# コーパスによる難解語・重要語の抽出　―医療用語を例に―

田中牧郎（国立国語研究所）　金愛蘭（国立国語研究所）
桐生りか（国立国語研究所）　近藤明日子（国立国語研究所）

## 1. 背景と目的

現在，人々の生活に密着した領域でも高度に専門化が進み，日々の暮らしの中で生活に必要な情報を得ようとする際にも，専門語に出会う機会が多い。一方で，個人の価値観が尊重され，自己責任が問われがちになってきている昨今，専門語で提供される情報をも，的確に理解し判断を迫られる場合が多くなってきている。ところが，非専門家である一般の人々は，専門語の難解さに困惑を感じる場合が多く，職業としてサービスを提供する側の専門家も，一般の人々とコミュニケーションを取る際に，専門語が障壁になる場合があると感じている。このような，専門家から非専門家に情報を伝える際の，用語の分かりにくさは，現代における言語問題の一つである。この問題のありようを把握し，解決の方向を探ることは，社会言語学や言語政策の研究課題となる。

この課題に取り組むには，いくつかの段階を用意しなければならないが，その初期段階で必要になる作業の一つに，問題を引き起こしている語彙を特定することが挙げられる。この作業を行うのに有効な方法は，①専門用語集の調査，②専門家及び非専門家に対するアンケート調査，③専門家と非専門家のコミュニケーション場面の収録，④コーパスの活用，の四つが考えられる。

本発表ではこのうち，④の方法により，医療の分野を対象に作業を行った経過を報告し，コーパスによって，ある専門分野の難解語と重要語を抽出する方法について考えたい。

## 2. コーパスの設計と作成

### 2.1 基本的考え方

上記の目的の実現のためには，（1）専門分野に特有な語彙（専門語）を抽出すること，（2）専門語のうち，難解な語彙・重要な語彙（難解語・重要語）を抽出すること，の二つの手続きが必要である。

コーパスを使った（1）の実践としては，専門分野のコーパスの語彙と，一般分野のコーパスの語彙を比較し，専門分野に特徴的な語彙を取り出すことが考えられる。同じく（2）の実践としては，一般の人々が専門情報に接する際に，難解と感じる語彙，情報の理解のために重要な語彙を，抽出することが考えられる。

### 2.2 専門語の抽出のためのコーパスの二類別

まず，（1）を実践するために，専門分野のコーパスと一般分野のコーパスとの二種を作成することとし，これを次のように設計した。語数は短単位で集計している。

A医療コーパス　約1,890万語
- ①　医療雑誌 33種　10箇月分
- ②　新聞の医療記事　全国紙1種　23箇月分
- ③　ウェブ上で製薬会社が提供する医療情報　60社分

B一般コーパス　約2,320万語
- 国立国語研究所で開発中の「現代日本語書き言葉均衡コーパス」のうち書籍分

A医療コーパスには，一般人が医療にかかわる情報に接する書き言葉の媒体のうち，代表的だと考えられる，①雑誌，②新聞，③ウェブを取り上げた。①雑誌は，『雑誌新聞総カタログ』（メディアリサーチセンター）で，「厚生・医療」の大分類に掲載されている雑誌から，発行部数が 10,000 部以上の月刊誌をリストアップし，小分類（「内科」「リハビリテーション」「医療管理」など）のバランスに配慮して，33誌を対象にした。2006年9月から2007年6月までの発行分で，約1,370万語の規模である。紙媒体をもとに，スキャナや手入力により，電子データを作成した。

②新聞は，よく読まれている全国紙のうち，ネット上で医療記事をひとまとめにして掲載しているものとして，「読売オンライン」（読売新聞社）の「医療・介護」を対象にして，ダウンロードしてコーパスに含めた。2006年1月～2007年11月までで，約260万語の規模である。

③ウェブは，良質の情報を組織的に提供しているサイトとして，製薬会社が提供する患者向け情報を対象にした。そこには，医薬品に関する情報にとどまらず，医療にかかわる多様な情報が分かりやすくまとめられている。日本薬剤師会のサイトのリンク集に挙げられた製薬会社のうち 60 社のサイトを対象にしてダウンロードし，約250万語分のコーパスとした。

B一般コーパスには，国立国語研究所で開発中の「現代日本語書き言葉均衡コーパス」のうち，2007年11月時点で電子化されている書籍データ 2,320 万語を対象にした。この均衡コーパスは，現代日本語の書き言葉の実態を反映させて設計され，構築が進められているものであり（http://www2.kokken.go.jp/kotonoha/），一般的な書

き言葉を代表するコーパスとして扱った。

Aの語彙とBの語彙とを比較し，Aに特徴的な語彙が，医療分野の専門語と見なすのである。

### 2.3 難解語・重要語の抽出のためのコーパスの二類別

（2）を実践するためには，専門家でない一般人が医療情報に接する媒体を，対象化する必要がある。そこで，A医療コーパスについて，

ａ専門家を読者とするコーパス

ｂ一般人を読者とするコーパス

の二つに分けた。①雑誌は，医師・看護師など医療の専門家を読者に想定した雑誌（『日経メディカル』『エキスパートナース』など）と，一般人を読者に想定した雑誌（『きょうの健康』『がんサポート』など）に分けられる。また，②新聞，③ウェブの患者向け情報は，一般人を読者に想定している。この二類別を行うと，ａ専門家対象の医療コーパスが約1010万語，ｂ一般人対象の医療コーパスが約880万語という規模になる。

ａの語彙とｂの語彙とを比較し，ａに特徴的な語彙は一般人にとってなじみがないので「難解な」語彙であろうし，ｂによく使われている語彙は，一般人が理解しておく必要性の高い「重要な」語彙であろうと見込まれる。

## 3．形態素解析と専門語の抽出

### 3.1 形態素解析

2で設計し作成したコーパスに対して，コンピューターによる「形態素解析」を施した。解析器は「茶筌」（http://chasen-legacy.sourceforge.jp/），解析辞書はUnidic Ver.1.36（http://www.tokuteicorpus.jp/dist/）を用いた。Unidicは，国立国語研究所で設計された「短単位」に基づき，一貫した規則にしたがって斉一に単位切りがされる点で，語彙研究にとって非常に有益な解析辞書である。

しかし，「短単位」は，単純語が基本となっており，複合名詞の多い専門語の分析には不向きな面がある。そこで，Unidicによって切り出された短単位について，我々の研究にとって，一単語と認めることが必要な場合を一つの単位にまとめる規則を立て，作業を行った。具体的には，名詞や形状詞，接辞などが一定のルールで連接するものを，一単位にまとめる規則を立て，この規則に合致するものは，一つの単位に認定し直した。

例：　熱中／症　→　熱中症

　　起立／性／低／血圧　→　起立性低血圧

このようにして作業を行ったデータをもとに，コーパスの全語彙について，語彙頻度表を作成した。全体の異なり語数は，約719,000語となった。

### 3.2 医療の専門語の抽出

2で述べた，（1）専門語の抽出を目的として，A医療コーパスの語彙とB一般コーパスの語彙を比較し，Aに特徴的な語彙を特定したい。Bと比較してAで有意に高頻度な語彙をAに特徴的な語彙とし，その抽出の指標として対数尤度比（log-likelihood ratio：LLR）を用いる。LLRは，英語学におけるコーパス研究でも，分野の特徴語を抽出する指標として用いられている。単語Wの LLRは以下の式で算出する。

$$LLR=2(a\ln a+b\ln b+c\ln c+d\ln d -(a+b)\ln(a+b)-(a+c)\ln(a+c)-(b+d)\ln(b+d) -(c+d)\ln(c+d)+(a+b+c+d)\ln(a+b+c+d))$$

a：Aでの単語Wの頻度　b：Bでの単語Wの頻度

c：Aの延べ語数−a　d：Bの延べ語数−b

ただし，単語WのAでの使用率がBでの使用率より低い場合，×(-1)の補正を行う。

LLR＞0でLLRの高い語ほどAにおいて特徴的な語と見なされる。ここでは，LLR≧10という基準を立て，これを満たす語を医療の専門語として抽出した。約24,000語となった。

## 4．専門度による語彙のレベル分け

### 4.1 二つの専門度

2で述べた，（2）難解語と重要語を抽出する作業のうち，はじめに，難解語について考える。

難解語を直接，コーパスから抽出することは困難であるので，難解語に近似する語彙を取り出す手法を工夫する必要がある。一般の人にとって難解な語彙は，次の二つの場合が考えられる。①専門分野でよく使われ，一般分野ではあまり使われないもの，②専門家の間でよく使われ，一般人が接触する機会があまりないもの。この二つはそれぞれ，①分野から見た専門度が高い語彙，②読者から見た専門度が高い語彙，ということになろう。

### 4.2 分野の専門度

まず，分野の専門度は，A医療コーパスの語彙と，B一般コーパスの語彙を比較し，Aに特徴的である語ほど，分野の専門度が高いと見ることができる。

表1　分野の専門度

| レベル | 語数 | LLRの区間 |
|---|---|---|
| 1 | 5,490 | 10.0 ～ 12.9 |
| 2 | 4,710 | 12.9 ～ 18.1 |
| 3 | 4,711 | 18.1 ～ 29.0 |
| 4 | 4.711 | 29.0 ～ 61.4 |
| 5 | 4,713 | 61.4 ～ |
| 計 | 24,335 | |

表1は，3で抽出した，医療の専門語，約24,000語について，LLR値の大小によって序列化し，語数がほぼ均等になるように，五つに区切ったものである。レベルが1から5へと進むにつれて，専門度が高くなる。

### 4.3 読者の専門度

次に，読者の専門度は，想定する読者によって医療コーパスを二類別した，a専門家読者のコーパスと，b一般人読者のコーパスの語彙を比較することによって測ることができる。やはり，この二つのコーパスの語彙頻度を比較してLLRを算出し，表2のようにレベルに分けた。

表2 読者の専門度のレベル

| レベル | 語数 | LLRの区間 |
|---|---|---|
| 1 | 5,102 | 〜 -12.1 |
| 2 | 4,808 | -12.1 〜 -0.3 |
| 3 | 4,969 | -0.3 〜 4.4 |
| 4 | 4,740 | 4.4 〜 14.0 |
| 5 | 4,716 | 14.0 〜 |
| 計 | 24,335 | |

医療の専門語として抽出した約24,000語の全体を対象としたもので，LLR値が負の場合も含んでいる。レベルが1から5へと進むにつれて，読者の専門度が高くなる。

### 4.4 二つの専門度の組み合わせ

表3 専門度のレベル

| 専門度 | 語数 |
|---|---|
| 2 | 489 |
| 3 | 2,170 |
| 4 | 3,433 |
| 5 | 5,129 |
| 6 | 5,216 |
| 7 | 2,518 |
| 8 | 2,139 |
| 9 | 1,692 |
| 10 | 1,549 |
| 計 | 24,335 |

以上の手続きによって，医療の専門語について，分野の専門度のレベルと，読者の専門度のレベルが与えられた。各語における二種のレベルの数値の和を，「専門度のレベル」として扱えば，2〜10 の九つのレベルに分けられる。各レベルに属する語数は，表3の通りである。

## 5．必要度による語彙のレベル分け

今度は，一般人にとっての重要語を取り出す方法を考えたい。コーパスから直接，難解語を抽出することが難しいように，重要語を抽出することも難しい。近似的な方法として，一般人が医療情報に接する媒体によく出てくる語彙を抽出することが考えられる。そうした媒体によく使われている語彙は，一般人にとっても，理解する必要に迫られる場合が多いと考えられるからである。

b一般人読者の医療コーパスにおいてよく使われている語彙を抽出するために，「よく使われている」ということを，使われる回数の多さ（度数）と，使われる場面の広さ（記事数）との二つでとらえる。度数だけでは，特定の場面に偏って使われる語が過剰に評価されてしまうおそれがあり，記事数だけでは，繰り返し使われる重要な語が評価できないおそれがあるからである。ここでいう，「記事」とは，雑誌や新聞の記事，ウェブでは，一つのページに表示される，情報のひとまとまりを指す。

表4 度数から見た必要度のレベル

| レベル | 語数 | 度数の区間 |
|---|---|---|
| 1 | 17,865 | 1 |
| 2 | 22,573 | 2 |
| 3 | 20,865 | 3 〜 5 |
| 4 | 14,916 | 6 〜 16 |
| 5 | 13,465 | 17 〜 |
| 計 | 89,684 | |

表5 記事数から見た必要度のレベル

| レベル | 語数 | 度数の区間 |
|---|---|---|
| 1 | 30,028 | 1 |
| 2 | 19,785 | 2 |
| 3 | 13,816 | 3 〜 4 |
| 4 | 13,502 | 5 〜 12 |
| 5 | 12,553 | 13 〜 |
| 計 | 89,684 | |

表6 必要度のレベル

| 必要度 | 語数 |
|---|---|
| 2 | 17,865 |
| 3 | 8,092 |
| 4 | 17,812 |
| 5 | 5,122 |
| 6 | 12,312 |
| 7 | 3,905 |
| 8 | 10,625 |
| 9 | 2,167 |
| 10 | 11,784 |
| 計 | 89,684 |

A医療コーパスに2回以上使われ，b一般読者のコーパスに1回以上使われる語彙を集計した。表4は，必要度を度数によって五つのレベルに分けた結果，表5は同じく記事数によって分けた結果である。二種のレベルの数値の和を「必要度のレベル」と扱って，各レベルに所属する語数をまとめたのが，表6である。

## 6．他の抽出法との比較

### 6.1 医療用語集掲載語彙との比較

4と5で，コーパスによって抽出した専門度と必要度のレベルが高い語彙は，本当に，医療分野における難解語・重要語になっているのだろうか。このことを検証するために，一般向け医療用語集での掲載状況，医師を対

象とするアンケート調査の結果と比較したい。

医療用語を一般向けに解説した辞書形式の用語集は，書籍やウェブによって，様々なものが公開されている。そのうち，収録語彙が 100 語以上で，医療の広い分野を扱い，内容が良質なもの，書籍版 7 種・ウェブ版 22 種を対象に，見出し語の調査を行った。異なりで約 17,000 語，延べで約 25,000 語が得られた。

コーパスから得られた専門度，必要度それぞれのレベル別の語彙について，上記の用語集に掲載されている語の比率（掲載率）と，掲載されている語の場合，平均していくつの用語集に出ているのか（平均掲載数）をまとめると，表７・表８になる。

表７専門度と医療用語集掲載状況

| レベル | 掲載率 | 平均掲載数 |
|---|---|---|
| 2 | 4.9% | 1.92 |
| 3 | 7.1% | 1.69 |
| 4 | 11.1% | 1.73 |
| 5 | 10.9% | 1.91 |
| 6 | 24.0% | 2.85 |
| 7 | 20.1% | 2.31 |
| 8 | 18.6% | 2.26 |
| 9 | 19.5% | 2.08 |
| 10 | 30.6% | 2.66 |

表８必要度と医療用語集掲載状況

| レベル | 掲載率 | 平均掲載数 |
|---|---|---|
| 2 | 3.4% | 1.53 |
| 3 | 2.5% | 1.49 |
| 4 | 2.9% | 1.46 |
| 5 | 5.7% | 1.53 |
| 6 | 4.7% | 1.78 |
| 7 | 8.6% | 1.91 |
| 8 | 8.2% | 2.14 |
| 9 | 20.0% | 2.30 |
| 10 | 15.0% | 2.85 |

全体としてレベルが高くなるほど，掲載率・平均掲載数ともに，数値が高くなっていく傾向が認められよう。特に，必要度においてその傾向が著しい。なお，専門度の場合，レベル６において，平均掲載数で最も高く，掲載率で２番目に高くなっていることが目を引くが，その理由は，今のところ不明である。

### 6.2. 医師に対するアンケート調査との比較

医師 300 人に対して，患者とのコミュニケーションで問題を感じたことがある語を挙げてもらう調査を行った。こうした問題語は，難解語や重要語と重なる性質があると考えられる。この調査では，異なり語数で 806 語，延べ回答件数で 1,510 件の回答があった。

各レベル別の語彙について，調査において医師から回答された語の比率（回答率）と，回答された場合は平均して何人の医師から回答されたか（平均回答件数）をまとめたものが，表９・表 10 である。

表９・表 10 によれば，回答率については，全体としてレベルが高くなるほど，数値が高くなっていく傾向があり，専門度の場合，レベル６が最も高くなっている。これは，医療用語集と比較した場合（表７・表８）と同じである。一方，平均回答件数は，必ずしもそうした傾向は見られない。

表９　専門度と医師調査回答状況

| レベル | 回答率 | 平均回答件数 |
|---|---|---|
| 2 | 0.2% | 1.00 |
| 3 | 0.4% | 1.75 |
| 4 | 0.5% | 1.67 |
| 5 | 0.8% | 1.44 |
| 6 | 4.2% | 2.42 |
| 7 | 3.1% | 2.05 |
| 8 | 2.0% | 1.72 |
| 9 | 2.1% | 3.60 |
| 10 | 4.1% | 2.06 |

表 10　必要度と医師調査回答状況

| レベル | 回答率 | 平均回答件数 |
|---|---|---|
| 2 | 0.2% | 1.22 |
| 3 | 0.1% | 1.14 |
| 4 | 0.1% | 1.53 |
| 5 | 0.4% | 1.00 |
| 6 | 0.3% | 1.28 |
| 7 | 0.4% | 2.00 |
| 8 | 0.7% | 1.55 |
| 9 | 1.9% | 1.64 |
| 10 | 3.1% | 2.62 |

## ７.コーパスによる抽出法の利点と課題

コーパスに基づいて判定した専門度・必要度のレベルは，レベルが高くなるほど，医療用語集によく取り上げられ，医師が問題に感じる場合も多くなる傾向があることが，確かめられた。この傾向が認められない指標もあるので，なお検証を要するが，おおむね，コーパスに基づいたレベル分けによって，難解語や重要語を抽出する目的を実現できると考えてよいだろう。

医療用語集や医師調査による抽出方法と比較した，コーパスによる抽出方法の利点としては，一度に語彙全体を対象に抽出が行えることと，レベル分けが行えることが挙げられる。医師対象の調査では，この二つのことは，ともに困難であり，医療用語集による場合も，レベル分けは多くを望めない。

一方，コーパスによる抽出方法には課題も多い。例えば，高レベルと認定された語彙の中には，一見して，難解でも重要でもないのではないかと思われる語が混在している。コーパスでの頻度のありようが，必ずしも，難解度や重要度を反映するわけではないところもあるのだと予想される。頻度のありようと，語彙の性質との関係は，特に深く追究すべき課題である。

**付記**　本研究は，文部科学省科学研究費補助金特定領域研究「日本語コーパス」（領域代表：前川喜久雄）における計画研究（言語政策班），「言語政策に役立つ，コーパスを用いた語彙表・漢字表の作成と活用」（代表：田中牧郎），および国立国語研究所研究開発部門言語問題グループの研究課題「病院の言葉を分かりやすくする提案のための資料作成」の成果の一部である。また，短単位解析の結果を組み上げて複合語を一つにまとめる方法は，石井正彦氏（大阪大学大学院）から示唆を得た。

**連絡先**　田中牧郎　〒190-8561　立川市緑町 10-2
国立国語研究所研究開発部門　mtanaka@kokken.go.jp