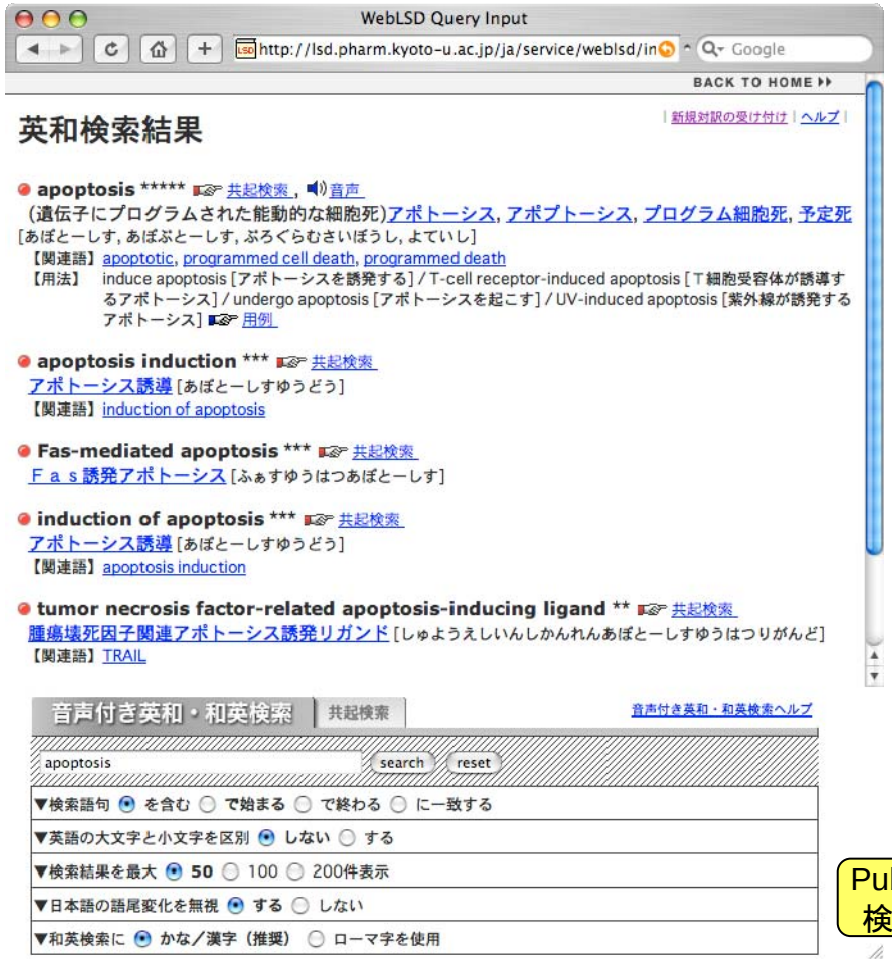
文献情報の解析に基づく対訳シソーラスの評価

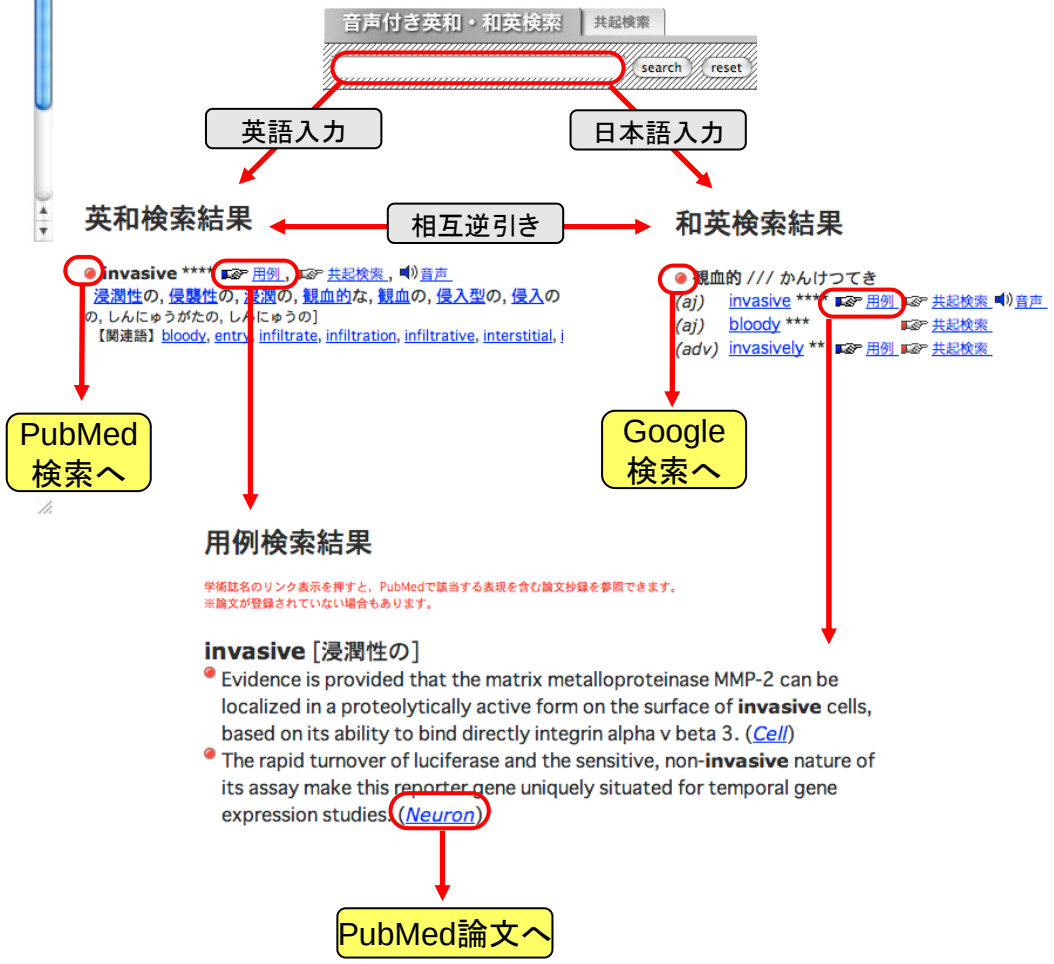
京都大学大学院薬学研究科
生体機能解析学分野
金子　周司
製品評価技術基盤機構
ゲノム解析部門
藤田　信之

- 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）

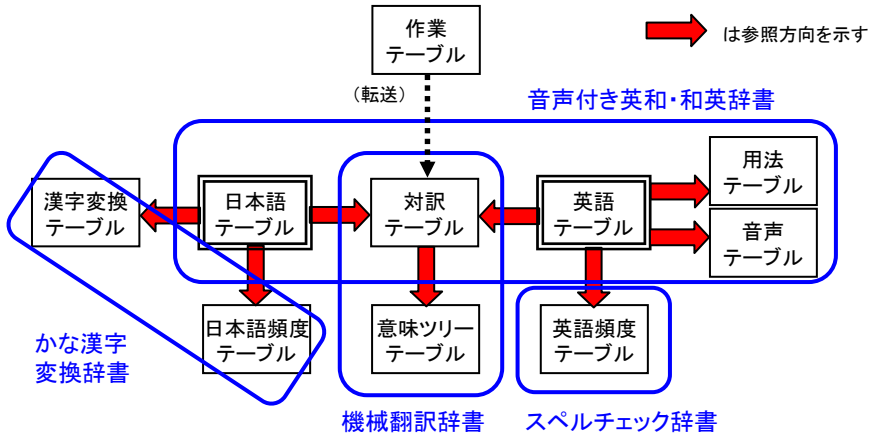
オンライン辞書サービス
WebLSD

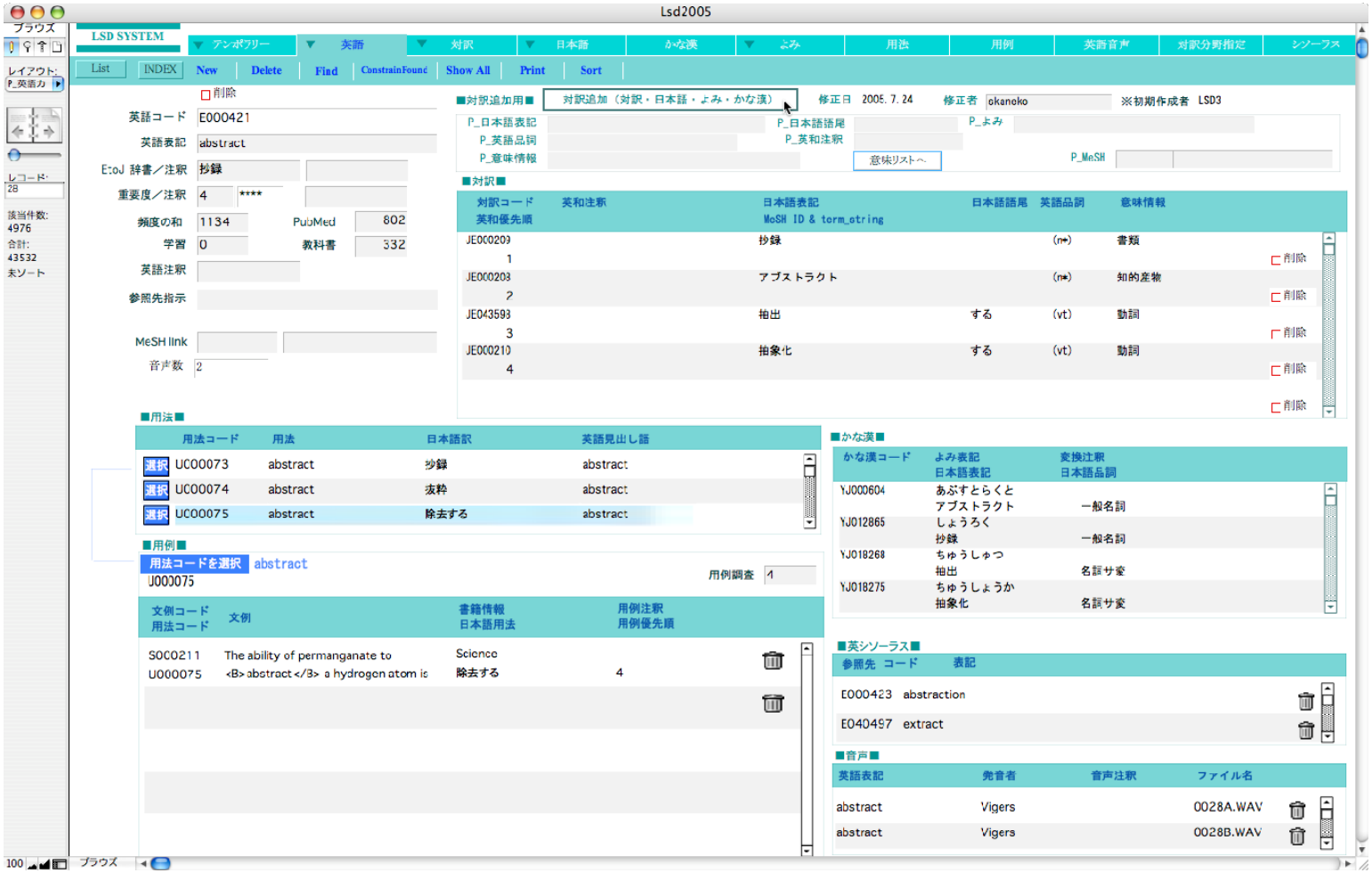




LSDデータベース構造

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





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]

英語コーパスを用いた頻度解析（1）

英語コーパス

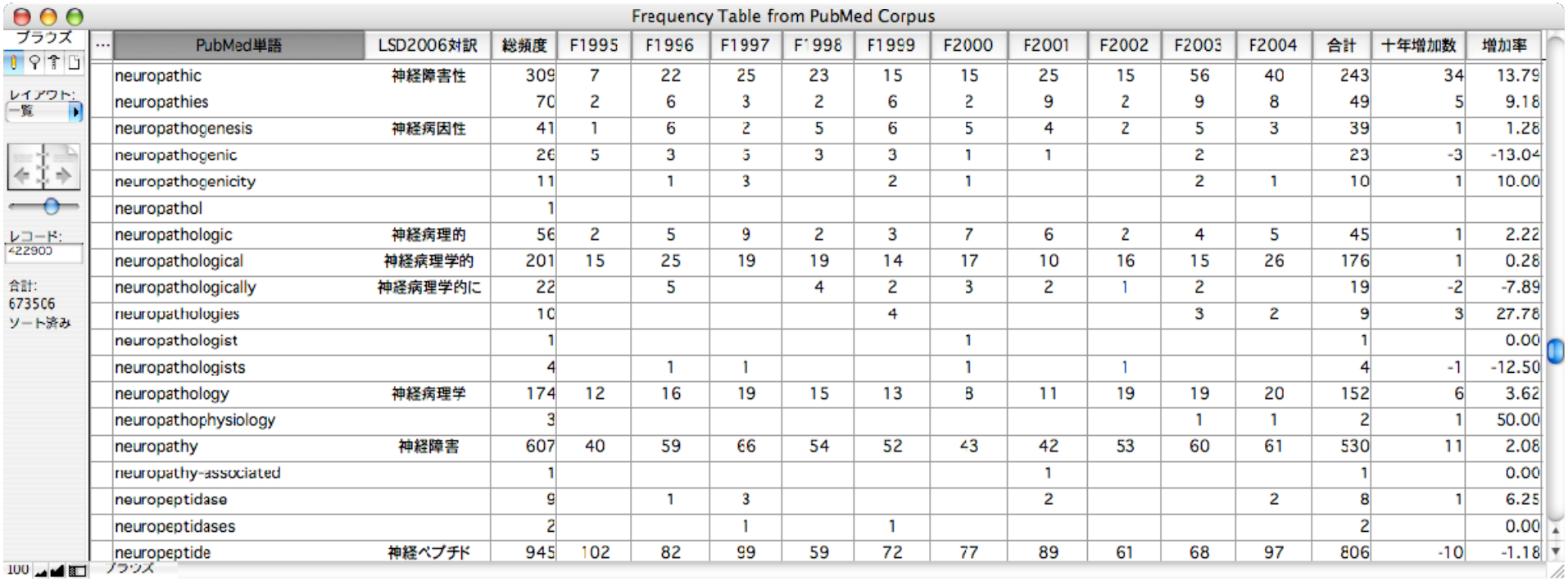
PubMed 収録のインパクトファクターの高い学術誌（89誌）にアメリカ・イギリスの研究機関から1995–2004年に発表された論文抄録

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



Frequency Table from PubMed Corpus

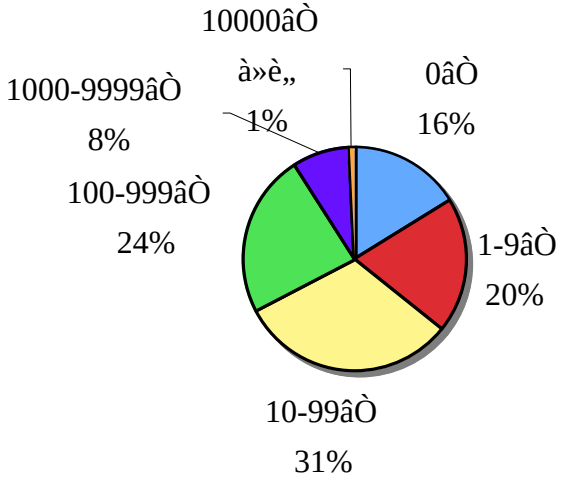
PubMed単語	LSD2006対訳	総頻度	F1995	F1996	F1997	F1998	F1999	F2000	F2001	F2002	F2003	F2004	合計	十年増加数	増加率
neuropathic	神経障害性	309	7	22	25	23	15	15	25	15	56	40	243	34	13.79
neuropathies		70	2	6	3	2	6	2	9	2	9	8	49	5	9.18
neuropathogenesis	神経病因性	41	1	6	2	5	6	5	4	2	5	3	39	1	1.28
neuropathogenic		26	5	3	5	3	3	1	1		2		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）

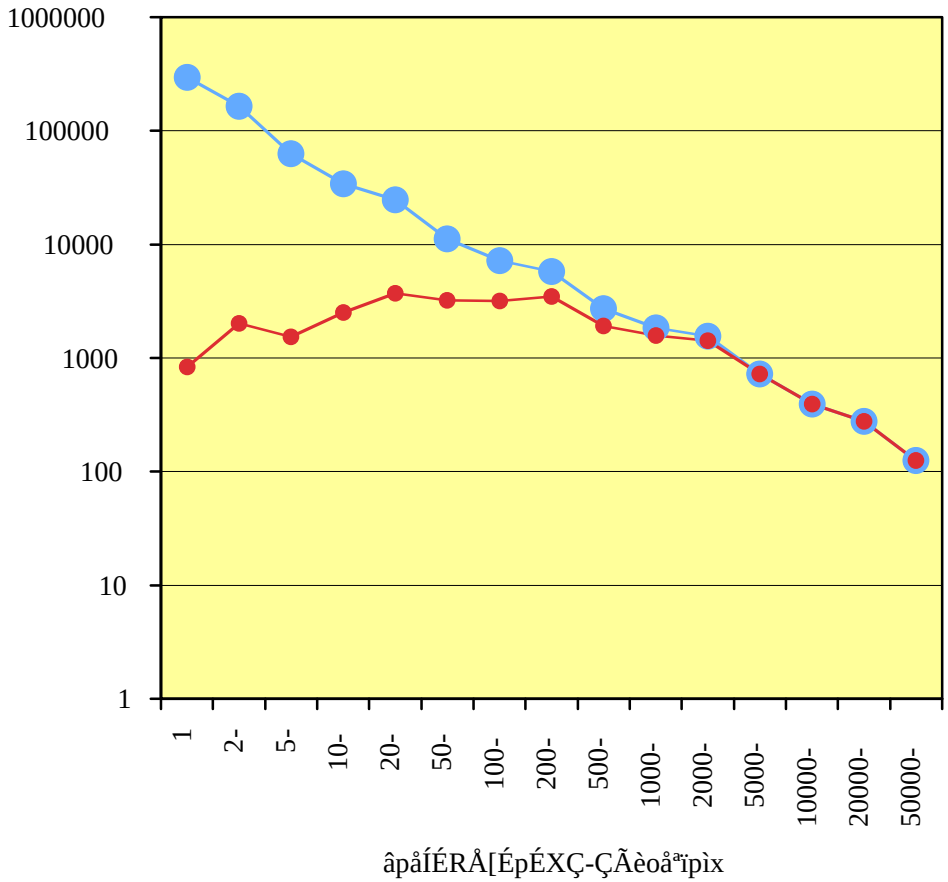
解析手法

– 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

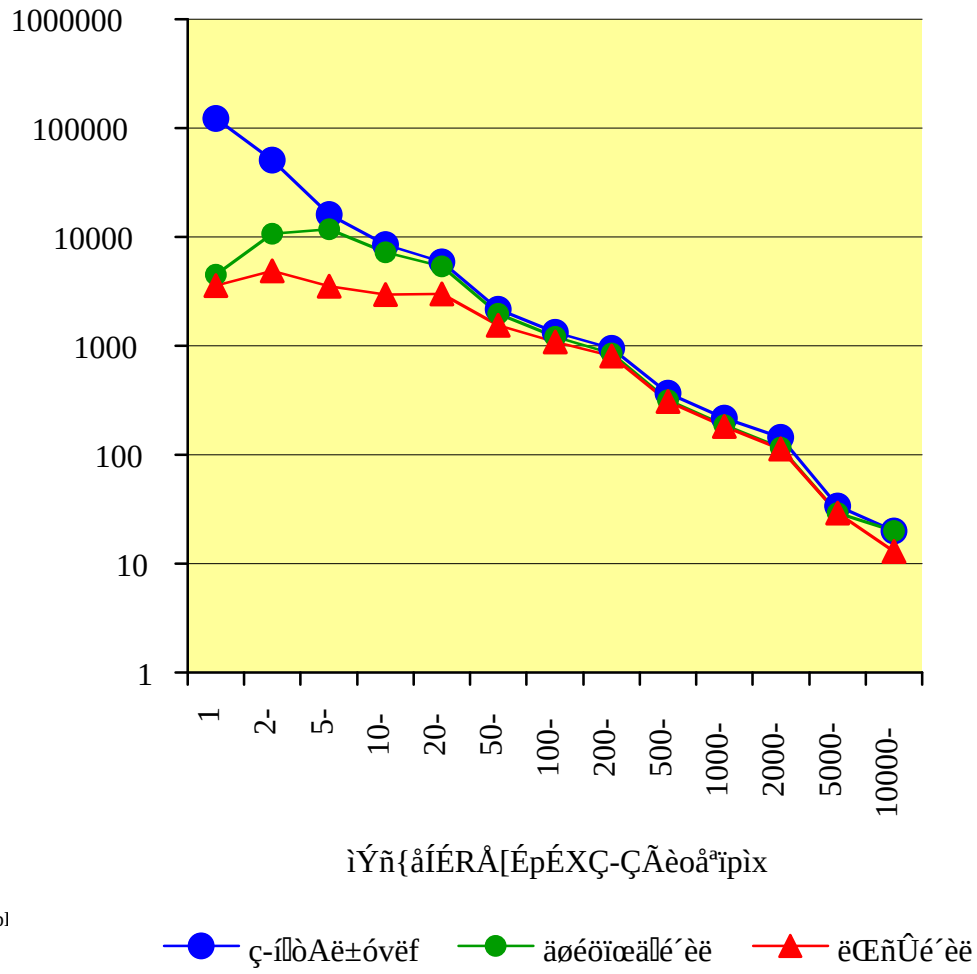
一部，臨床医学テキストも使用

合計 34 MByte (2000万文字)

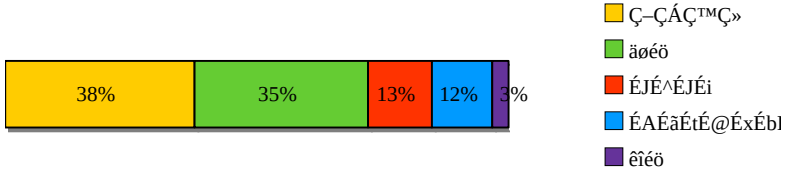
解析手法

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

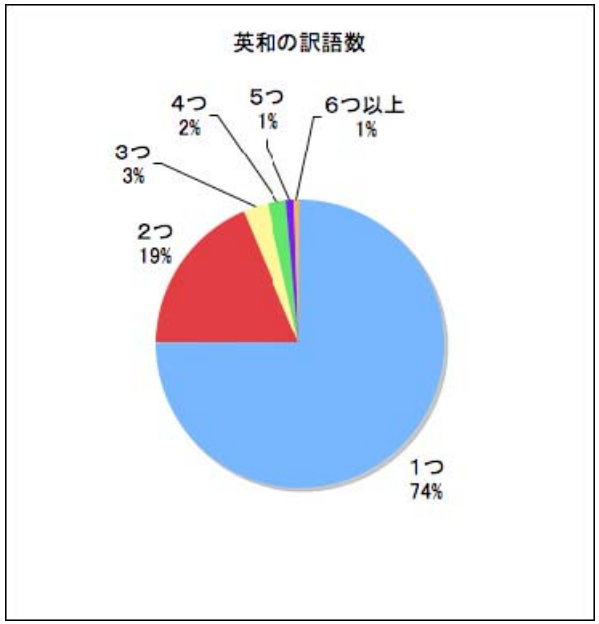


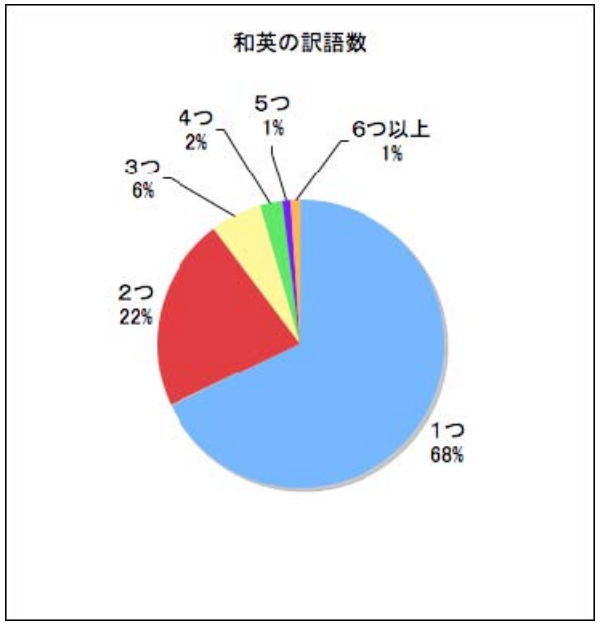
英語と日本語の関係

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





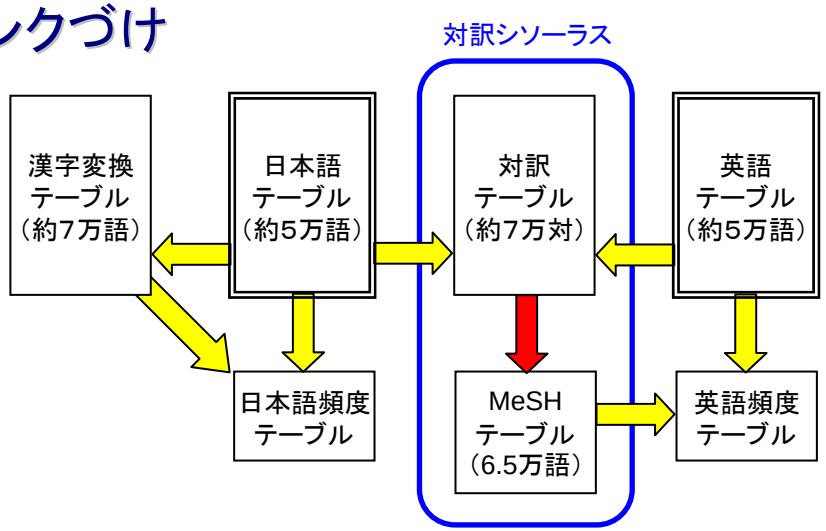
対訳関係は複雑

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

ブラウズ

レイアウト: tree_num

レコード: 1

合計: 44501

ソート済み

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	Peritoneum	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	Peritoneal Cavity	腹腔
3	A01.047.025.750	D012187	Retroperitoneal Space	Retroperitoneal Space	後腹膜腔
2	A01.047.050	D034861	Abdominal Wal	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	Sacrococcygeal Region	Sacrococcygeal 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)

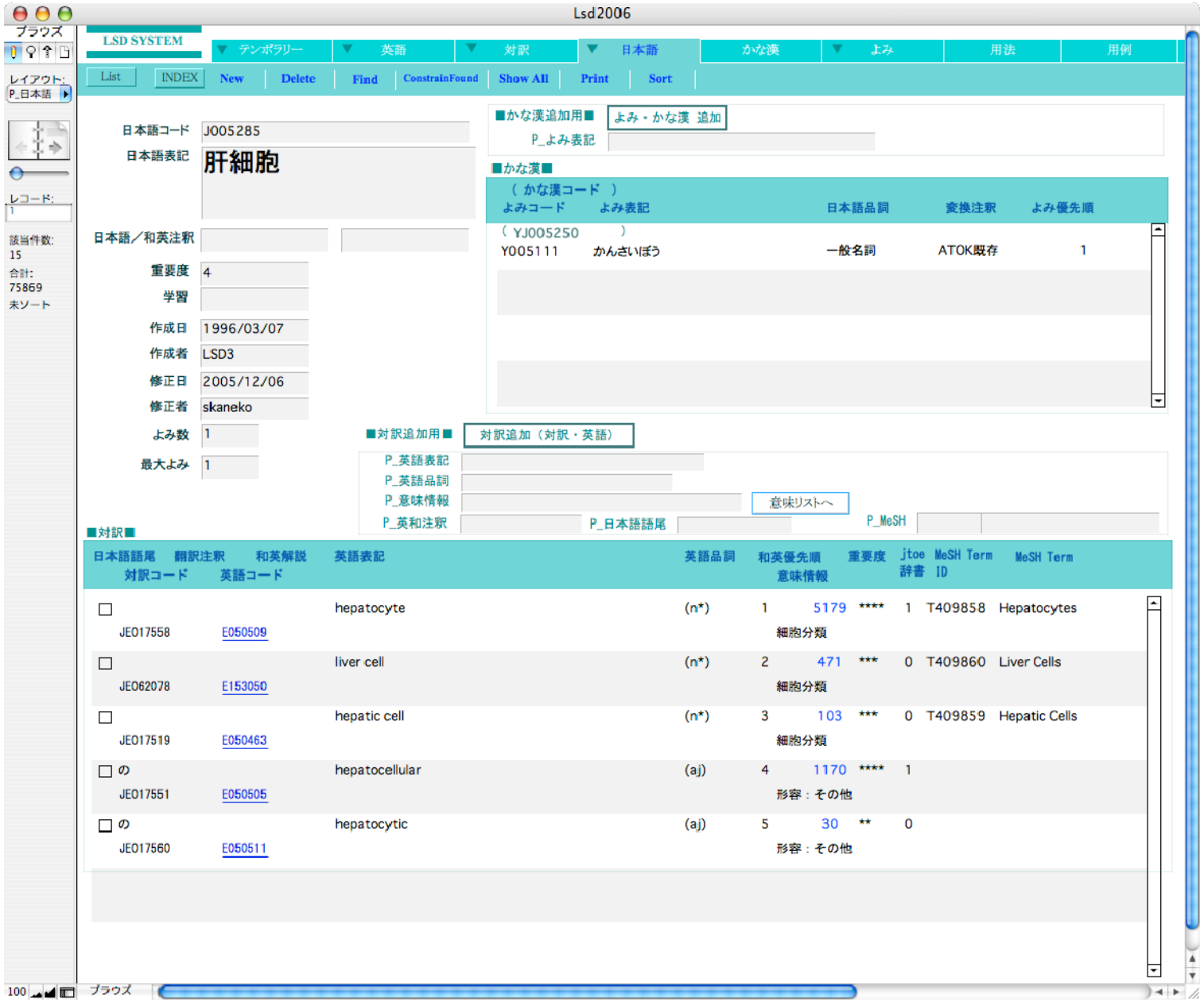
LSD英語　　LSD日本語　　MeSH Descriptor, Term と Tree

MeSH term ベースでの網羅率

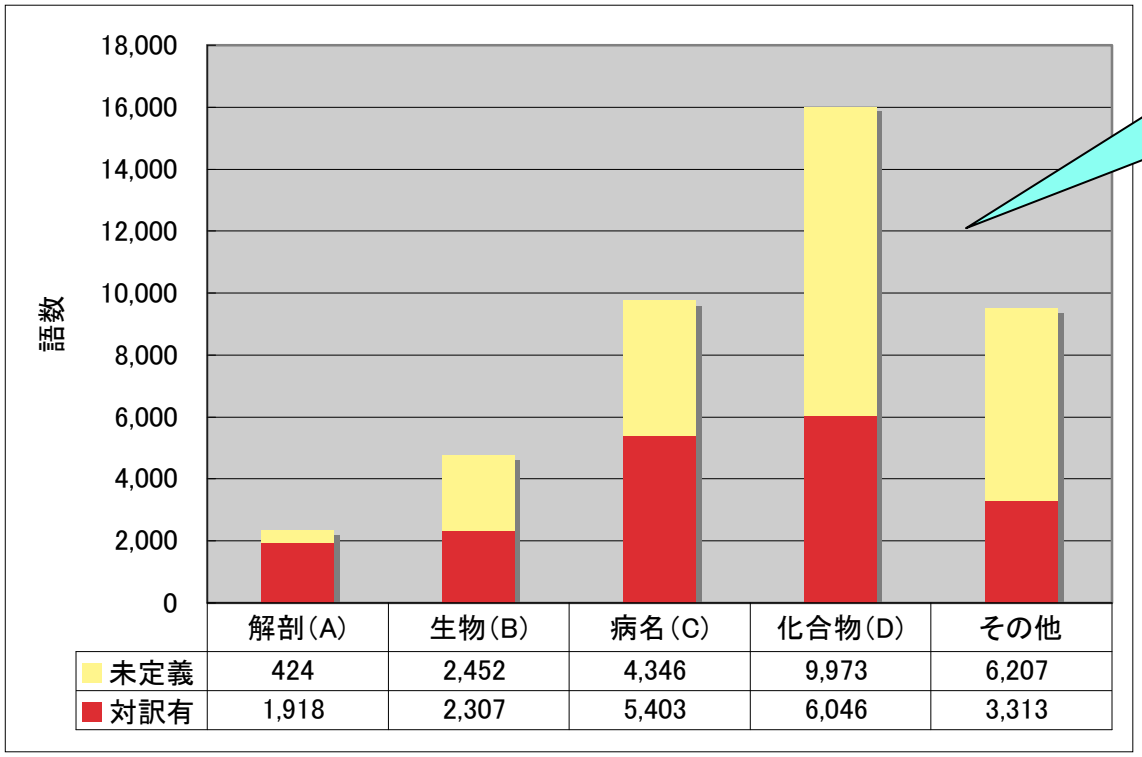
分類 【 MeSHカテゴリー】	LSD英語 [A] 対訳数	MeSH* [B]	共通語* [C] = [A]∩[B]	カバー率* [D] = [C]/[B]	未マップ語 [E] = [A]-[C]
Anatomy【A】	4,102 4,970	2,576 (2,024)	1,604 (1,528)	62% (75%)	2,498
Organisms【B】	2,505 3,101	5,635 (3,254)	1,215 (1,193)	22% (37%)	1,290
Diseases【C】	5,403 7,400	12,095 (6,338)	3,700 (3,604)	31% (57%)	1,703
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

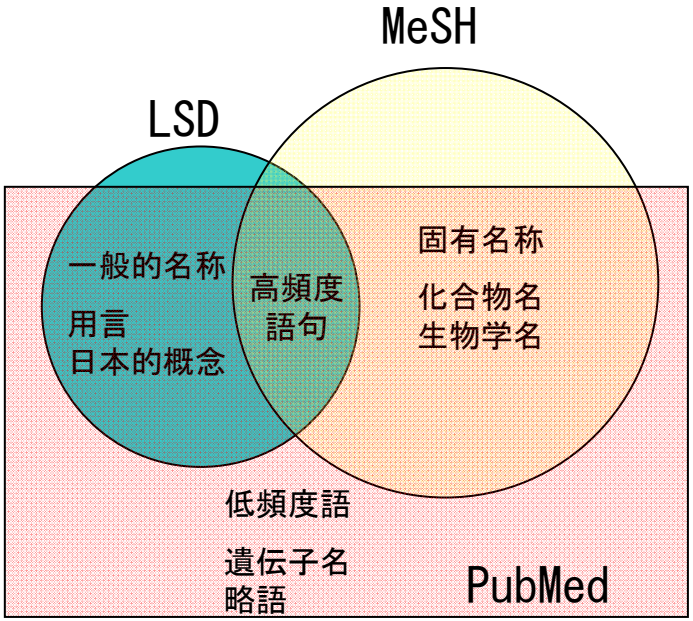


MeSH ツリーベースでの LSD 網羅率



	解剖(A)	生物(B)	病名(C)	化合物(D)	その他
未定義	424	2,452	4,346	9,973	6,207
対訳有	1,918	2,307	5,403	6,046	3,313

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）