したい用語がシステム辞書によって複数の形態素に分割されてしまい、うまく抽出することができなかった。

そこで、コスト値を 10 に設定しシステム辞書よりもユーザー辞書を優先して使用するよう調整した。

4.1 試作問題の分析における旧手法との比較

すでに旧手法で分析した情報科入試の試作問題に対して、ユーザー辞書を用いた用語の抽出を改めて行い、その結果を表 6 に示す。

表 6 : 試作問題の旧手法と新手法の教科書用語数の比較

	旧	新				
試作問題	a		b		c	
大学入試センター 試作問題	124	129	72	75	51	51
大学入試センター 参考問題	32	20	17	14	15	12
大学入試センター 旧情報	200	190	125	113	88	76
電気通信大学 試作問題	15	20	12	13	10	9

広島市立大学 模擬問題 A	74	74	40	49	36	39
広島市立大学 模擬問題 B	57	47	30	29	25	20
京都産業大学 情報入試模擬試験	108	127	66	81	53	55

a : 試作問題に出現する単語のうち情報科全教科書用語に含まれる用語の数

b : a に該当する用語の中で、情報 I の教科書に登場する用語の数、

c : a に該当する用語の中で、すでに 4 つの教科書領域に分類されている用語の数

抽出できた教科書用語数だけ見るとあまり変化していないようにみられる。そこで具体的にどのような変化が見られたのか c の領域に分類されている用語を以下の表 7 に列挙する。

表 7 : 旧手法でのみ抽出できた用語

第 1 領域	
(ア)	共有, 伝送媒体, 媒体, ブレイクストリーミング
(イ)	情報セキュリティ, 信憑性, 改竄, 認証, コンピュータウイルス, パスワード, 著作物
(ウ)	ソーシャルネットワークワーキングサービス, IT, 情報技術
第 2 領域	
(ア)	パリティビット, パリティチェック, 検索エンジン, E-mail, 圧縮, アナログ
(イ)	ヒストグラム
(ウ)	
第 3 領域	
(ア)	コンピュータ
(イ)	入力, パリティビット, 真理値表, 制御, レジスタ, プログラム, 演算子, 論理回路
(ウ)	アベレージ, 添え字, 標準偏差, アルゴリズム, アプリ, ソート, プログラム, 演算子, 制御, リスト
第 4 領域	
検索エンジン	
(ア)	HTTPS, パリティビット, パリティ検査, 情報セキュリティ, 公開鍵, パケット, サーバ, プロトコル, IT, 認証, 公開鍵暗号, サイト
(イ)	サービス
(ウ)	アベレージ, 散布図, メジアン, 標準偏差, 表計算ソフト, 相関関係, インタビュー, ヒストグラム, 外れ値, 相関
特別な領域	目的, 定義, 設計, 知識, 評価, 要素, 分析

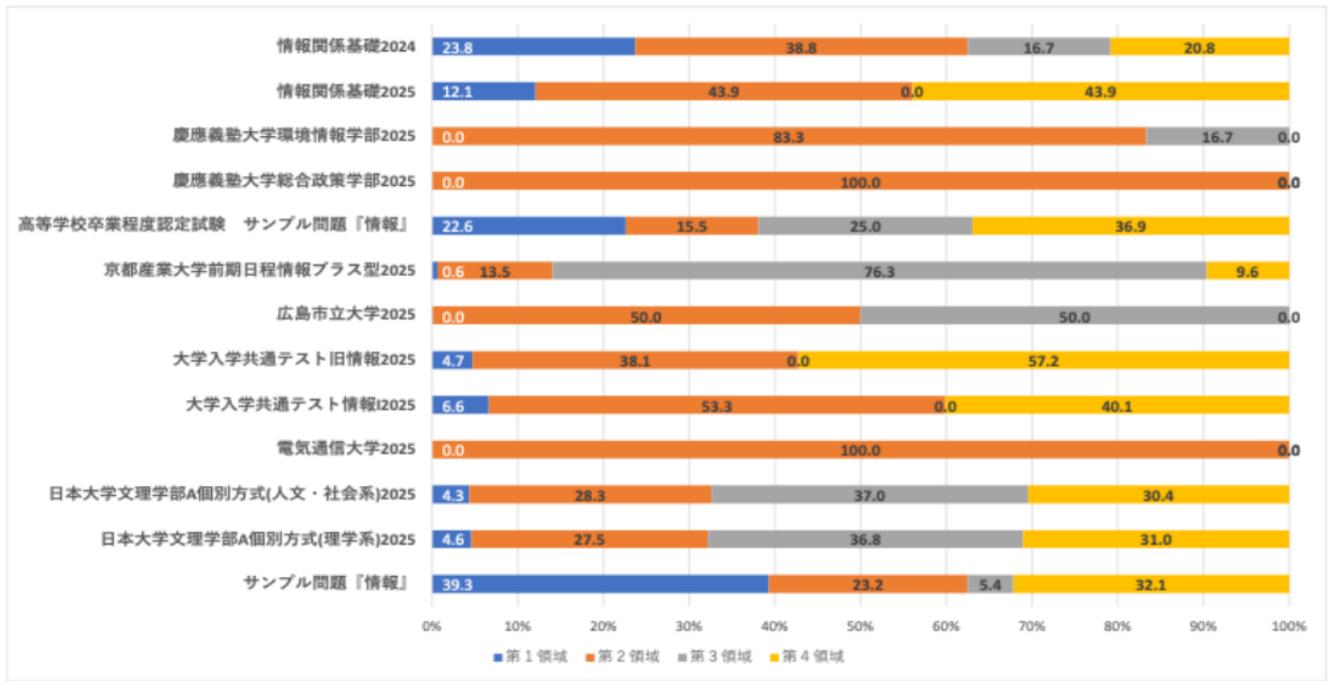


図5: 最近の情報科入試問題の領域分布

表8: 新手法で新たに抽出できた用語

第1領域	
(ア)	
(イ)	機密性, なりすまし, セキュリティ
(ウ)	
第2領域	
(ア)	デジタル, ファイル形式, 情報の量, 符号化, 量子化, bps, アナログ, トリミング, ファイル, 動画
(イ)	HTML, アイコン, ピクトグラム
(ウ)	HTML, トリミング
第3領域	
(ア)	入力装置, 出力装置, OS, ソフト, ハードディスク, 主記憶装置
(イ)	最大値, 比較演算子, ソフト, 並べ替え, 昇順, 選択
(ウ)	
第4領域	
通信	
(ア)	http, 鍵, 共通鍵, LAN, URL, WAN, Wi-Fi, bps, ネットワーク, ハブ, プロバイダ, ルータ, 暗号化
(イ)	キー, ファイル, 昇順, 選択
(ウ)	箱ひげ図, 負の相関, 欠損値, 質的データ, 量的データ, キー, ファイル
特別な領域	
問題,	

表7は旧手法では抽出できたが、新手法では抽出できなかった用語である。ユーザ辞書にまだまだ改良の余地があることを除くと、ここに現れるのはシステム辞書によって分割されてしまった別の用語が考えられる。例えば、第4領域の(ウ)には「相関」という用語が現れる。類似する用語として表8の第4領域の(ウ)には「負の相関」という用語が現れる。ベースのシステム辞書では負の相関

という用語は分割されてしまう(図4)。これに対しユーザ辞書を用いた新手法ではこれが分割されずに新たな教科書用語として抽出された。

負の相関	
負	名詞, 一般, **, **, *, 負, フ, フ
の	助詞, 連体化, **, **, **, の, ノ, ノ
相関	名詞, サ変接続, **, **, *, 相関, ソウカン, ソーカン
EOS	

図4: システム辞書の使用例

4.2 情報科入試の領域分布

新手法を用いて河合塾:キミのミライ発見 [4]で公開されている情報科入試問題に対し、教科書用語の抽出を行い、その領域分布を調べた。以下の図5を参考にされたい。

5. まとめと今後の課題

本研究では情報科入試問題を分析するため、Mecabのユーザ辞書を作成し、教科書用語を抽出した。システム辞書では検出できなかった用語も新たに見つけることができたが、これまで検出できていた一部の用語を見落としてしまうこととなった。また、現在ユーザ辞書で設定しているコストの値は適切とは言えない。

今後の展望としては、ユーザ辞書の更なる改良を進め、情報科入試問題に限らず、一般文書にどのくらい情報科教科書用語が用いられているのか汎用辞書の作成を目指したい。

参考文献

- [1] 赤澤紀子, 赤池英夫, 柴田雄登, 角田博保, 中山泰一: 情報科教科書に現れる用語の変遷 - 情報ABCから情報I・IIまで -, 情報処理学会論文誌「教育とコンピュータ」, Vol.10, No.1, pp. 13-24 (2024).

- [2] 角田博保: 高等学校情報科全教科書用語リスト, 情報処理, Vol. 65, No. 12, pp.622-626 (2024).
- [3] 情報処理学会情報入試委員会: 情報科全教科書用語リスト(オンライン), 入手先, <https://www.ipsj.or.jp/topics/20240412_word.htm>(参照 2025-07-25)
- [4] 河合塾:キミのミライ発見 学部・学科/入試方式別掲載問題一覧(オンライン), 入手先<<https://www.wakuwaku-catch-mondai.net/question/#s2>>(参照 2025-07-25)
- [5] 岩崎翼, 赤池英夫, 赤澤紀子, 角田博保, 中山泰一: 情報科の入試問題に現れる用語と教科書用語との比較検討, 情報処理学会 第 86 回全国大会論文集, 7ZL-04(2024. 3. 17).
- [6] 岩崎翼, 赤池英夫, 赤澤紀子, 角田博保, 中山泰一: 大学入学共通テスト「情報関係基礎」に現れる用語と情報科教科書用語との比較検討, 情報処理学会研究報告コンピュータと教育(CE), 2024-CE-174(2), pp.1-6(2024. 3. 2).
- [7] MeCab: Yet Another Part-of-Speech and Morphological Analyzer (online) , available from<<https://taku910.github.io/mecab/>>(参照 2025-4-20)
- [8] Mecab-ipadic-neologd (online) , available from<<https://github.com/neologd/mecab-ipadic-neologd>>(参照 2025-4-20)
- [9] MeCab: 単語の追加方法(オンライン), 入手先<<https://taku910.github.io/mecab/dic.html>>(参照 2025-4-18)
- [10] MeCab のコスト計算について(オンライン), 入手先<https://qiita.com/Taka_input/items/ae500a5c4440fa2e091e>(参照 2025-4-18)