文献情報の解析に基づく対訳シソーラスの評価

金子周司1）藤田信之2）

生命科学および医学の教育研究を支援する目的で，筆者らは広範な生命科学の諸領域で使われる英語および日本語の専門用語を文献コーパスの定量的解析から抽出し，独自の対訳辞書をライフサイエンス辞書（LSD）として公開してきた。今回，テキストマイニング等に応用できるシソーラスへの発展を目標に，LSDで対訳と意味情報を定義した5万語の英語と5万語の日本語について文献情報による評価を行い，続いて，MeSHツリーとのマッピングによる体系化を試みた。主にPubMed抄録からなる英語コーパスをLSD収録語で解読できる割合は80%であった。MeSHツリーから得られた6.5万語のうち，LSDと一致したのは20%であったが，PubMed中に出現するMeSH termについては40%がカバーされていた。MeSHにないLSD収録語としては略語，名詞以外の品詞，未分類の学問分野の用語などが浮かび上がった。今回の解析から，我が国で今後，医療現場等で発生する大量のテキストをコンピュータで解析するためには新たな対訳シソーラスの必要性が示唆された。

キーワード：シソーラス，オントロジー，電子辞書，ライフサイエンス，用語法

Evaluation of an English-Japanese thesaurus based on the analysis of biomedical corpora : KANEKO Shuji1), FUJITA Nobuyuki2)

Life Science Dictionary (LSD) is a versatile database of English and Japanese terms based on the quantitative analyses of biomedical corpora. To develop a thesaurus of LSD terms for future application to computer-assisted text mining, we have evaluated the frequency of LSD terms in the literature-based corpora, and mapped the LSD terms to the MeSH tree. Coverage of LSD English terms in a PubMed-based corpus was 80%. In 65,000 MeSH tree terms, LSD-matched terms were 20%, which was increased to 40% in a subpopulation of terms occurred in the English corpus. The MeSH-unmatched LSD terms included abbreviations, verbs, adjectives, adverbs and MeSH-unclassified terms. These results indicate the requirement of new comprehensive thesaurus tree covering complex English-Japanese translations.

Keywords: thesaurus, ontology, electric dictionary, life sciences, terminology

1. はじめに

ライフサイエンス辞書（Life Science Dictionary，以下LSD）は医学を包含する生命科学の教育研究を支援する目的で，様々な学問領域の研究者が協力しあい，1993年以来，制作されてきた電子対訳辞書である1）。LSDの特長は，学術論文の計量的な解析を行って作成したデータに基づき，頻度の高い用語を網羅的に収録している点や，音声や共起表現の提示など，学習者にとって有用な機能を実装している点などが挙げられる2）3）。

一方，生命科学の研究が加速し，テキスト情報量が指数関数的に増大している現状において，すべての情報を人間が解読する努力はすでに限界を超えている。また，ゲノムやタンパク質の配列情報が正規化データベースとして蓄積され，コンピュータによって情報の解読や推測が可能となっている現状とはきわめて対照的に，医学研究報告に記述された人智は有効利用されないまま蓄積され続けている。さらに，インターネットの普及によって一般化した検索エンジンによる情報検索において，無駄な情報を排除して求める情報を提示するためには，分野に特化したセマンティックウェッブ辞書の開発が求められる。

我々は将来LSDを英語と日本語を包括したテキストマイニング，機械翻訳，セマンティックウェッブ等に応用できる「コンピュータのための辞書」に発展させることを新たな研究目標としている4）5）。生命科学領域における対訳シソーラスとしてはUMLSの日本語化6）7）などに先例がある が，実際

*1）京都大学大学院薬学研究科生体機能解析学分野
*2）製品評価技術基盤機構ゲノム解析部門
〒606-8501　京都市左京区吉田下阿達町
TEL：075-753-4541
FAX：075-753-4542
E-mail：skaneko@pharm.kyoto-u.ac.jp

に英語および日本語テキストの解析を行った結果に基づき，対訳における多様性と各々の言語の特徴にまで踏み込んで対訳シソーラスを構築した例はまだない。そこで本研究では，まずLSDに収録された5万語の英語と5万語の日本語が，実際に最近の論文や総説で用いられている専門用語を網羅する割合を評価し，到達目標を明確にしようとした。次に，これら英語と日本語の関係を定義している約7万対の対訳を体系化するため，PubMed検索への応用性を考慮して，代表的な既存シソーラスとしてMeSH (Medical Subject Headings)との照合を行い，その結果から実用性の高い対訳シソーラスを構築するにあたっての問題点と方向性を考察した。

2. 方法

2.1 LSD用語

2006年1月にWebLSDとして公開した辞書1）の元となったFileMaker Proのリレーショナルデータベースより，英語49,034語（対訳定義ずみ），日本語73,103語（うち，対訳の定義されたもの48,024語），およびそれらを結合する中間テーブルから70,622対の英語および日本語を出力して用いた。なお，LSDにおいて英語とは英単語と語句を含んでおり，規則変化を伴う名詞，形容詞，動詞ではそれぞれ単数形，原級，現在形が，ラテン語由来の名詞は単数形と複数形が別個のレコードとして収録されており，不規則変化を伴う動詞は現在形のみを収録している。また，LSD対訳には従来より意味情報および品詞が付与されているので，それらをMeSHに合わせて「A:解剖」「B:生物」「C:疾患」「D:薬物」「E:技術」等に再分類した。

2.2 コーパスの作成

英語については，PubMedに抄録が収録されている学術誌のうち，インパクトファクターなどを考慮して生命科学の各分野から選んだ代表的な学術誌（89種類）にアメリカおよびイギリス国内の研究機関から1995年から2004年に報告された論文抄録テキスト（年間23000ー24000抄録，368 MB）を蓄積した。原著論文だけでは教科書に記載されるような基本的理解に必要な語彙が不足するため，NCBI Bookshelfで公開されている教科書や，協力を得られた出版社から提供された電子テキストなどを加えて，463 MB（約6千万単語）の英語コーパスを作成した。

一方，日本語については，基礎医学・ゲノム科学の最新研究成果に関する総説誌を発行している出版社の協力を得て，1996年から2002年にかけて出版された総説誌原版からタイトルを含む本文テキスト26 MBを抽出した。これに臨床医学の教科書テキストを合わせて34 MB（約2千万文字）の日本語コーパスを作成した。

2.3 解析プログラム

英語コーパスからの英単語の抽出と出現頻度解析には，単語間のスペースを認識して切断し，単語毎に計数するPerlスクリプトを作成した。この結果をFileMaker Proに読み込み，LSDデータベースに対して同一見出し語間でリレーションを設けることによって語尾変化を考慮しないLSD収録語とのマッチングを行った（図１）。続いて，ラテン語由来の名詞の不規則変化には対応しないが名詞複数形の規則変化や規則動詞および不規則動詞の変化形に対応する公開逐語訳エンジンEtoJ8）を改変し，LSDに収録された英単語および英語句の英語コーパス中での出現頻度解析を語尾変化を含めて行った。

日本語の抽出および出現頻度解析には，漢字，カタカナ，ひらがな，アルファベットおよび数字の境目を認識して最長連続する要素（仮に単語と呼ぶ）を抽出するPerlスクリプトを作成し，コーパスにおけるそれぞれの文字種の割合とユニークな単語の種類を計数した。また，日本語コーパス中で日本語の単語が出現する頻度の計数もPerlスクリプトを作成して行った。

PubMed単語	LSD2006対訳	総頻度
neuropathic	神経障害性	309
neuropathies		70
neuropathogenesis	神経病因性	41
neuropathogenic		26
neuropathogenicity		11
neuropathol		1
neuropathologic	神経病理的	56
neuropathological	神経病理学的	201
neuropathologically	神経病理学的に	22
neuropathologies		10
neuropathologist		1
neuropathologists		4
neuropathology	神経病理学	174
neuropathophysiology		3
neuropathy	神経障害	607
neuropathy-associated		1
neuropeptidase		9
neuropeptidases		2
neuropeptide	神経ペプチド	945

図1 英語コーパスの単語頻度解析とLSDとのマッチング

2.4 MeSH加工と対訳リンク

MeSH 2006 （XML 形 式 ） か らTree, Descriptor, Concept, Termテキストを抽出し，FileMaker Pro上でリレーショナルデータベースとして再構築した。MeSH Term表記の語順を自然な形に置き換え，規則的な名詞の語尾変化を吸収してLSDで用いている標準的な辞書表記と一致させるスクリプトを作成し，規則変化を伴うMeSH Term表記の変換を行った（図2）。MeSH Term 483,674語のうち，ツリーに帰属できたTermは76,836語であったが，語順変換などによって同一の文字列となった重複レコードを除外すると65,733語となった。一方，LSDに収録されている対訳レコード中 の英語について，FileMaker Pro上で表記の一致するMeSH Termへのリレーションを設定することによってLSDとMeSHツ

リーの照合を行った。なお，FileMaker Proのリレーションでは大文字/小文字は同一視される 。対訳テーブルあるいはMeSHテーブルでは重複する見出しが存在するため，それらの重複は後で除外した。

以上の関係を整理すると図3のようになる。7万対のLSD対訳は5万語の日本語と5万語の英語から参照されており，それらはコーパス解析結果である頻度情報へリンクしている。日本語の場合は対訳を有しないレコードを含む7万語の漢字変換テーブルも頻度解析に 用いた。

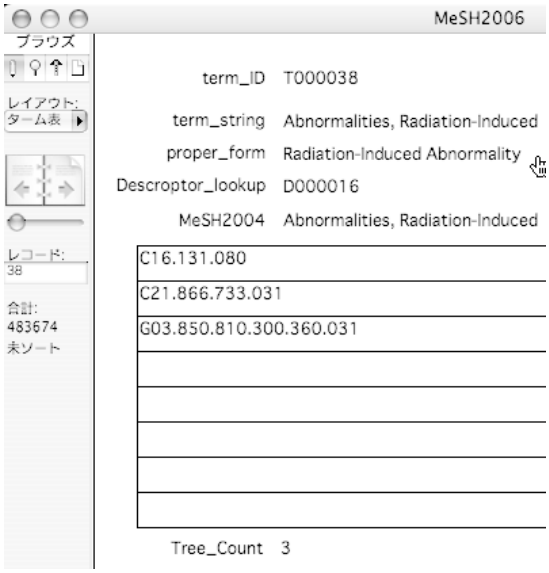

図2 変換されたMeSH2006

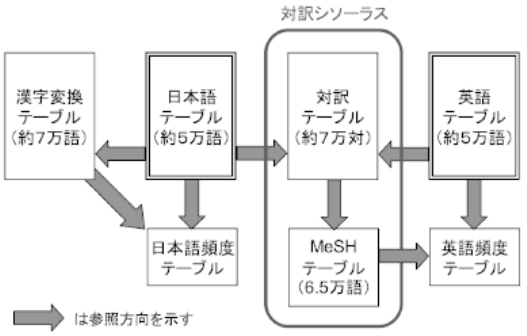

図3 LSDとMeSHのリレーション概略

3. 結果

3.1 英語コーパスを用いた頻度解析

まず，LSDに対訳が収録されている49,034語の英単語および語句について，語尾変化も含めて英語コーパスでの出現頻度を計数した（図4A）。全体の84%がコーパスに1回以上出現したが，6,714語（16%） はコーパスで出現しない用語であった。内容としては大学や機関名などの固有名詞，希少な動植物の学名，化学や物理学など周辺分野の語句が多く見られた。スペースを含む複合語を除いた31,218語の英単語に限ると，29,473語（94%）が英語コーパス中に出現していた。

次に，英語コーパスの全単語の頻度解析を行った。語尾変化を考慮せずにすべてのユニークな文字列を抽出する手法を用いると，英文コーパスから67万種類の単語が抽出された。このうち，数字と記号のみからなる5万語を除外した62万語について出現頻度ランクごとに単語の種類を数えると，両対数プロットで負の直線的相関性（Power law）が認められた（図4B）。回帰直線の傾きは-0.67（r2=0.99）となった。出現頻度が低くなるにつれてカバー率は低下したが， ほぼ半数を占める30万語はコーパスで１回しか出現しない記号やスペルミス等であった。

さらに，LSDで対訳を規定した英語が実際に研究者の書く英語のどの程度をカバーしているかを評価するために，英語コーパスのテキストをLSDで訳せるかどうかを規則的な語尾変化に対応する逐語訳で検討 したところ，1000回以上出現する英単語はほぼすべて変換された。英語コーパス全体に対してLSD収録単語で解読できるテキストの割合は88%と算出された。

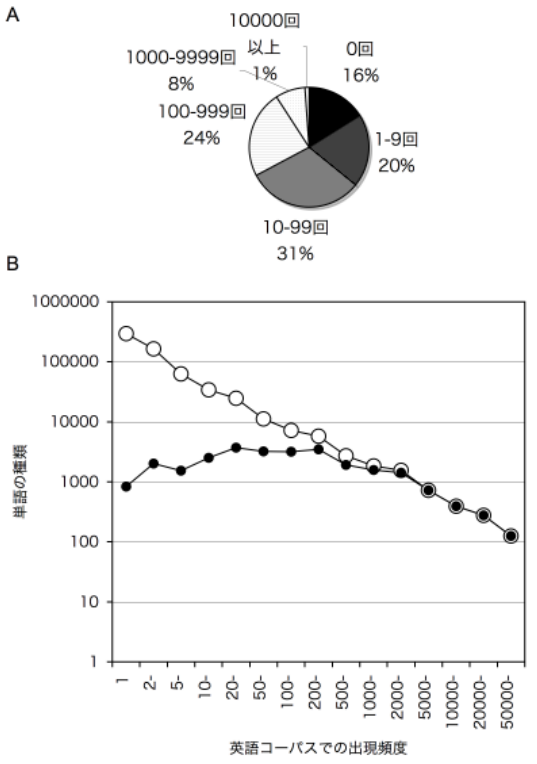

図4 英語コーパスを用いた頻度解析
A，LSD収録英語（49,034語）の頻度分布；B，英語コーパスの全単語（○）とLSD収録英単語（●）の出現頻度分布

3.2 日本語コーパスを用いた頻度解析

医学・生命科学領域の総説および教科書から作成した大規模な日本語コーパスを解析した例はこれまでにほとんど報告されていない。今回，まず日

本語コーパスの漢字，カタカナ，ひらがな，英数字という文字種の割合を解析したところ，漢字の割合が35%，英数字とカタカナで記述される割合が全テキストの28%にも及ぶことが明らかになった（図5A）。また，ひらがなのみから成る文字列を除去した上でコーパス中のユニークな単語の種類を数えたところ，全体で26万語が抽出されたが，20%に相当する5.4万語は英語ないし英数字記号であった。なお，LSDに対訳が収録されている48,024語の日本語について，日本語コーパスに1回以上出現している単語を数えると44,286語（92%）であった。コーパスに出現しない単語は化学，物理学，数学などの周辺科学領域に多く見られた。

次に，コーパスに出現する漢字およびカタカナを含む21万語について出現頻度ランクごとに単語の種類を数えると，英語の場合と同様に両対数プロットで負の直線的相関性が認められ，回帰直線の傾きは-0.89（r2=0.99）であった（図5B）。また，LSDに収録されている日本語について，漢字変換テーブルに収録されている語と対訳を有する語のそれぞれについて頻度を解析すると，頻度5以上ではいずれもほぼすべて収録されていた。しかし，対訳を有する日本語は頻度100未満において，出現頻度が低くなるにつれて収録率は低下していた（図5B）。単純なテキストマッチングによって日本語コーパス中の漢字・カタカナ語に対してLSD収録語がカバーする割合は80%，さらに対訳が規定されている日本語に限定した場合は 66%と算出された。

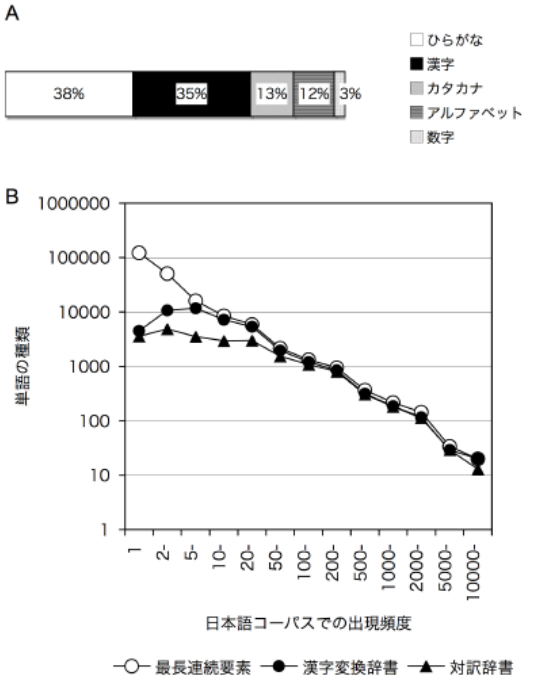

図5 日本語コーパスを用いた頻度解析

A，コーパスを校正する文字種の割合；B，全単語（○）とLSD収録語（●全日本語，▲対訳あり）の頻度分布

3.3 MeSHツリーへのマッピング

LSD対訳テーブルで70,622対の訳語を定義している49,034語の英語に対して， 自然な表記に変換したMeSH termの照合を行ったところ（図6），ツリーに帰属できる65,733語のMeSH TermのうちLSDと一致した共通語は13,462語（20%）であった（表1最下段）。MeSHのうち，英語コーパスに1回以上出現した33,195語に限定すると，共通語は12,971語（40%）となった。MeSH収録語でLSD未収録の用語としては，化合物名や生物学名などの事物を特定する固有の名称 が多く見られた。一方，LSD対訳のうちMeSHと照合できない英語は 35,572語であり，その中にはMeSHカテゴリーに掲載されていない一般的な名詞（例：increase，cooperation）や形容詞や動詞といった用言が特徴的に見られた。

英語表記	日本語コード	日本語表記	MeSHリンク		
acute aortic dissection	J041788 (n)	解離性大動脈瘤	t p		
acute B-cell leukemia	J051255 (n)	急性B細胞白血病	t T045626 p D015450	B-Cell Leukemia, Acute B-Cell Leukemia, Acute	C04.557.337.428.500.1
acute brain injury	J053144 (n*)	急性脳損傷	t T005485 p D001930	Acute Brain Injuries Acute Brain Injuries	C10.228.140.199
acute cholecystitis	J041369 (n)	急性胆嚢炎	t T521696 p	Acute Cholecystitis	
acute coronary occlusion	J043680 (n)	急性冠閉塞	t p		
acute coronary syndrome	J041222 (n)	急性冠動脈疾患	t p		
acute coronary syndrome	J052884 (n)	急性冠状動脈疾患	t p		
acute disease	J041283 (n*)	急性疾患	t T000603 p D000208	Acute Disease Acute Disease	C23.550.291.125
acute disease	J051557 (n*)	急病	t T000603 p D000208	Acute Disease Acute Disease	C23.550.291.125
acute exacerbation	J041070 (n)	急性増悪	t p		
acute hemorrhagic conjunctivitis	J071103 (n)	急性出血性結膜炎	t T008421 p D003232	Conjunctivitis, Acute Conjunctivitis, Acute	C02.325.250.250
acute hemorrhagic leukoencephalitis	J053369 (n)	急性出血性白質脳炎	t T014357 p D004684	Leukoencephalitis, Acute Leukoencephalitis, Acute	C10.114.375.362
acute hepatic failure	J046828 (n)	急性肝不全	t T061087 p D017114	Hepatic Failure, Acute Hepatic Failure, Acute	C06.552.308.500.750
acute hepatitis	J032500 (n)	急性肝炎	t p		
acute infantile hemiplegia	J031354 (n)	急性乳児片麻痺	t p		
acute infection disease	J032501 (n*)	急性伝染病	t p		
acute inflammatory demyelinating	J072348 (n)	急性炎症性脱髄性多発ニューロパチー	t T385254 p D011128	Polyradiculoneuropathy Polyradiculoneuropathy	C10.114.750
acute intermittent porphyria	J071727 (n)	急性間欠性ポルフィリン症	t T061121 p D017118	Porphyria, Acute Intermittent Porphyria, Acute Intermittent	C06.552.630.150
acute leukemia	J044723 (n)	急性白血病	t p		

図6 LSD対訳とMeSHリンク例

次にMeSHカテゴリーおよびLSD意味分類別に分けて詳細を検討した（表1）。解剖学用語（カテゴリーA）はLSD収録数がMeSHを超えており，MeSHの62%はLSDに収録されていた。LSD収録語が多いのは，主として発生学用語が多いためであった。生物学名（カテゴリーB），病名（カテゴリーC），化合物名（カテゴリーD）について，MeSHのLSD一致率はそれぞれ22%，31%，13%と低かったが，英語コーパスに出現する語に限定すると，その割合は37%，57%，29%に高まった。生物学名および病名ではLSD収録語のうちでMeSH未収録の割合は比較的低かったが，化合物名では日本国内でのみ用いられる薬物や生薬などでMeSH未収録語の割合が高かった。技術・装置（カテゴリーE）以下の分類においては，LSDとMeSHの一致率はどちらから見ても低く，LSDでは現象や性質を表現する用語が多いのに対して，MeSHでは事物の名称が多く収録されている対照的な違いが見られた。すべてのカテゴリーについて略語を抽出すると，LSDに収録された1895語のうち，MeSHに収録されているのは325語のみであった。

また，表2で示すように，MeSHとLSD収録語につい

て平均単語長と平均文字バイト数を比較すると，MeSHには複合語が多くLSDには単語が多い差異が明らかであり ，共通語ではそれらの中間的な値を示した。英語コーパスでの出現頻度はMeSHよりLSDで顕著に高かった。さらに，75%パーセンタイル値では共通語が最も高い値を示した。

表1 LSDとMeSH収録語数と分類別比較

分類 【MeSHカテゴリー】	LSD英語 [A] <対訳数>	MeSH* [B]	共通語* [C] = [A]∩[B]	カバー率* [D] = [C]/[B]	未マップ語 [E] = [A]-[C]
Anatomy【A】	4,102 <4,970>	2,576 (2,024)	1,604 (1,528)	62% (75%)	2,498
Organisms【B】	2,505 <3,101>	5,635 (3,254)	1,215 (1,193)	22% (37%)	1,290
Diseases【C】	5,403 <7,400>	12,095 (6,338)	3,700 (3,604)	31% (57%)	1,703
Chemicals & Drugs【D】	8,001 <9,806>	32,259 (13,835)	4,211 (4,015)	13% (29%)	3,790
Techniq. & Equip.【E】	3,099 <4,214>	4,900 (2,740)	1,013 (965)	21% (35%)	2,086
その他の名詞【F-Z】 および略語	14,648 <23,418>	8,268 (5,004)	1,719 (1,666)	21% (33%)	12,929
形容詞	7,984 <12,053>				7,984
動詞	2,144 <3,974>				2,144
副詞	1,148 <1,686>				1,148
合計	49,034 <70,622>	65,733 (33,195)	13,462 (12,971)	20% (40%)	35,572

*丸カッコ内は英語コーパスで頻度1以上の語について集計した値

表2 LSDとMeSH収録語の平均語句長と頻度

	LSD英語 [A]	MeSH [B]	共通語 [C] = [A]∩[B]
語数	49,034	65,733	13,462
平均単語長	1.51	2.43	1.70
平均文字バイト数	12.7	19.5	14.7
平均頻度 （75%パーセンタイル）	1,141 (257)	194 (16)	868 (328)

4. 考察

今回，LSDに対訳と意味情報 が定義された英語および日本語について，それぞれ十分な大きさを有する専門文書コーパスを制作して出現頻度およびテキスト網羅率を検討し，さらにMeSHとの照合により対訳シソーラス構築の第一歩を試みたことは，LSDの現状評価と将来の課題について，いくつかの重要な示唆を与える。

4.1 頻度解析による評価

まず，LSDは元来，PubMed等の英文コーパスの頻度解析結果に基づいて収録すべき用語を選択しており，今回，新たに最近10年間に発表されたPubMed抄録や教科書を用いて頻度解析を行った結果，84%の語句がコーパス中に出現し，コーパステキスト中の全単語の88%がLSDでヒットしたことによって，その方針の妥当性が裏付けられた。また，同様の解析を初めて日本語に対しても行ったが，LSD収録語の92%がコーパス中に出現したことは，LSDが生命科学の日本語を記述するために適切なボキャブラリーを有していることを意味している。しかし，単純なテキストマッチングでは日本語コーパスにある漢字・カタカナの66%のみがLSD収録語に一致する結果となった。このように英語に比べて網羅できる割合が低いのは「起こる」「明らかになる」等の漢字1文字とひらがなの組み合わせで構成される高頻度の用言が漢字1文字と解釈されて除外されたことによるものであり，今後は形態素解析プログラムにLSDを辞書として組み合わせることによって，日本語辞書の正しい評価を行う必要がある 。

4.2 それぞれに含まれる用語

今回，研究者が公表する論文を対象にしたテキストマイニングを想定し，その抄録データベースであるMEDLINE（あるいはPubMed）の統制語MeSHとLSDのマッピングを行うことによって，広範な語彙で記述される生命科学の学術用語を英語・日本語を含めたシソーラスツリーに当てはめることを試みた。しかしながら，実際にその共通集合として得られた語数は，LSDとMeSHのいずれから見ても20%台に留まった。一方，前述したようにLSDはPubMedを主体とする英語コーパスに対して88%の網羅性を有している。この差異は，LSDがコーパスで高頻度に出現する用語を収録していることで説明される。

LSDとMeSH，さらにPubMedにおいて特徴的に見られる語句の種類について，図7で概念的に示した。PubMedとの重なりはLSDがMeSHより高く，LSDとMeSHの共通語にはコーパスに 高頻度の語句が含まれている。LSDには動詞，形容詞，副詞といった用言や日本語独特の概念（例：生活習慣病や再生医療に対応する訳語）が固有に収録されている一方，MeSHは化合物名や生物学名を網羅的に収録している。これらのシソーラスに含まれないのは主として低頻度の用語であり，特に遺伝子名のような記号や略語は多種多様である。これらについて，以下でより詳細に考察する。

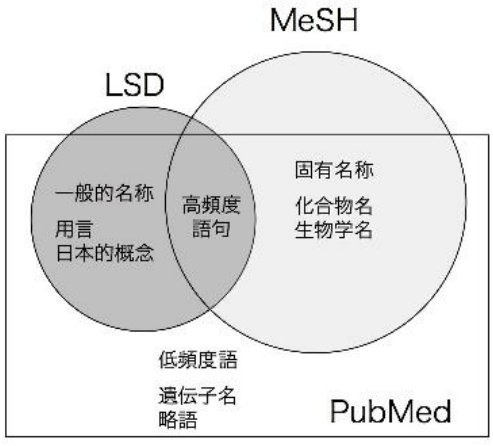

図7 LSD，MeSH，PubMed語彙の特徴

4.3 LSD未収録の英語

LSD未収録語には非常に多様かつ低頻度の語句が残されており，中でも記号とも言える遺伝子名については，対訳を定義することは不可能に近く，実際に日本語コーパスにおいてもそれらは英語のまま記述されていた 。同様に，化合物名や生物学名についても日本語でカタカナで表記される一般名（慣用名）を収録することは意味あるが，日本語でも英語のままで表記される化学名やラテン語学名をカタカナ表記で収録することはあまり意味がない（少なくともテキストマイニング辞書としては効果が低い）と考えられる。また，単なるカタカナへの変換は，日本語の視認性というメリットを低下させる要因でもあろう。むしろ日本語コーパスを用いた解析によって日本人が英語のまま記述しやすい用語を定量化かつ判別し，それらの用語については対訳シソーラスではなく英語で構築されたGene Ontology9）など他のシソーラスの応用を考えるのが現実的と考えられる。

4.4 多義性への対応

英語と日本語の対訳関係は1対1でない場合が多い。LSDでも複数の訳語を定義している語句が英和，和英いずれの方向でも3割程度存在しており，出現頻度の高い語句では一般的に見られる。例えば，dischargeは生理学では「発火」，内分泌では「放出」「分泌」，臨床では「退院」を意味するため，単語レベルで意味を一義的に定めることは困難である。このような多義性の問題について，LSDでは連接する単語を含めた複合語を収録することで対応する方針をとっていたが，今回，MeSHとの比較によって平均単語長と文字数のいずれも短く，単語を中心に収録している事実が明らかになったことで，現状のLSDでは意味を理解するための辞書としてはまだ不十分であることが指摘される。
略語も含めた多義性への対応に関しては，多くの推定手法が報告されている10）11）12）が，その精度を高めるためにも多義性を有する語句の実例を多くカタログすることは価値がある。今後，LSDオンラインサービスで実装している共起検索技術1）を応用して，全単語の連接する用語について頻度を調査し，一定値以上の共起関係が検出される複合語を網羅的に収録することで対応する必要があるだろう。また，生命科学は広範な学問領域を包含しているため，各々の専門領域に特化したコーパスに基づく用語の抽出13）も多義性の解決に有用と考えられる 。

4.5 日本語に特有な語彙と表記の多様性

LSDには約7万語の漢字変換辞書に対して，対訳を有するものは約5万語であるが，この差分は図5Bの結果から主として低頻度の日本語で構成されていることがわかる。今回，英語と日本語の頻度解析データを比較してみたところ，英訳のない（あるいは頻度の少ない）日本語には日本語独特の概念や事物（例：生薬，漢方薬など）が多く含まれていることがわかった。日本の科学を外国人だけでなくコンピュータに理解させるためには，これらの訳語を定義し，シソーラスにマッピングする新たな努力が必要であろう。
また，日本語の頻度解析からわかる日本語の特徴は，同一概念に対する表記の多様性である。典型的な例として「protein」に対して「タンパク質」「タンパク」「蛋白質」「蛋白」「たんぱく質」「たん白質」「プロテイン」等の訳語があり，事実 これらが日本語コーパスから検出された。多様性の一因として 異なる学問領域で異なる表記が推奨されているが，実際には隣接する語句によって用いられる表記には大きな偏りがある場合が多い。上記の例ではプリオン蛋白(prion protein)，プロテインキナーゼ（protein kinase），蛋白尿（proteinuria）などが研究分野に関係なく慣用的な表記として用いられていた。したがって，対訳シソーラスにおいては連接語も含めて頻度の高い表記法を優先的に収録するのが好ましいように思われる。

4.6 コーパスの選択

今回用いた英語コーパスはPubMed論文抄録を80%，教科書テキストを20%含んでいる。抄録という限られたスペースで序論から結論までを記述する文章が主体である結果，定型的な表現が多くなる代わりに詳細な記述が少なく，頻度解析プロットにおいて傾斜が緩い，すなわち高頻度の用語（例：expression，suggestなど）が多く 用いられる結果が得られたと考えられる。研究者が記述する自然な文章を解析する目標を考えると，今後は英語についても論文の全文テキストを用いた解析が必要であろう。 一方，日本語コーパスはほとんどが教科書の全文で詳細な記述が含まれる結果として，頻度解析プロットではより典型的な分布を示したと思われるが，コーパスサイズを考慮しても日本語の 語彙数は英語のそれより小さかった 。このことは，前述したように遺伝子名や学名などが翻訳し得ない事実を反映していると考えられる。
最近は対訳コーパスを用いて対訳を抽出する試みが数多く行われている。生命科学における対訳テキストは特許において最も典型的に見られるため，特許テキストを利用した評価系にも興味が持たれる。専門文書の日本語コーパスの構築にあたっては電子化や著作権といった現実的障害が横たわっているが，webにおける専門情報が著しく増大している現状からは，それらの有効な利用も考えられる。

4.7 より良い対訳シソーラスへ

MeSHはカテゴリーAからDにわたる事物および病名の分類はマクロからミクロ的な視点，あるいは物質の構造等に基づいて分類学的かつ体系的に行われている。しかし，現象や概念を記述するその他のカテゴリーに関しては，必ずしも体系的に分類されていない印象を受けた。発生学や物理化学などの周辺領域についてもMeSHは収録語彙が乏しい。さらにMeSHには名詞のみが収録されているが，LSDは略語，動詞，形容詞，副詞を収録しているため，これらは既存シソーラスとはリンクしない。しかしながら，シソーラスの応用先として実際のテキスト解析を考慮すると，これらを克服する方策を考案する必要がある。形容詞や副詞に対して

，共起する名詞や動詞の統計値を収録することも有用であろう 。

5. おわりに

本研究で制作した対訳シソーラスの実効性については，セマンティックウェッブ検索ページや機械翻訳辞書への実装を行って検証していきたいと考えている。また，今回明らかになった問題点や課題を踏まえて，他の様々なシソーラスも参考にしつつ， LSDを新しい機能的な対訳シソーラスとして体系化することを目指していきたいと考えている。

6. 謝辞

本研究は，21世紀COEプログラム「ゲノム科学の知的情報基盤・研究拠点形成」，学術振興会科学研究費補助金 研究成果公開促進費，文部科学省科学研究費補助金 特定研究（応用ゲノム），カシオ科学振興財団からの助成を得て行われた。

参考文献

1) ライフサイエンス辞書.
http://lsd.pharm.kyoto-u.ac.jp/ja/

2) 金子周司. ライフサイエンス辞書とは. 情報管理. 2006; 印刷中.

3) 金子周司. 無料ライフサイエンス辞書の活用と効能. ファルマシア. 2006; 127(6): 印刷中.

4) 金子周司. ライフサイエンス辞書から生命科学オントロジーへ. 情報知識学会誌. 2005; 15(4): 1-10.

5) ライフサイエンス辞書プロジェクトの現在と展望.
http://www.apple.com/jp/medical/dictionary/kaneko/
アップル社医療サイト（2005年12月公開）

6) Onogi Y, Ohe K, Tanaka M et al. Mapping Japanese medical terms to UMLS Metathesaurus. Medinfo. 2004; 11(1): 406-410.

7) UMLSと連携した日本語医学用語シソーラスの作成http://jumls.h.u-tokyo.ac.jp/平成15年度厚生科学研究費補助金医療技術評価総合研究事業

8) 藤田信之，金子周司. ライフサイエンスのための英和変換ツール. コンピュータサイエンス. 1995; 2(1): 41-45.

9) Gnene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006; 34: D322-D326.

10) Campbell DA, Johnson SB. A technieuq for semantic classification of unknown words using UMLS resources. Proc. AMIA Symp. 1999: 716-720.

11) Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. J Am Med Inform Assoc. 2002; 9(6): 621-636.

12) Nenadic G, Spasic I, Ananiadou S. Terminology-driven mining of biomedical literature. Bioinformatics 2003; 19(8) 938-943.

13) 竹村匡正，松井弘子，芦田信之用例に基づく医療用語知識の体系化について医療情報学. 2004; 24(1): 139-145.