

文献情報の解析に基づく 対訳シソーラスの評価

京都大学大学院薬学研究科
生体機能解析学分野

金子 周司

製品評価技術基盤機構
ゲノム解析部門

藤田 信之

- ライフサイエンス辞書とは
- 頻度解析の手法
- MeSHリンクによる
シソーラス構築と評価

2006年7月1日

医療情報学会春期学術大会
(神戸)

The screenshot shows the homepage of the Life Science Dictionary Project. The browser address bar displays 'http://lsd.pharm.kyoto-u.ac.jp/ja/index.html'. The page features a navigation menu on the left with categories like 'サービス' (Services), '資料' (Materials), 'About Us', and '辞書ダウンロード' (Dictionary Download). The main content area is divided into several sections: 'オンライン辞書サービス' (Online Dictionary Service) with a 'WebLSD' link and update date '2006.03.30'; 'オンデマンド英語教材' (On-demand English Textbook) with a 'Go!' link and update frequency 'weekly'; 'オンライン変換サービス' (Online Conversion Service) with links for 'E to J Vocabulary', 'E to J', and 'WebSpell', all with 'Go!' links and update dates; and '携帯用WebLSD' (Mobile WebLSD) with a QR code and a '送信' (Send) button. A footer contains copyright information for 2006 and a 'PAGE TOP' link.

WebLSD Query Input

http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/in

BACK TO HOME

英和検索結果

最新対訳の受け付け | ヘルプ

- **apoptosis** ***** [共起検索](#) [音声](#)
 (遺伝子にプログラムされた能動的な細胞死) **アポトーシス**, **アポトosis**, **プログラム細胞死**, **予定死**
 [あぼとーしす, あぼふとーしす, ぶろくらむさいぼうし, よていし]
 【関連語】 [apoptotic](#), [programmed cell death](#), [programmed death](#)
 【用法】 induce apoptosis [アポトーシスを誘発する] / T-cell receptor-induced apoptosis [T細胞受容体が誘導するアポトーシス] / undergo apoptosis [アポトーシスを起こす] / UV-induced apoptosis [紫外線が誘発するアポトーシス] [用例](#)
- **apoptosis induction** *** [共起検索](#)
アポトーシス誘導 [あぼとーしすゆうどう]
 【関連語】 [induction of apoptosis](#)
- **Fas-mediated apoptosis** *** [共起検索](#)
F a s 誘発アポトーシス [ふあすゆうはつあぼとーしす]
- **induction of apoptosis** *** [共起検索](#)
アポトーシス誘導 [あぼとーしすゆうどう]
 【関連語】 [apoptosis induction](#)
- **tumor necrosis factor-related apoptosis-inducing ligand** ** [共起検索](#)
腫瘍壊死因子関連アポトーシス誘発リガンド [しゅようえいしんしかんれんあぼとーしすゆうはつりがんど]
 【関連語】 [TRAIL](#)

音声付き英和・和英検索 [共起検索](#) [音声付き英和・和英検索ヘルプ](#)

apoptosis [search](#) [reset](#)

▼検索語句 を含む で始まる で終わる に一致する

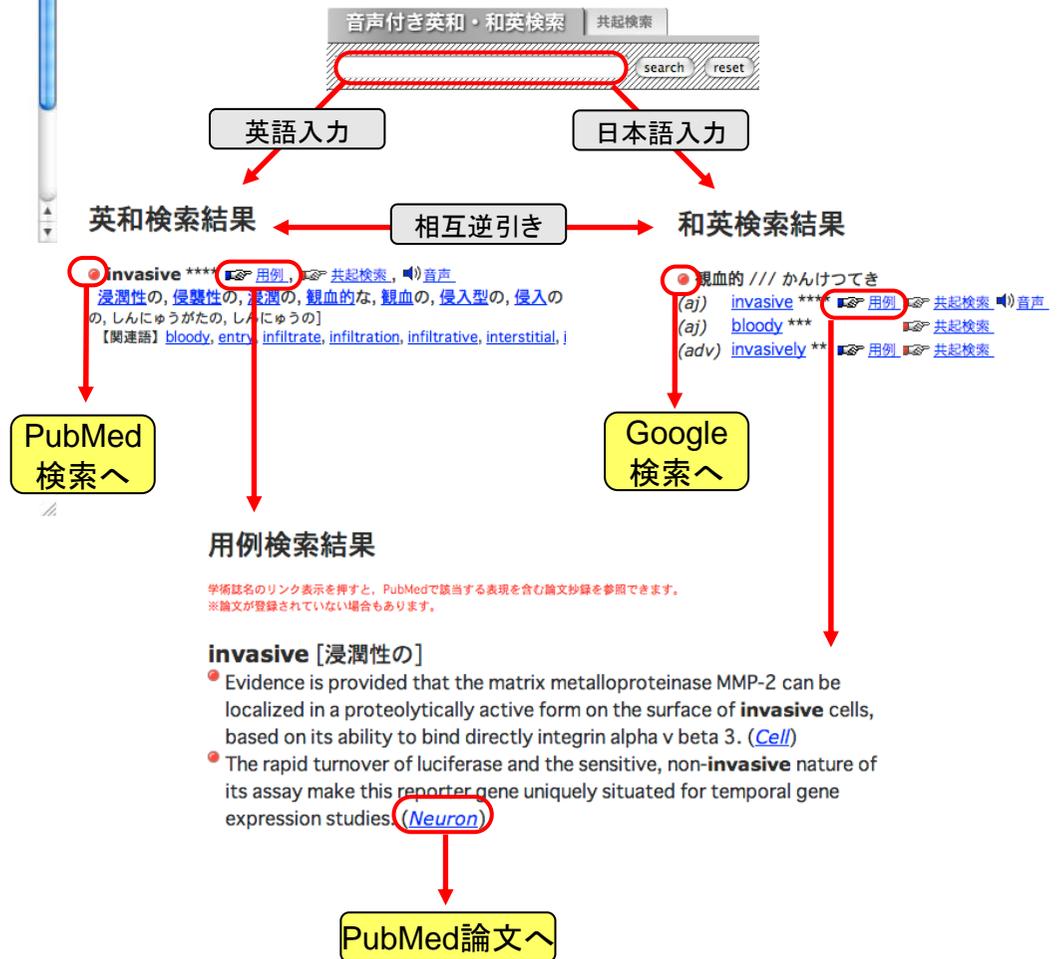
▼英語の大文字と小文字を区別 しない する

▼検索結果を最大 50 100 200件表示

▼日本語の語尾変化を無視 する しない

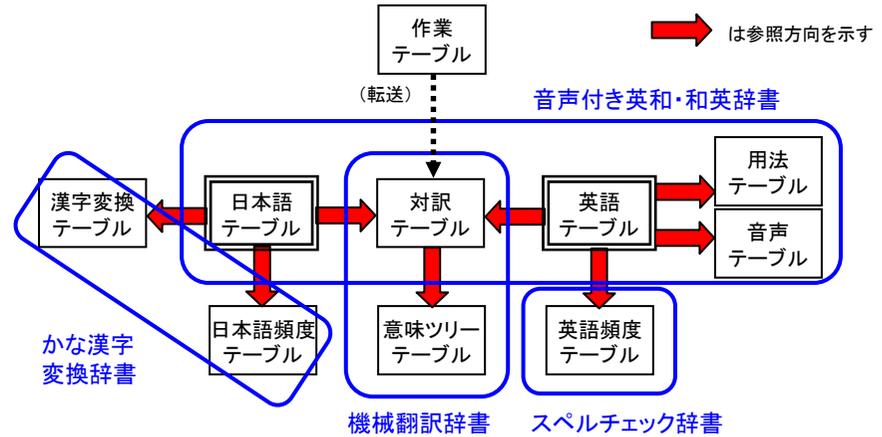
▼和英検索に かな/漢字 (推奨) ローマ字を使用

オンライン辞書サービス WebLSD



LSDデータベース構造

英語テーブルはコーパス解析から
対訳テーブルは手作業で規定
シノニムはあるが、ツリーではない



LSD SYSTEM

アンボアラー 英語 対訳 日本語 かな漢 よみ 用法 用例 英語音声 対訳分析指定 シノニムス

List INDEX New Delete Find ConstrainFounc Show All Print Sort

削除 対訳追加 (対訳・日本語・よみ・かな漢) 修正日 2005. 7. 24 修正者 okanoko ※初期作成者 LSD3

英語コード E000421
英語表記 abstract
EtoJ 辞書/注釈 抄録
重要度/注釈 4 ****
頻度の和 1134 PubMed 802
学習 0 教科書 332
英語注釈
参照先指示
MeSH link
音声数 2

対訳追加用
P_日本語表記 P_日本語語尾 P_よみ
P_英語品詞 P_英和注釈
P_意味情報 P_MeSH

対訳

対訳コード	英和注釈	日本語表記 MoSH ID & term_string	日本語語尾	英語品詞	意味情報	
JE000209		抄録		(n*)	書類	
1						<input type="checkbox"/> 削除
JE000208		アブストラクト		(n*)	知的産物	<input type="checkbox"/> 削除
2						<input type="checkbox"/> 削除
JE043593		抽出	する	(vt)	動詞	<input type="checkbox"/> 削除
3						<input type="checkbox"/> 削除
JE000210		抽象化	する	(vt)	動詞	<input type="checkbox"/> 削除
4						<input type="checkbox"/> 削除

用法

用法コード	用法	日本語訳	英語見出し語
<input type="checkbox"/> UC00073	abstract	抄録	abstract
<input type="checkbox"/> UC00074	abstract	抜粋	abstract
<input type="checkbox"/> UC00075	abstract	除去する	abstract

用法コードを選択 abstract
U000075 用例調査 1

文例コード 文例 書籍情報 用例注釈
用法コード 日本語用法 用例優先順

文例コード	文例	書籍情報	用例注釈
SOC0211	The ability of permanganate to	Science	
U000075	 abstract </3> a hydrogen atom is	除去する	4

かな漢

かな漢コード	よみ表記 日本語表記	変換注釈 日本語品詞
YJ000604	あぶすとらくと アブストラクト	一般名詞
YJ012865	しょうろく 抄録	一般名詞
YJ018268	ちゅうしゅつ 抽出	名詞サ変
YJ018275	ちゅうしょうか 抽象化	名詞サ変

英シソーラス

参照先 コード 表記

E000423	abstraction	<input type="checkbox"/>
E040457	extract	<input type="checkbox"/>

音声

英語表記	発音者	音声注釈	ファイル名
abstract	Vigers		0028A.WAV <input type="checkbox"/>
abstract	Vigers		0028B.WAV <input type="checkbox"/>

研究目的

日本語コーパスの解析による英語との比較

LSD の現状評価

日本語と英語の間における概念の相違

MeSH とのリンクによるシソーラスの構築

独自のシソーラスへの足がかり

MeSH Tree Structures

[Biological Phenomena, Cell Phenomena, and Immunity \[G04\]](#)

[Cell Physiology \[G04.335\]](#)

[Cell Death \[G04.335.139\]](#)

▶ [Apoptosis \[G04.335.139.160\]](#)

[Anoikis \[G04.335.139.160.060\]](#)

[DNA Fragmentation \[G04.335.139.160.200\]](#)

[Necrosis \[G04.335.139.638\]](#)

The screenshot shows a web browser window titled "WebLSD Query Input" with the URL "http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/it". The search results list various interleukin terms with their English descriptions and MeSH IDs. Below the results is a search control panel with a search box containing "interleukin" and buttons for "search" and "reset". The control panel includes several options for refining the search, such as "検索語句" (Search terms) and "和英検索" (J-E search).

WebLSD Query Input
http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/it

- **interleukin-10** *** [共起検索](#)
(サイトカイン産生抑制因子の一種) [インターロイキン 1.0](#) [いんたーろいきんてん, いんたーろいきんじゅう]
【関連語】 [IL-10](#)
- **interleukin-11** ** [共起検索](#)
(造血機能を活性化するサイトカインの一種) [インターロイキン 1.1](#) [いんたーろいきんいれぶん, いんたーろいきんじゅういち]
【関連語】 [IL-11](#)
- **interleukin-12** *** [共起検索](#)
(NK細胞を刺激するサイトカインの一種) [インターロイキン 1.2](#) [いんたーろいきんとうえるぶ, いんたーろいきんじゅうに]
【関連語】 [IL-12](#)
- **interleukin-13** *** [共起検索](#)
(活性化ヘルパーT細胞が産生するB細胞刺激因子の一種) [インターロイキン 1.3](#) [いんたーろいきんじゅうさん]
【関連語】 [IL-13](#)
- **interleukin-14** *
(高分子量のB細胞増殖因子の一種) [インターロイキン 1.4](#) [いんたーろいきんじゅうよん]
【関連語】 [IL-14](#)
- **interleukin-15** *** [共起検索](#)
(IL-2類似のT細胞増殖因子の一種) [インターロイキン 1.5](#) [いんたーろいきんじゅうご]
【関連語】 [IL-15](#)
- **interleukin-16** ** [共起検索](#)
(リンパ球走化性サイトカインの一種) [インターロイキン 1.6](#) [いんたーろいきんじゅうろく]
【関連語】 [IL-16](#)
- **interleukin-17** ** [共起検索](#)
(間質系細胞や滑膜細胞に作用するサイトカインの一種) [インターロイキン 1.7](#) [いんたーろいきんじゅうなな]
【関連語】 [IL-17](#)
- **interleukin-18** ** [共起検索](#)
(IFN γ 産生誘導性サイトカインの一種) [インターロイキン 1.8](#) [いんたーろいきんじゅうはち]

音声付き英和・和英検索 [共起検索](#) [音声付き英和・和英検索ヘルプ](#)

interleukin search reset

▼ 検索語句 を含む で始まる で終わる に一致する

▼ 英語の大文字と小文字を区別 しない する

▼ 検索結果を最大 50 100 200件表示

▼ 日本語の語尾変化を無視 する しない

▼ 和英検索に かな/漢字(推奨) ローマ字を使用

英語コーパスを用いた頻度解析(1)

英語コーパス

PubMed 収録のインパクトファクターの高い学術誌(89誌)にアメリカ・イギリスの研究機関から1995-2004年に発表された論文抄録
Bookshelf公開の教科書テキスト等も使用
合計 463 Mbyte(6000万語)

解析手法

1. 単語間のスペースのみを認識して切断するPerlスクリプトを用いて単語および年度毎に出現頻度を計数した。
2. 語尾変化を考慮しないで, LSD収録語とのマッチングを行った

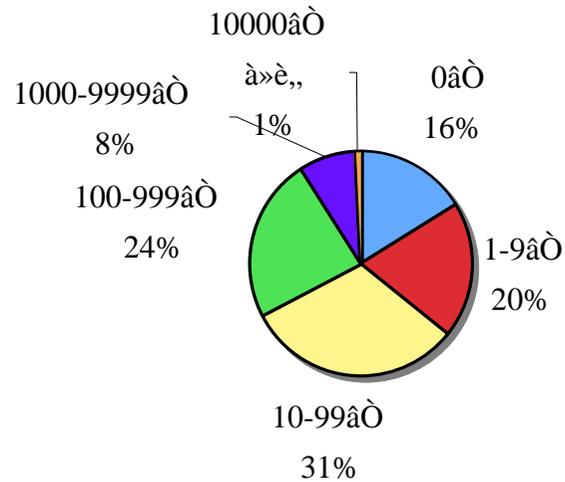
PubMed単語	LSD2006対訳	総頻度	F1995	F1996	F1997	F1998	F1999	F2000	F2001	F2002	F2003	F2004	合計	十年増加数	増加率
neuropathic	神経障害性	309	7	22	25	23	15	15	25	15	56	40	243	34	13.79
neuropathies		70	2	6	3	2	6	2	9	2	9	8	49	5	9.18
neuropathogenesis	神経病因性	41	1	6	2	5	6	5	4	2	5	3	39	1	1.28
neuropathogenic		26	5	3	5	3	3	1	1		2		23	-3	-13.04
neuropathogenicity		11		1	3			2	1		2	1	10	1	10.00
neuropathol		1													
neuropathologic	神経病理的	56	2	5	9	2	3	7	6	2	4	5	45	1	2.22
neuropathological	神経病理学的	201	15	25	19	19	14	17	10	16	15	26	176	1	0.28
neuropathologically	神経病理学的に	22		5		4	2	3	2	1	2		19	-2	-7.89
neuropathologies		10					4				3	2	9	3	27.78
neuropathologist		1						1					1		0.00
neuropathologists		4		1	1			1		1			4	-1	-12.50
neuropathology	神経病理学	174	12	16	19	15	13	8	11	19	19	20	152	6	3.62
neuropathophysiology		3									1	1	2	1	50.00
neuropathy	神経障害	607	40	59	66	54	52	43	42	53	60	61	530	11	2.08
neuropathy-associated		1							1				1		0.00
neuropeptidase		9		1	3				2			2	8	1	6.25
neuropeptidases		2			1		1						2		0.00
neuropeptide	神経ペプチド	945	102	82	99	59	72	77	89	61	68	97	806	-10	-1.18

英語コーパスを用いた頻度解析(2)

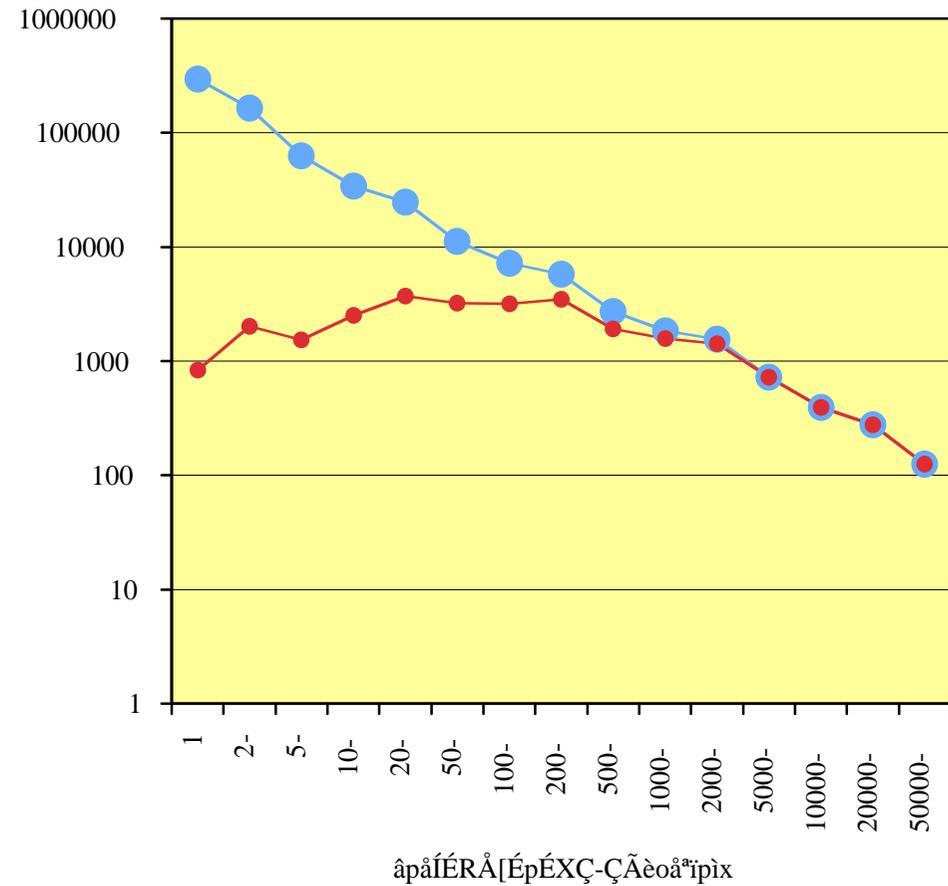
解析手法

- 名詞や動詞の規則変化に対応する逐語訳 EtoJエンジンを用いて, LSD収録語の出現頻度解析を行った。

LSD収録英語の頻度分布



英語コーパス全単語とLSD収録英単語の頻度分布



→ LSD収録語はコーパスを88%網羅

日本語コーパスを用いた頻度解析

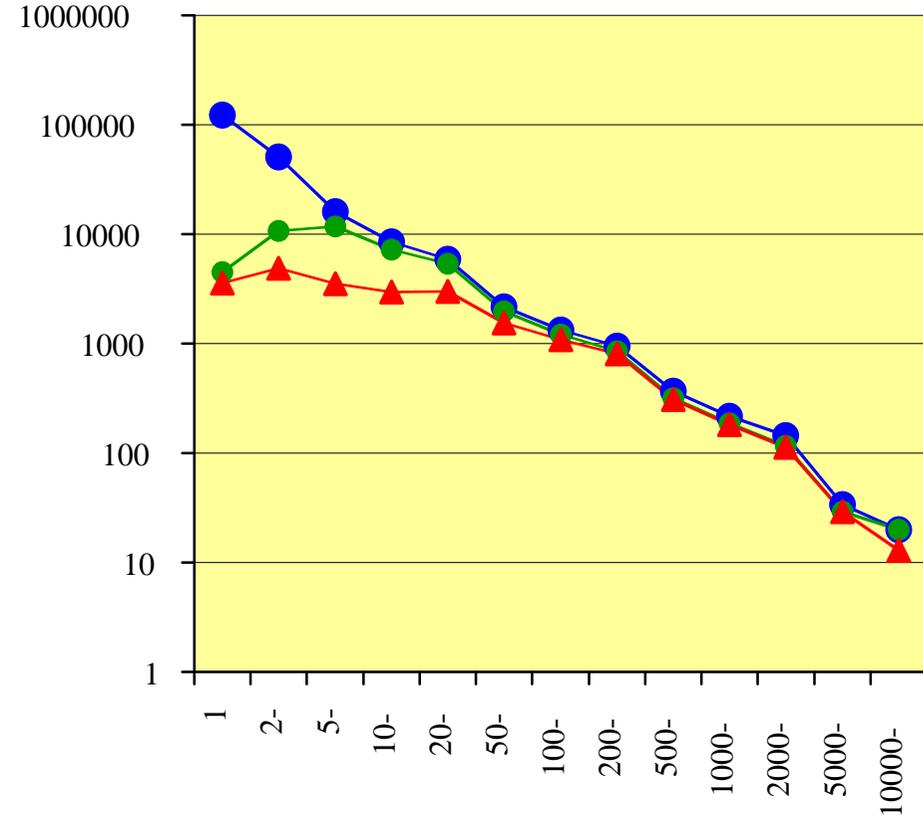
日本語コーパス

ある出版社の協力により提供された医学
 総説誌1996-2002年の全文
 一部, 臨床医学テキストも使用
 合計 34 MByte (2000万文字)

解析手法

1. 漢字, カタカナ, ひらがな, アルファベット, 数字の境目を認識して最長連続する要素(単語)を抽出するPerlスクリプトで計数
2. 日本語コーパス中でLSD収録語および1で得られた単語の出現頻度を計数するPerlスクリプトを使用

全単語とLSD収録語の頻度分布



コーパスを構成する文字種の割合



ìÝñ{áíÉRÅ[ÉpÉXÇ-ÇÃèòã*ipìx

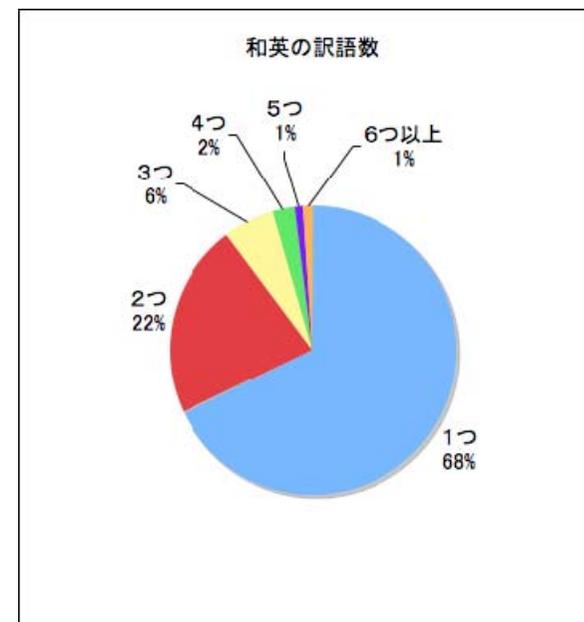
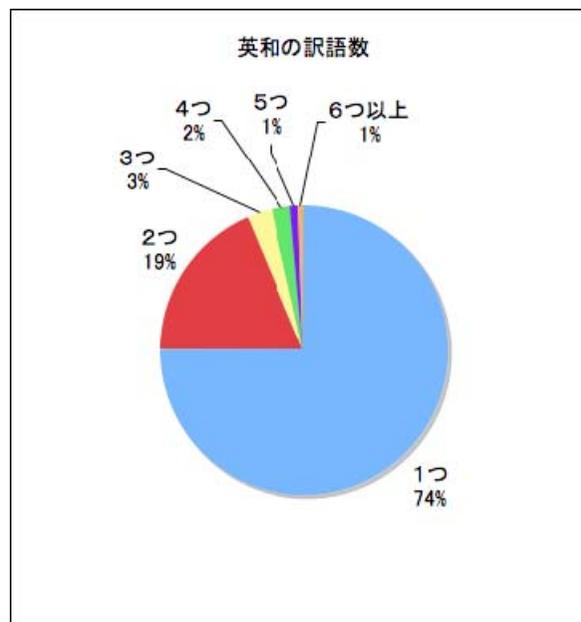
● ç-íllòAè±óvëf ● äøööiöæállé`èè ▲ ëÇñÛé`èè

英語と日本語の関係

英和, 和英ともに1対1関係にある語句は全体の7割程度

Metabolism 代謝
Transcription 転写
Cancer 癌
などなど

しかし癌は「がん」「ガン」とも

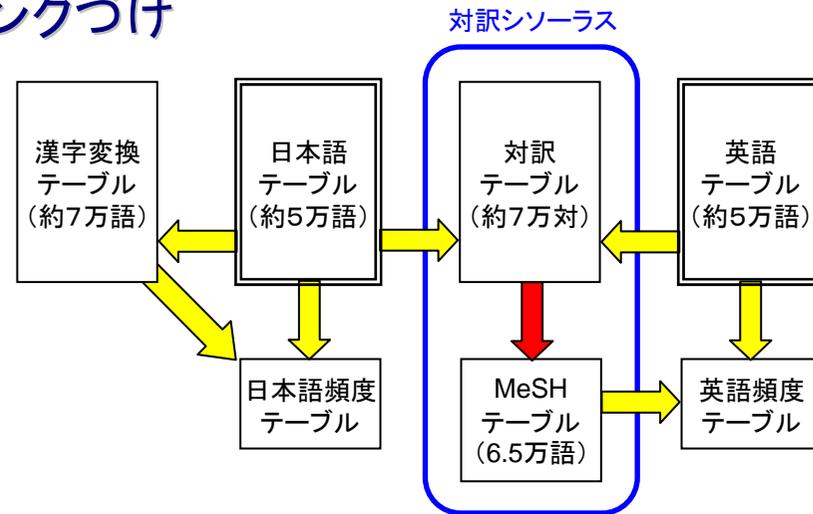


対訳関係は複雑

英語 (* MeSH term)	agent	drug	日本語	剤	薬
immunosuppressive	263*	229	免疫抑制	102	6
antihypertensive	107*	82*	降圧	2	71
antibacterial	131*	55	抗菌	3	313
anti-inflammatory	197*	411	抗炎症	6	102
anticancer	125	121	抗癌	633	6

MeSH term の標準表記化と LSD へのリンクづけ

実際に用いられる語順への変更
 複数形から単数形への統一



MeSH2006						
Depth	tree_number	descriptor_ID	Descriptor_lookup	descriptor::proper_form_REV	英語::EtoJ 辞書	
0	A01	D001829	Body Regions	Body Region	身体領域	
1	A01.047	D000005	Abdomen	Abdomen	腹部	
2	A01.047.025	D034841	Abdominal Cavity	Abdominal Cavity	腹腔	
3	A01.047.025.600	D010537	Peritoneum	Pertoneum	腹膜	
4	A01.047.025.600.225	D004312	Douglas' Pouch	Douglas' Pouch		
4	A01.047.025.600.451	D008643	Mesentery	Mesentery	腸間膜	
5	A01.047.025.600.451.535	D008646	Mesocolon	Mesocolon	結腸間膜	
4	A01.047.025.600.573	D009852	Omentum	Omentum	網	
4	A01.047.025.600.678	D010529	Peritoneal Cavity	Pertoneal Cavity	腹腔	
3	A01.047.025.750	D012187	Retroperitoneal Space	Retroperitoneal Space	後腹膜腔	
2	A01.047.050	D034861	Abdominal Wall	Abdominal Wall	腹壁	
2	A01.047.365	D006119	Groin	Groin	鼠径部	
2	A01.047.412	D007264	Inguinal Canal	Inguinal Canal	鼠径管	
2	A01.047.849	D014472	Umbilicus	Umbilicus	臍	
1	A01.176	D001415	Back	Back	背中	
2	A01.176.519	D008161	Lumbosacral Region	Lumbosacral Region	腰仙部	
2	A01.176.780	D012445	Sacroccocygeal Region	Sacroccocygeal Region	仙尾骨部	
1	A01.236	D001940	Breast	Breast	乳房	
2	A01.236.249	D042361	Mammary Glands, Human	Human Mammary Gland		
2	A01.236.500	D009558	Nipples	Nipple	乳頭	
1	A01.378	D005121	Extremities	Extremity	末端(きわみ)	
2	A01.378.100	D000672	Amputation Stumps	Amputation Stump		

MeSH リンクによる LSD 対訳シソーラス化

	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)

LSD SYSTEM													
テンポラリー		英語	対訳	日本語	かな漢	よみ	用法	用例	英語音声	対訳分野指定	シソーラス	英語頻度	
n 対訳コード	英語コード・品詞	英語表記	日本語コード	日本語表記	日本語語長	優先順・解説	意味情報	翻訳注釈	MeSHリンク				
JE068736	E157067	2 acute aortic dissection	J041788	解離性大動脈瘤	英和 1 (病名)	病名	T						
JE061408	E152552	1 acute B-cell leukemia	J051255	急性B細胞白血病	英和 1 (病名)	病名	T T045626 B-Cell Leukemia, Acute	C04.557.337.428.500.100					
JE064820	E154567	2 acute brain injury	J053144	急性脳挫傷	英和 1 (病名)	病名	T T005495 Acute Brain Injuries	C10.228.140.199					
JE061409	E152553	2 acute cholecystitis	J041369	急性胆嚢炎	英和 1 (病名)	病名	T T521696 Acute Cholecystitis						
JF064431	F154273	2 acute coronary occlusion	J043680	急性冠閉塞	英和 1	病名	T						
JF064478	F154272	3 acute coronary syndrome	J041222	急性冠症候群	英和 1 (不安定狭心症から血栓形成)	病名	T						
JE064429	E154272	3 acute coronary syndrome	J052884	急性冠状動脈症候群	英和 2	病名	T						
JE061849	E152871	3 acute disease	J041283	急性疾患	英和 1	病名分類	T T000603 Acute Disease	C23.550.261.125					
JE061850	E152871	3 acute disease	J051557	急病	英和 2	病名分類	T T000603 Acute Disease	C23.550.261.125					
JE068506	E156937	2 acute exacerbation	J041070	急性増悪	英和 1	現象	T						
JE065732	E155236	2 acute hemorrhagic conjunctivitis	J071193	急性出血性結膜炎	英和 1 (結膜下出血を伴う急性濾胞)	病名	T T009421 Conjunctivitis, Acute	C02.325.250.250					
JE065143	E154803	1 acute hemorrhagic leukoencephalitis	J053359	急性出血性白質脳炎	英和 1 (病名)	病名	T T014357 Leukoencephalitis, Acute	C10.114.375.382					
JE000037	E155452	2 acute hepatic failure	J046828	急性肝不全	英和 1 (病名)	病名	T T051097 Hepatic Failure, Acute	C03.552.308.500.750					
JE047244	E141426	3 acute hepatitis	J032500	急性肝炎	英和 1 (病名)	病名	T						

LSD英語

LSD日本語

MeSH Descriptor, Term と Tree

MeSH term ベースでの網羅率

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

LSD データベースから MeSH 参照例

Lsd2006

ブラウザ LSD SYSTEM

テンポラリー 英語 対訳 日本語 かな漢 よみ 用法 用例

List INDEX New Delete Find ConstrainFound Show All Print Sort

レイアウト: P_日本語

日本語コード: J005285

日本語表記: 肝細胞

日本語/和英注釈

重要度: 4

学習

作成日: 1996/03/07

作成者: LSD3

修正日: 2005/12/06

修正者: skaneko

よみ数: 1

最大よみ: 1

該当件数: 15
合計: 75869
未ソート

■かな漢追加用■ よみ・かな漢 追加

P_よみ表記

■かな漢■

(かな漢コード)

よみコード	よみ表記	日本語品詞	変換注釈	よみ優先順
(YJ005250)				
Y005111	かんさいぼう	一般名詞	ATOK既存	1

■対訳追加用■ 対訳追加 (対訳・英語)

P_英語表記

P_英語品詞

P_意味情報

P_英和注釈

P_日本語語尾

意味リストへ

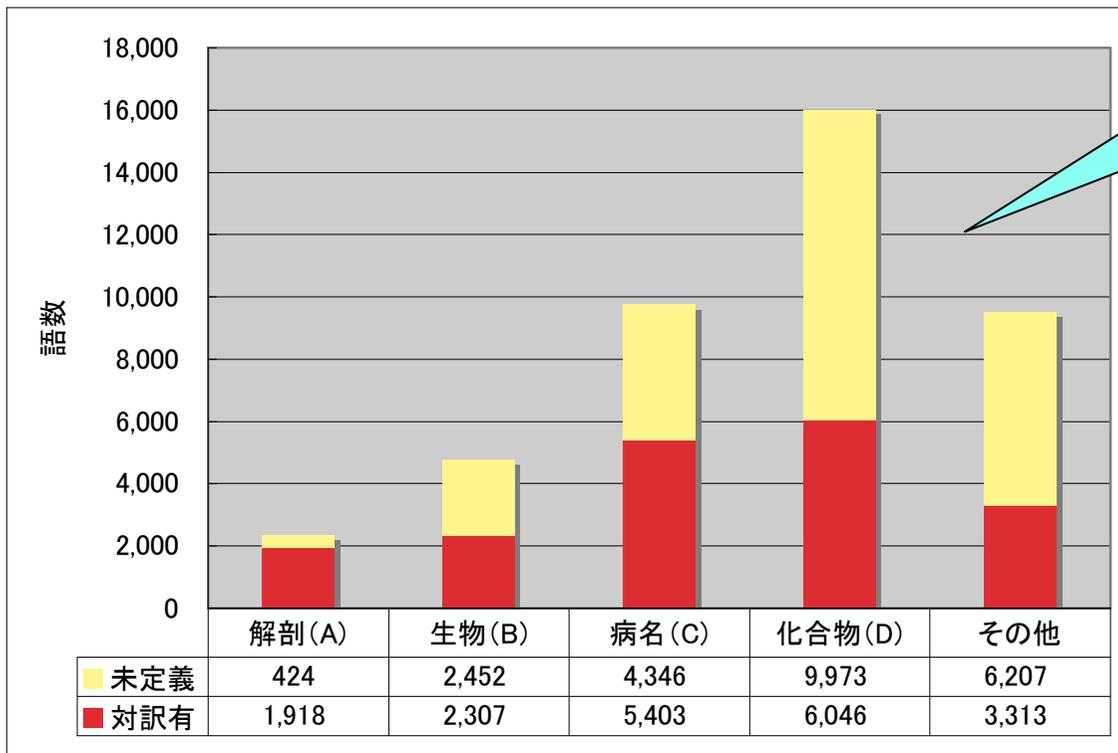
P_MeSH

■対訳■

日本語語尾 対訳コード	翻訳注釈 英語コード	和英解説	英語表記	英語品詞	和英優先順 意味情報	重要度	jtoe 辞書 ID	MeSH Term	MeSH Term
<input type="checkbox"/> JE017558	E050509		hepatocyte	(n*)	1	5179 ****	1	T409858	Hepatocytes
									細胞分類
<input type="checkbox"/> JE062078	E153050		liver cell	(n*)	2	471 ***	0	T409860	Liver Cells
									細胞分類
<input type="checkbox"/> JE017519	E050463		hepatic cell	(n*)	3	103 ***	0	T409859	Hepatic Cells
									細胞分類
<input type="checkbox"/> の JE017551	E050505		hepatocellular	(aj)	4	1170 ****	1		
									形容: その他
<input type="checkbox"/> の JE017560	E050511		hepatocytic	(aj)	5	30 **	0		
									形容: その他

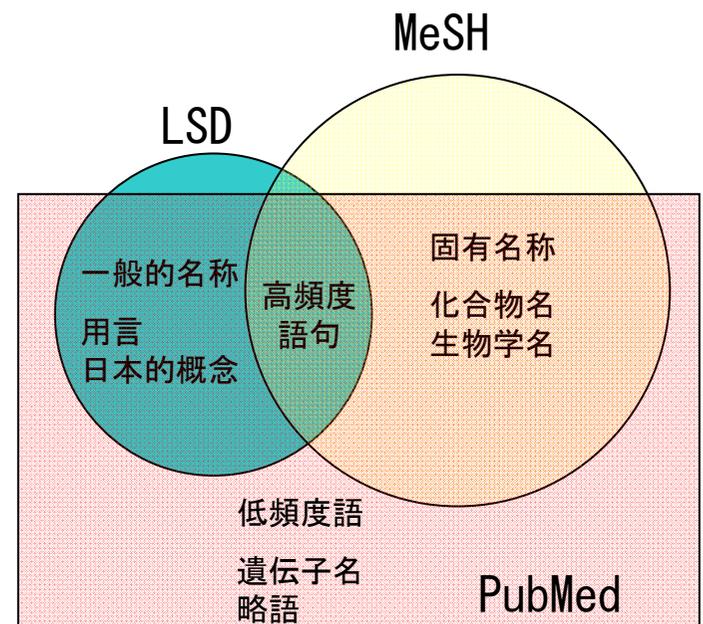
100 ブラウズ

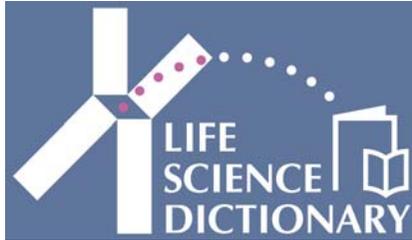
MeSH ツリーベースでの LSD 網羅率



MeSH ツリー (45,000カテゴリー) の
45%をカバー
(2006年6月)

課題
LSD収録PubMed頻出語のマッピング
コーパスを用いた関係抽出実験





ライフサイエンス辞書プロジェクト

1993年～



- 辞書制作
 - 金子周司
(京都大学大学院薬学研究科, 薬理学)
- 技術開発
 - 藤田信之
(製品評価技術基盤機構ゲノム解析部門, 生物遺伝学)
 - 鵜川義弘
(宮城教育大学環境実践研究センター, 情報教育学)
- 教材作成, 出版
 - 大武 博
(京都府立医科大学, 英語教育)
 - 河本 健
(広島大学医歯薬総合研究科, 生化学)
- 評価, 利用促進
 - 竹内浩昭
(静岡大学理学部, 行動生理学)
 - 竹腰正隆
(東海大学医学部, 分子生物学)