医学用語の選択に見られる特徴

金子 周司（京都大学大学院薬学研究科）†

Characteristics of the Choice of Japanese Medical Words in the Corpora of Scientific and Clinical Documents

Shuji Kaneko (Kyoto University Graduate School of Pharmaceutical Sciences)

１．はじめに

医療や生命科学の急激な進歩は、莫大な数の専門用語を新たに生み出している。筆者は医学系学生や研究者が英語の専門用語を学習・活用するための電子辞書の開発に 20 年来取り組んできたが、日本語については訳語として位置づけ、あまりその特性について深く考察してこなかった（金子 2006）。しかし今後、医療や教育の電子化がますます進展し、自然言語処理が医療サポートや知識発見に応用されていくことを考えると、医学用語の日本語表記について理解を深めることが必要と思われる（金子、大武 2010）。

本研究ではどのようにして和文で医学用語が選択されているかを少しでも知るために、医学文献や医薬品解説書を元にしたコーパスを構築し、専門用語を抽出した上で異表記を収集、解析した。我が国で医学用語をどのように表記するかについては、決まり事や法則があるわけではなく、研究者が独自に編み出したり、すでにある文書を参考に用語が選ばれたりしている。コーパスの解析結果からも、日本語では漢字、カタカナ、ひらがな、英語綴りなどを混在して用いることができるため、異表記が非常に多いという特徴があることがわかってきた。編集者や許認可者による修正を経た後の文書においても、医学用語の多様性は維持されている。いくつかの例を紹介して考察してみたい。

２．コーパスの概要

筆者は、出版社である株式会社羊土社の協力を得て、医学研究者が書く総説の全文コーパスを以前に構築した（金子 2006）。本研究ではそれをさらに拡張し、1996 年から 2005 年の 10 年間にわたって『実験医学』誌に発表された全総説をテキスト化することで実験医学コーパス（37.3Mbyte）とした。本コーパスについては、用語の解析目的でのみ使用できる許諾契約を締結した。また、財団法人日本医薬情報センター（通称 JAPIC）が有料で販売している医療用医薬品全 13,000 種の添付文書情報（2008 年版）について、解析目的での使用許諾を得てテキスト化し JAPIC コーパス（49.6Mbyte）とした。

解析としては、ライフサイエンス辞書に収録している 157,347 語の日本語をクエリーとして、コーパス中で一致する文字列の頻度を Perl スクリプトにより求めた。これら 2 種類のコーパスの概要を表 1 に示す。JAPIC コーパスのほうがサイズ的には大きいが医薬品固有名が多いこともあり、以下においては同規模のコーパスとして頻度（語数）の比較を行う。

表 1　本研究で用いたコーパスの概要

	実験医学	JAPIC
文字数	20,235,504	25,247,795
読点数	271,158	418,839
句点数	504,847	687,669
「など」頻度	22,897	26,997
「血管」頻度	13,146	12,345
「高い」頻度	4,265	4,087

† skaneko@pharm.kyoto-u.ac.jp

３．各コーパスの特徴

表２はそれぞれのコーパスで求めた頻度のうち、一方での値が他方の 100 倍以上であった特徴語を示している。

実験医学コーパスにおいては、最先端の成果を研究者自らが執筆していることもあり、「遺伝子」「タンパク質」「配列」といった生体分子の名称や物性を表す語が多く、「シグナル」や「ドメイン」のように専門家の間でのみ通用する jargon とも考えられるカタカナ語が多用されている点が特徴的である。

一方、JAPIC コーパスで最も頻度が高いのは「本剤」であるが、これは医薬品添付文書における主語として多用されるためである。その他にも「経口投与」「血中濃度」など添付文書における解説として専門家に注目されるべき特徴語が見られる。医薬品は同一作用機序をもつ類似薬が多いこともあり、それらの添付文書間では記述も似ている傾向がある。このことは実験医学コーパスからは 50,257 種類の語が抽出されたのに対して、JAPIC コーパスからは 36,449 語しか抽出されなかった解析結果に反映されている。

図１には各コーパスを構成している文字種の割合を示した。いずれのコーパスにおいても英数字とカタカナが 3〜4 割、漢字の割合も 3〜4 割を占めており、きわめて専門用語に満ちた文書であることがわかる。

表２　コーパスの特徴語（頻度データ）

語	実験医学	JAPIC
遺伝子	45, 676	294
タンパク質	31, 544	53
シグナル	18, 755	14
ドメイン	14, 474	7
配列	10, 640	16
本剤	35	99, 945
経口投与	90	21, 550
血中濃度	135	19, 973
既往症	2	14, 879
妊婦	11	14, 007

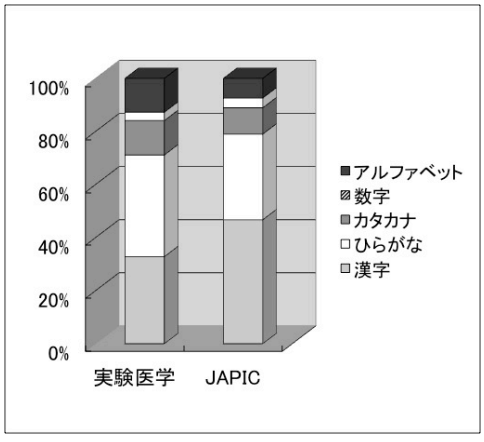

図１　コーパスを構成する文字種

４．用語の選択

以上のように構築したコーパスを用いて専門用語の頻度を見ていくと、いくつかの問題点に気づく。それらをデータと共に説明していく。

４.１.「protein」使い分けの実情

Protein とはアミノ酸がペプチド結合によって連なり、そのアミノ酸の性質や場の状況によって特異的な立体構造をとる最も重要な生体構成分子種である。英語においては protein という語の他には比較的短い鎖を指す peptide（ペプチド）という語があるが、protein に異表記は存在しない。しかしながら日本語においては protein が卵白に多く含まれることに起源をもつ「蛋白」から「蛋白質」という語を生み出し、日本医学会は用語集で「蛋白質」を推奨している。しかし文部科学省は学術用語として「タンパク質」という表記を標準としており、新聞や報道等においては「たんぱく質」という表記が多く採用されている。それぞれから「質」を除去した表記も多く用いられ、さらには「プロテイン」と表記すれば一般社会においてはサプリメントとして用いられる補助栄養食品を指すかのように微妙に使い分けられている。

ひとつの英語に対して複数のカタカナ外来語が生じることは、例えば vector が数学の世界では「ベクトル」、分子生物学や情報学では「ベクター」と書かれるように、ドイツ語由来と英語の発音に近い複数のカタカナ語が時代を異にして生じるため珍しいことではない。しかし医学用語の場合には、ひらがなや漢字での異表記まで加わって非常に多様になっている。今回、構築したコーパスにおいて調べてみた結果を表３に示すが、それぞれ編集者や許認可者の手が入った文書であるにもかかわらず、多様な表記が検出された。

表３　「protein」の日本語表記における選択

語	実験医学	JAPIC
たんぱく質	0	14
タンパク質	***31, 544***	53
タンパク	4, 330	174
蛋白質	2, 693	489
蛋白	1, 067	***4, 337***
プロテイン	801	77

実験医学コーパスで「タンパク質」が多く、JAPIC コーパスで「蛋白」が多いのは基礎医学と医療という分野間の差異であると考えられるが、より詳細に見ていくと用語はさらに精密に選択されている（表４）。

表４　「protein」の連接語に応じた選択

コーパス	連接語	タンパク質	タンパク	蛋白質	蛋白	プロテイン
実験医学	結合～	***1, 327***	199	139	29	0
	プリオン～	***53***	9	2	***54***	0
	～分解	***676***	128	18	11	0
	～キナーゼ	75	42	3	0	***489***
JAPIC	～結合	0	43	0	***1, 693***	0
	糖～	2	20	***114***	76	0

実験医学コーパスにおいて、前に「結合」や後に「分解」が連接する場合はいずれも「タンパク質」が多く用いられていた。「プリオン」との連接においては「プリオン蛋白」という表記が特異的に高い傾向が見られた。このことはカタカナ同士が連接した場合に元の語の境界が分かりづらくなることを避けている表現なのかもしれない。しかし、タンパク質をリン酸化する酵素である protein kinase を表す際には、そのままカタカナ語として「プロテインキナーゼ」が最も頻出した。

JAPIC コーパスにおいては、「蛋白結合」のように「蛋白」という表記が全般的に好んで

用いられていたが、この用語はいずれの省庁や団体も推奨している表記ではない。一部においては「糖蛋白質」のように「質」をつけた表記が集中しているケースも見られたが、これは他の類似薬で用いられた文書をそのまま流用して使っているために複製増幅効果が現れたものと推察される。

４.２．薬物のカテゴリーを表す名称

先行研究において筆者は、英語圏で発達した医学や生命科学が日本へ「輸入」された際に必ずしも専門用語を直訳するのではなく、日本人が理解しやすいように「意訳」を行ってきた実例をいくつか提示した（金子 2006)。この結果は PubMed で公開されている英語の医学文献抄録と実験医学コーパスの前身である日本語テキストを比較解析して得られたものであったが、今回、JAPIC コーパスを新たに解析することによって、これらの指摘が準公的な医薬品添付文書においても適用できることが明らかになってきた。

その一例として、表５は腫瘍の増殖に対して抑制的に作用するカテゴリーの薬物に与えられる一般的な名称を調査した結果である。この結果から、いずれのコーパスにおいても多様な表記が混在していることがわかる。専門的には「癌≠悪性腫瘍」であり「癌＝上皮細胞の（つまり一部の）悪性腫瘍」であることを加味すると、このように階層の異なる概念を同一視している現状は好ましいとは言えない。

表５　腫瘍増殖を抑制する薬物の名称

語	実験医学	JAPIC
抗癌薬	19	0
抗癌剤	***763***	32
抗がん薬	0	9
抗がん剤	3	6
制癌剤	37	3
抗腫瘍薬	13	0
抗腫瘍剤	12	52
抗悪性腫瘍薬	10	13
抗悪性腫瘍剤	1	***741***
悪性腫瘍治療薬	1	0

５．まとめ

医学用語は長らく標準化の方向性で議論されていた。しかしながら、本研究で編集者や許認可者の修正を経た文書コーパスを解析した結果、コントロールされた状況においても医学用語の多様性は失われていないことが明らかになった。実際に現場で作成される文書（例えば電子カルテや学会抄録など）はさらに多様で混沌としているであろうことは容易に想像できる。今後、医療文書の電子化などによって情報の利活用を目指す場合、このように多様な異表記に耐えうる（かつ英語表記や略記にも対応した）頑強なシソーラスを早急に整備することが必要と思われる。

文　献

金子周司 (2006)「ライフサイエンス辞書とは」情報管理, 49:1, pp.24-35.
金子周司、大武 博 (2010)「ライフサイエンス辞書からクリニカルインフォマティクスへ」情報管理, 53:9, pp.473-479.

関連 URL

ライフサイエンス辞書プロジェクト　http://lsd.pharm.kyoto-u.ac.jp/