厚生労働科学研究費補助金
（地域医療基盤開発推進研究事業）

テキストマイニングによる薬物有害事象の自動抽出を目的とした オントロジー構築とシステム開発に関する研究

平成18～20年度　総合研究報告書

Annual Report
Grant-in-Aid for Research on Medical Safety and Medical Technology Evaluation
Supported by the Ministry of Health, Labor and Welfare, Japan in 2006-2008
(Chief Researcher: Shuji KANEKO, Ph.D.)

平成21年(2009)年3月

主任研究者　金子　周司

厚生労働科学研究費補助金
（地域医療基盤開発推進研究事業）

テキストマイニングによる薬物有害事象の自動抽出を目的とした オントロジー構築とシステム開発に関する研究

平成18～20年度　総合研究報告書

Annual Report
Grant-in-Aid for Research on Medical Safety and Medical Technology Evaluation
Supported by the Ministry of Health, Labor and Welfare, Japan in 2006-2008
(Chief Researcher: Shuji KANEKO, Ph.D.)

平成21年(2009)年3月

主任研究者　金子　周司

目　　次

厚生労働科学研究費補助金（地域医療基盤開発推進研究事業）
総合研究報告書

テキストマイニングによる薬物有害事象の自動抽出を目的とした オントロジー構築とシステム開発

主任研究者　金子周司　京都大学大学院薬学研究科生体機能解析学分野

研究協力者：大武 博（京都府立医科大学医学研究科），藤田信之（製品開発技術評価機構）

[研究要旨]

ライフサイエンス辞書（LSD）は，広範な生命科学の論文や教科書で用いられる英語（9.5 万語）および日本語（10.5 万語）の専門用語を，用語の出現頻度を重要性の目安として選出した対訳辞書である。本研究では，LSD に収録されている名詞のうち，病名/症候名，薬物/生体内分子名，解剖/発生部位名，生物学名，方法や研究技術，概念や現象を意味する日英約 14 万語について，用語の同義性や上下関係を整理し，既存の専門用語シソーラスである MeSH と動的に関連づけ，随時更新できるリレーショナルデータベースを完成させた。また，公開されている病名分類（ICD-10 準拠標準病名マスターおよび ICH 国際医薬用語集 MedDRA/J）や薬効分類（WHO ATC および MeSH Pharmacological Actions）へのリンクも動的に設けることで用語に意味と属性を付与した。用語を補充した結果，ツリー状に整理した 2.5 万語の統制語に日英約 18 万語の専門用語を割り当てた LSD シソーラスを完成させた。また，専門用語の共起関係を医学論文から求め，統制語間の関係性を数値化したオントロジーを制作し，データを公開した。次に，これら語彙資源を有害事象の自動抽出に応用するため，FDA が公開している副作用報告システム AERS に収録された世界中の医薬品名についてほぼすべての名前解決を行える辞書を制作した。最後に，医薬品添付文書のテキスト解析によって，自由記述される医療情報から正しく医薬品名および疾患・症状名を抽出できるかを詳細に検討した。この LSD シソーラスを利用したテキスト処理は，医療文書からの有害事象の検出に極めて有用な手段であるのみならず，医療情報の解読や入力エキスパートシステムに応用できる優れた方策になると考えられる。

A. 研究目的

本研究は，ゲノム科学における情報科学的手法として発展・応用されつつあるテキストマイニング技術を医薬品の副作用（有害事象）のレポートや医療情報の解析に最適化し，日本語と英語を網羅する医療関連の用語オントロジーをテキスト解析エンジンに実装して，その評価を行いつつ，実効性のある情報解析システムを開発することによって，情報電子化時代を迎える医療における効率良く確かな安全体制の実現を，情報技術的に支援することを目的とした。

本研究は，過去 15 年にわたって広範な生命科学の論文や教科書で用いられる英語と日本語をそれぞれ用語の出現頻度を重要性の目安として選んできた対訳辞書であるライフサイエンス辞書（LSD）を構造化してオントロジーとするとともに，テキストマイニングを簡便に行うためのプログラムの開発を中心に据えて行ってきた。

研究初年度にあたる 18 年度は，テキストマイニングを行うために必要な用語について，統制語を定め，同義語を集約することを第一の目標とした。また，構築した辞書を用いたテキストマイニングエンジンの試作および高速化を行い，情報抽出に向けての第一段階を実現した。次に 19 年度は，マイニング辞書について，統制語を概念の上下関係が記述された頑強なシソーラスに配置して，実用レベルにすることを目標とした。また，LSD とシソーラスツリーを独立したデータベースとして管理することで，動的な関連づけを可能にし，将来にわたってデータ更新やメンテナンスが簡便にできるような基本設計を行った。さらに，構築した辞書を用いたテキストマイニングエンジンの制作と評価を行い，情報抽出に向けての具体的ステップを確認した。最終 20 年度は，構築したシソーラス辞書について，副作用情報の宝庫である米国 FDA の Adverse Event Reporting System (AERS)として公開されている膨大なデータを用いて，構築した辞書の網羅性を確認するとともに，我が国で市販されているすべての医薬品の添付文書を整理した財団法人日本医薬情報センター JAPIC の医療用医薬品解説の全データをテキストマイニングの材料として実際に解析を行い，解読の実効性と残された課題およびその解決のための方策について検討した。

研究はほぼ計画通り進み，後述するように実用レベルの辞書と，高速なマイニングエンジンの開発を完成できた。それらの詳細について，以下に説明していく。

B. 研究方法

1．LSD シソーラス構築

LSD には，生命科学の学術論文で用いられる専門用語が英語，日本語の見出し語がそれぞれ 9.5 万および 10.5 万語収録されている（2009 年 3 月現在）。このうち，医療情報を解読するために重要な意味をもつ病名および症候名，医薬品・化合物名，生体内分子名，解剖・発生部位名，生物学名，方法・手法，尺度，現象や概念を意味する日英あわせて 14 万語について，基本的には自作の Perl スクリプトによる集計と手作業での修正によって同義語 synonym を整理し，代表的表記としての統制語 descriptor を定め，統制語の上下関係をツリー状のシソーラス thesaurus に整理し，統制語の関連性に基づくオントロジーontology の構築を目指した。シソーラスとしては対象分野の広さから NLM が定期的に更新している MeSH に準拠することにした。その際，LSD に収録されている単数形で自然テキスト順の用語と，MeSH に多く見られる複数形および階層性を残した語順の表記を一致させるため，MeSH 用語を標準テキスト順に戻し（例：leukemia, acute → acute leukemia），かつ複数形を単数形に揃える Perl スクリプトを制作し，自動処理を施せるようにした。

2．オントロジー構築

医薬品に対する適応症と有害事象は，ともに病名・症候名によって記述される。これらを区別するためには，医薬品の薬効分類についての知識を医薬品名称と対応させることが必要である。そこで LSD シソーラスの統制語を用いて，医薬品の作用点データベースを作成した。また，医薬品名を WHO ATC 薬効分類および MeSH Pharmacological Actions に関連づけ，各薬物に薬効分類および化学構造タグを付与した。なお，既存の病名分類である ICD-10 準拠標準病名マスターおよび ICH 国際医薬用語集 MedDRA/J と関連づけた。

3．AERS による辞書の評価

米国 FDA が公開している世界規模の医薬品有害事象データベース AERS では，薬物処置を施した症例の適応症と有害事象の記述に MedDRA の Preferred Term (PT) が用いられて一貫性を保っているが，医薬品の名称には商品名や一般名など多様な表記のゆれを含むため情報の統計的な処理が難しい。そのため，LSD シソーラスから医薬品名を抽出し，4 万種類以上の名称を約 4,500 種類の統制語にマッピングした医薬品名称解決のための辞書を制作した。

次に 2004 年から 2008 年第二四半期までに公開された 4 年半分 150 万件の AERS レポートをリレーショナルデータベースに再構築し，医薬品名称の解決を試みた。なお，適応症や反応事象の解決は MedDRA ver.11 を用いた。

4．JAPIC 医薬品情報のテキストマイニング

本研究の目的は電子カルテなど，医療現場で実際に発生する医療文書からの知識抽出であるが，特定の症例に基づく実証を行うまでに，網羅的な知識抽出が行えるかどうかの検証を行う必要がある。そこで適応症や副作用の記述に特別な取り決めが無く自由記述がなされている医薬品添付文書を材料にテキストマイニングを行ってみた。

まず LSD シソーラスから，関係抽出のための辞書を試作した。英文および和文テキストに対して文中の専門用語を認識して自動的に統制語タグを付与し，タグづけされた統制語の頻度と統制語同士の共起頻度をカウントするスクリプトを制作した。

その上で，2008 年版の JAPIC 医療用医薬品データベースに収録された 18,517 件の医薬品添付文書テキストに対して，LSD シソーラスと自作スクリプトを用いて専門用語へのタグづけを行った。用語頻度の集計を行い，さらに一部分を抽出して適合率と再現率の評価を行った。

C. 研究結果

1．LSD シソーラス構築

LSD に収録された日英 20 万語のうち，医療情報の解読に必要となる概念としては，病名および症候名，薬物および生体内分子名，解剖・発生部位名，生物学名，方法や研究技術，現象や概念を意味する用語が必要十分であると考えられた。そこで本研究では 3 年間かけてそれらの概念を意味する日本語および英語の 14 万語について用語の同義性や上下関係を整理し，さらに既存の専門用語シソーラスである MeSH 2008 年版とリレーショナルデータベースで動的に関連づけた。また，公開されている病名分類や薬効分類へのリレーションを設けた。この結果，MeSH からの追加を含め，約 18 万語の専門用語を約 2.5 万語のツリー状に整理した統制語に割り当てた LSD シソーラスが完成した（表 1）。このデータから，18 万語の同義語は，LSD 収録の英語と日本語，および新たに加えた MeSH 英語が，それぞれほぼ 3 分の 1 ずつの割合を占めることがわかる。生体分子などの物質名，特に海外での医薬品商品名や化学一般名などの異表記を非常に数多く含む物質カテゴリにおいては，MeSH に由来する名称が半数に及んだ。

しかしながら一方で，LSD に収録されながら MeSH と照合できないため統制語に帰属されない用語が英語で 1 万語以上も存在することが明らかになった（表 2）。特に，病名・症候名や解剖部位名においては，国内で用いられている標準病名マスターや MedDRA にも収録されながら，MeSH に帰属できない用語が数多く残された（詳細については本報告書 p.23-26 参照）。また，医薬品としては国内医薬品において収録されていない用語が若干，存在した。今後は，これらの語句を帰属させるために統制語を拡張するとともに，複数のツリー体系に対応できるよう，データベース仕様を改良していく必要が示された。

表 1 ライフサイエンス辞書 LSD のシソーラス化(2009 年 3 月現在)

Tree[1]	カテゴリ	統制語数(a)	シノニム数(b)	平均異表記数 (b)/(a)	LSD 英語[2]		LSD 日本語		MeSH 由来	
A	解剖部位	1,557	7,237	4.6	3,376	47%	3,138	44%	723	9%
B	生物学名	3,531	18,532	5.2	6,254	34%	8,337	45%	3,941	21%
C+F03	病名・症候名[3]	4,404	28,678	6.5	10,080	35%	12,724	44%	5,874	21%
D+SC	物質名	12,675	104,612	8.3	24,990	24%	29,714	28%	49,908	48%
	(うち医薬品)[3]	(3,885)	(45,709)	(11.8)	(8,856)	(19%)	(12,038)	(26%)	(24,815)	(54%)
E	方法, 尺度	2,293	12,518	5.5	4,253	34%	5,549	44%	2,716	22%
G	知識, 現象	1,320	6,995	5.3	2,669	38%	2,886	41%	1,440	21%
計		25,780	178,572	6.9	51,622	29%	62,348	35%	60,079	36%

1) ツリー記号は MeSH Tree 2009 年版 (2008 年 11 月公開) に準じた。構築したツリーは資料編 1 に示した。

2) 百分率はシノニム数(b)に対する割合を表す。

3) 病名・症候名および医薬品に関しては, 構築したシソーラスを資料編 2,3 にすべて示した。

参照: **資料編1(T1〜T205 ページ) シソーラスツリー**
資料編2(D1〜D351 ページ) 病名シノニム
資料編3(C1〜C523 ページ) 医薬品シノニムおよび薬効分類

表 2 統制語に帰属できなかった LSD 収録語

カテゴリー	統制語に帰属されない語句		例
	LSD 英語	(日本語対訳)	
解剖部位	2,242	2,376	sacral cord, natural killer T-cell, iPS cell
生物名	757	807	Periplaneta japonica, avian influenza virus
病名・症候名	4,871	5,757	varicella zoster, ketoacidosis
物質名	2,719	3,163	hemoglobin A1c, mozavaptan
方法, 尺度	1,634	2,140	molecular imaging, chemoradiotherapy
計	12,223	14,243	

※ MeSH との照合は英語ベースで行ったため, 日本語対訳で帰属されなかった語数は参考値である。

2．オントロジー構築

医薬品と適応症との対応関係は JAPIC「添付文書記載病名集」などによって整理されつつある。本研究ではより分子作用に立脚した医薬品オントロジーを目指し，医薬品と生体の相互作用を数少ない相互作用様式とともに三項関係によって記述する新たなデータベースを構築した。

このデータを制作するため，すべての統制語について PubMed より代表的な学術誌に掲載された 10 年分の論文抄録(600 M バイト)を題材として，テキスト中に共起する統制語の頻度を収集し，関連概念としてデータベース化した。その中には医薬品の作用点あるいは薬効分類に属する用語が数多く集まっていたため，これらを材料にして手作業によるオントロジー化を行った。

LSD シソーラスに収録された薬理活性のある化合物の生体作用について，受容体，酵素，膜輸送タンパク質などの分子レベルでの相互作用が記述できる場合は，作用点となる生体分子を統制語あるいは RefSeq，酵素 EC 番号を用いて特定した。その相互作用様式については，阻害，活性化など少数の記述子を用いて三項関係として表現した。標的分子が明らかでない場合は，細胞ないし組織レベルにおいて報告されてきた知識を同じく統制語と作用様式を表す記述子を用いて整理した（詳細は本報告書 p.35-38 参照)。

これらの結果，ほとんどの化合物について，薬物・標的分子（あるいは標的細胞ないし組織）・相互作用様式の三項関係を記述することができた。さらに薬物を表す統制語は WHO ATC 分類および MeSH Pharmacological Actions と関連づけ，各薬物に薬効分類および化学構造タグを付与した。

一方，病名に関しても複数存在する病名の階層化に対応するため，既存の病名分類である ICD-10 準拠標準病名マスターおよび ICH 国際医薬用語集 MedDRA/J と関連づけた。

3．AERS による辞書の評価

AERS は米国 FDA が公開している医薬品の有害事象データベースである。自発報告であるが適応症と反応事象に対しては MedDRA による用語の統制がとれているため，医薬品名の解決が可能になれば有害事象のデータマイニングに有用な研究資源となる。そこで本研究では構築した LSD シソーラスでの医薬品の網羅性を検証するため，AERS の名前解決を試みた。

AERS レポート 150 万件に出現する医薬品を解析したところ，18 万種類の異なる名称が検出された。しかし 45,000 種類の医薬品名を 3,900 語の統制語に整理した LSD シソーラスとの関係づけを直接，加工せずに行った段階でも 80%以上のレコードで薬物の有効成分名を解決することができた。関係づけできなかったレコードの多くは，複数の医薬品名が記載されたり余分な注釈が記載されたレコードであったため，AERS の元データに適切な二次加工を施すことで対処が可能であり，最終的に AERS に出現する医薬品レコードの 96%（Primary Substance に限定すると 98%）が同定できた。また，すべての医薬品について，先に述べたように分子作用点，薬効分類，構造分類の属性が付与されているため，例えば，「ベンゾジアゼピン系催眠薬」，「フェノチアジン化合物」，といった，ある一群の医薬品を指す用語を用いても AERS の検索が可能になった。

なお，本研究で制作した AERS 医薬品名解決辞書はすでに JAPIC に導入され，有害事象検索サービスとして実用化されている。また，京都大学薬学部において制作している学生教育用医療データベース（薬学統合ナビゲーションシステム）においても本辞書を実装し，医薬品情報や教材に出現する医薬品名から AERS の有害事象を検索できるシステムを医療薬学教育に用いはじめたところである。（詳細は本報告書 p.44-48 参照）

4．JAPIC 医薬品情報のテキストマイニング

本研究の最終目標である医療文書のテキストマイニングによる有害事象の発見のためには，テキスト中のすべての専門用語が適切にもれなくタグづけられ，医薬品と病名・症状との共起関係が抽出される必要がある。

この処理を行うため，本研究では専門用語を最長一致で発見し，LSD シソーラスを用いて統制語タグを付与する Perl スクリプトを開発した。また，処理速度や内容の質を評価するため，本研究は文書の題材として英語用は世界標準の薬理学教科書 Goodman & Gilman's The Pharmacological Basis of Therapeutics 10th edition の全テキスト（8 M バイト），日本語用は JAPIC 医療用医薬品集（56 M バイト）を用意した。

スクリプトによる処理は，テキスト中に出現する専門用語に属性を持つ統制語 XML タグを付与し，さらに医薬品名と病名・症候名が 1 文中で共起した場合にそれらを抽出するようにした。これを用いた場合，市販のパソコン（Apple Mac Pro）でも 1 分間に約 7 M バイトもの大量のテキストから，医薬品と適応症や副作用にタグ付けを行い，それらの出現頻度や共起関係を計数することが可能であり，十分な速度が得られた。

次に，タグ付けテキストの内容をブラウザで確認しながら，曖昧性の排除と統制語の最適化を行った。この過程において，テキストでの一致のみによって統制語への変換を行う場合，曖昧性を排除するために多義性のある略語や商品名等，一部のシノニムをタグ付け辞書から除外する必要が生じた（約 200 語）。また，「ヒト human」，「病気 disease」，「酸 acid」等のように，非常に大きな概念は関連するキーワードとして不必要あるいは不適切と考えられたため，それら（約 360 語）もタグ付け辞書から除外した。

以上の最適化辞書を用いて，20 種類の医薬品添付文書に記載された副作用（有害事象）のテキストマイニングを行った結果，適合率，再現率ともに 90%以上の成績が得られた。

しかしながら，この解析の結果，医薬品添付文書やインタビューフォームに記載された相互作用や副作用情報は規制用語である MedDRA や標準病名等とは必ずしも一致せず，数多くの異表記を含むうえに，読み手の専門知識を前提とした記述で満ちていることが明らかになった。例えば，相互作用に注意すべき医薬品の呼び名として，「カルシウム拮抗剤」や「Ca 拮抗剤」のような表記のゆれとともに，「ジヒドロピリジン系薬剤」や「ニフェジピン等」などの表現があり，これが商品名「アダラート」を初めとする一群の「ニフェジピン」を中心とする一群の「降圧薬」との併用に注意を喚起している文章であることは，これらの関係を正確に解釈できることを前提としている。同様な表記の多様性は適応症や副作用についても指摘できる。幸いにも LSD シソーラスではそれらを関連づけるデータを有しているので，実用レベルのテキストマイニングにおいては名前解決を図るだけでなく，階層（粒度）の異なる用語を場合に応じて展開しながら解析を行う必要がある。

D.考察

医療情報化社会において医療等の安全を達成するためには，市販後の医療情報や調査データを解析することによる有害作用の知識発見を早期に，確実かつ網羅的に行う必要性がある。そのような医療情報のほとんどは，文章（テキスト）として記述される。医療現場において今後，急速に電子化が推進され，大量のテキスト情報が発生すると予想できる。しかし医療情報を記述する用語については，病名，医薬品名などで国際的協調によって表記の統一化の努力が続けられているが，実際に FDA 等で公開されている医療テキストを解析すると，様々な用語が統一されずに用いられ

ていることが本調査研究から分かってきた。我が国において状況はさらに深刻であり，医薬品副作用報告でMedDRA/Jなどの規制用語が用いられているものの，医薬品添付文書を初めとして，それ以外の医療文書では英語以上に多種多様の日本語が使われているのが実態であった。これまでの用語集は表記や分類を統一する方針で制作されており，網羅性に問題がある。医療情報の解析を行うためには，自然言語処理によって英語と日本語の語彙を網羅し，かつ事物（医薬品等）や概念（病名等）の同義性や関連性をツリー状に整理したオントロジーを構築し，常に最新の状態で維持することが最優先の課題である。

テキストマイニング技術をゲノム科学に応用し，遺伝子と発現プロファイルや代謝パスウェイとの関係から創薬標的の発掘や遺伝子の機能推定を行おうとする研究は情報科学の領域で盛んに行われ，一部は商品化もされている。しかし，テキストマイニングを副作用情報の発見に応用しようとする試みはほとんどなかった。機能が不明な遺伝子の機能推定とは異なり，薬品名や症候名は（表記は統一されていないものの）限られた数の語彙から構成されており，本来コンピュータによるテキストマイニング処理には適した材料である。しかしながら，過去に誰も着手しなかった最大の原因は，あらゆる文書での日本語・英語を網羅する「辞書」が存在しなかったためと考えられる。事実，医薬品名，遺伝子，病名の各用語はそれぞれ独自に国内外で規定されているが，これらの相互関係を多様なボキャブラリを含めて網羅的に記述したデータベースは今なお皆無である。専門領域に特化したオントロジーの構築は，検索エンジン技術を発展させる研究として情報科学でもきわめて注目されており，本研究で構築したLSDシソーラスは，医学オントロジーのプロトタイプとして有用な資源であると考えられる。

本研究によって，臨床現場から発生する大量の電子化された生の文書を早期に定量的に分析し，有害事象の早期発見を可能にするシステムの開発が可能性を帯びてきた。すなわち，様々な医療情報から有害事象が疑われるレポートを自動抽出し，人間による最終的な知識発見を支援する実用システムが完成されよう。

しかし日本語は表記のゆれが著しいため，日本語解析のためのスクリプトを形態素解析と融合させることによって改良する必要がある。また，シソーラス辞書は適合率，再現率90%以上の性能を有しているが，医療情報の解析において検索漏れは許されない。シソーラス制作時に1万語以上の該当漏れが生じたことからも明らかなように，今後さらに網羅性の高いシソーラスを完成させる必要がある。

本研究は，抽出される薬物（A）と有害事象（B）との関係抽出に留まらず，DNAアレイを用いた他の研究における薬物（A）と遺伝子（C）との関係と論理的に組み合わせることによって，有害事象（B）と遺伝子（C）との関係が示唆されることになり，副作用メカニズムを実験科学で立証するための着眼点が提案されうる。これらの諸観点から，本研究の成果はICT医療時代にあって極めて高い実現性と波及効果が期待される。

E. 結論

本研究によって，医薬品と疾患，症状に加えて，関連する技術や方法，解剖部位や生物名までを網羅した頑強な医学オントロジーがほぼ完成した。また，英語および日本語テキストについて文中でのキーワード共起解析を高速かつ簡便に行うための処理プログラムを開発し，テキストマイニングからデータマイニングへの橋渡しが可能であることを示した。

今後はさらに解析結果のフィードバックからシソーラス辞書の網羅性を高めると共に，共起解析の結果を二項関係から関連性の解読へと発展

させることで医療文書の解読や入力エキスパートシステムの構築に向けて製品開発が期待できる。その際，本研究で制作した医薬品と薬理作用点データベースに，JAPIC が制作した医薬品と適応症のデータベースを組み合わせることで，有害事象の判別が可能になると考えられる。実際の医療文書の解析と評価を繰り返すことによって，十分な実用性と有用性を有する医療情報システム設計が可能になると結論できる。また本研究は，電子カルテやオーダリングシステムにおいて医薬品添付文書の記載に基づく相互作用や禁忌症など使用上の注意に対する警告ないし助言を可能にする等，医療情報システムのインテリジェント化を推進するためにも有用な資源となると期待できる。

なお，本研究で構築した LSD シソーラスは，無料検索サービスとして京都大学サーバで公開している。今後，情報ポータルとしても有用性を高めていく予定である。

F.

G.

ライフサイエンス辞書の
シソーラス化とその応用

京都大学大学院　薬学研究科
金子 周司
skaneko@pharm.kyoto-u.ac.jp
2008年1月22日
統合データベースセミナー

LSDプロジェクト

NIFTY-Serve バイオフォーラム FBIO（1989年）
かな漢字変換辞書の公開
学術用語デジタル利用の予備調査（1992年）
内容が古い，死語が多い，電子化されていない
独自の専門用語辞書を制作する必要性
プロジェクト発足（1993年）
研究成果公開促進費や民間財団を財源とした活動
1．計量的な英文の解析に基づいて語彙を選択
2．電子辞書としての利用に最適化
3．表記や訳語を統一しない
教育研究支援のサーバ・ツール公開（1996年～）
WebLSDオンライン辞書，オンデマンド英語教材，EtoJ 逐語訳，かな漢字変換辞書

2008年1月22日 統合データベース

- 辞書制作
  - 金子周司（京大・薬）
- 技術開発
  - 藤田信之（製品評価技術基盤機構）
  - 越川義弘（宮教大）
- 教材作成，出版
  - 大武 博（京府医大）
  - 河本 健（広島大・歯）
- 評価，利用促進
  - 竹内浩昭（静岡大・理）
  - 竹腰正隆（東海大・医）

オンライン辞書サービス WebLSD

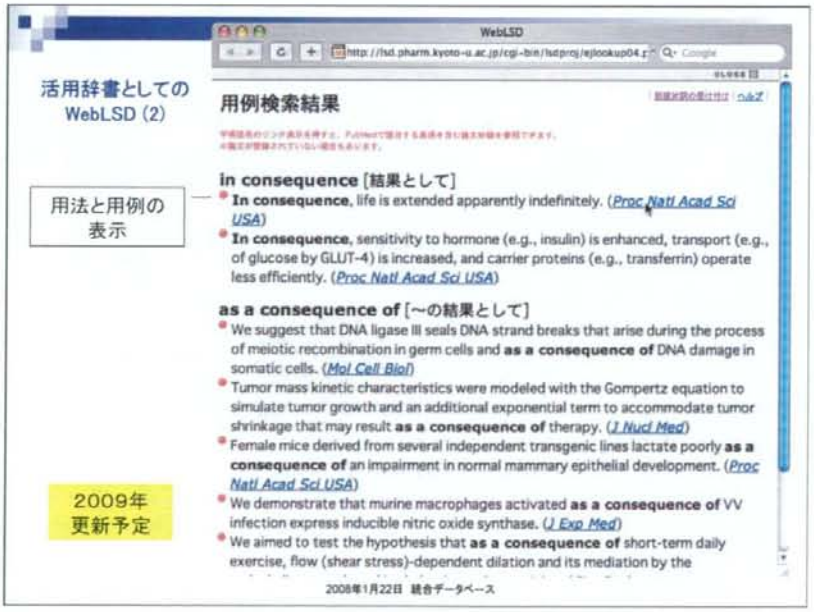
英和・和英・活用辞書

出現頻度，音声対訳，解説，関連語や用例を表示

WebLSD2007
英和 72,995語
和英 83,060語
音声 4,978語
例文 26,101文

2008年1月22日 統合データベース

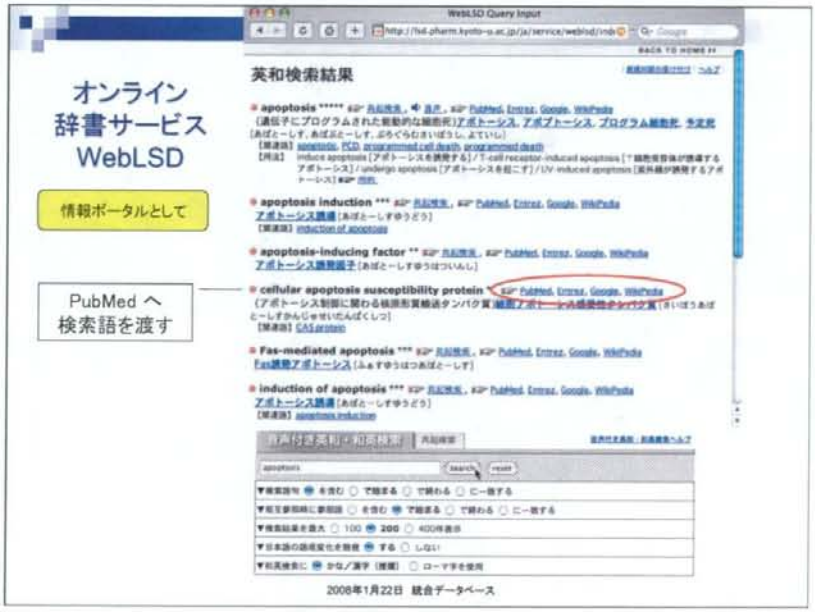

活用辞書としての
WebLSD (1)
用法・用例と
共起検索
を組み合わせて
標準的な英語を
知る
WebLSD Query Input
和英検索結果
2008年1月22日 統合データベース


活用辞書としての
WebLSD (2)
用法と用例の
表示
2009年
更新予定
WebLSD
用例検索結果
in consequence [結果として]
In consequence, life is extended apparently indefinitely. (Proc Natl Acad Sci USA)
In consequence, sensitivity to hormone (e.g., insulin) is enhanced, transport (e.g., of glucose by GLUT-4) is increased, and carrier proteins (e.g., transferrin) operate less efficiently. (Proc Natl Acad Sci USA)
as a consequence of [〜の結果として]
We suggest that DNA ligase III seals DNA strand breaks that arise during the process of meiotic recombination in germ cells and as a consequence of DNA damage in somatic cells. (Mol Cell Biol)
Tumor mass kinetic characteristics were modeled with the Gompertz equation to simulate tumor growth and an additional exponential term to accommodate tumor shrinkage that may result as a consequence of therapy. (J Nucl Med)
Female mice derived from several independent transgenic lines lactate poorly as a consequence of an impairment in normal mammary epithelial development. (Proc Natl Acad Sci USA)
We demonstrate that murine macrophages activated as a consequence of VV infection express inducible nitric oxide synthase. (J Exp Med)
We aimed to test the hypothesis that as a consequence of short-term daily exercise, flow (shear stress)-dependent dilation and its mediation by the
2008年1月22日 統合データベース


活用辞書としての
WebLSD (3)
共起リスト[1語前でソート]
6000万語の
PubMed corpus
から300例を
KWIC表示
2008年1月22日 統合データベース


活用辞書としての
WebLSD (4)
共起リスト[1語前でソート]
共起検索で
任意の語句の
前後関係を調査
2008年1月22日 統合データベース


隣接頻度の集計値
1. 正しい文法の理解を助ける
2. 制作ツールとしても内部使用
共起頻度の集計値
2008年1月22日 統合データベース


オンライン
辞書サービス
WebLSD
情報ポータルとして
PubMed へ
検索語を渡す
英和検索結果
2008年1月22日 統合データベース

情報ポータルとしての
WebLSD (1)

英和・和英から
PubMed 検索

2008年1月22日 統合データベース

情報ポータルとしての
WebLSD (2)

英和・和英から
Entrez 検索

2008年1月22日 統合データベース

情報ポータルとしての
WebLSD (3)

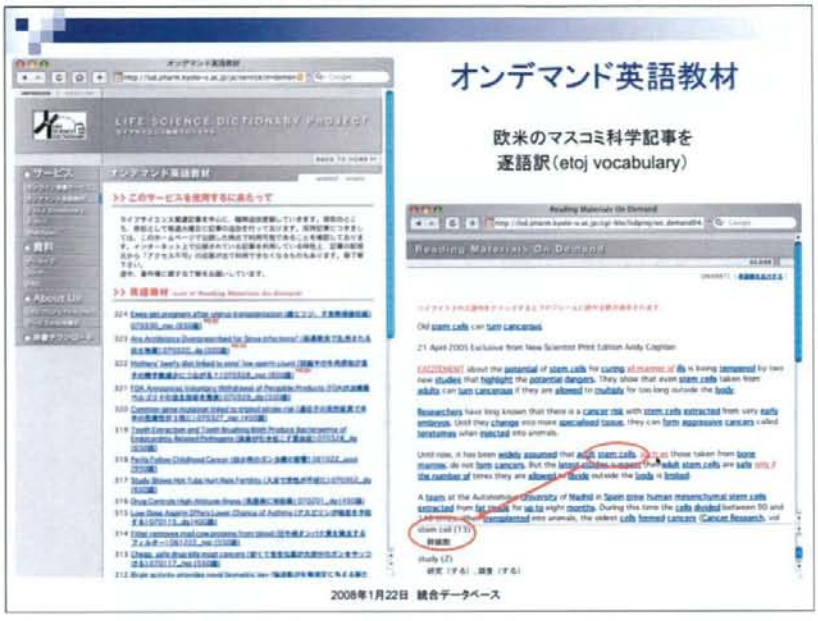
英和・和英から
Google 検索

2008年1月22日 統合データベース

情報ポータルとしての
WebLSD (4)

英和・和英から
Wikipedia 検索

2008年1月22日 統合データベース


オンデマンド英語教材
欧米のマスコミ科学記事を
逐語訳(etoj vocabulary)
LIFE SCIENCE DICTIONARY PROJECT
Reading Materials On Demand
2008年1月22日 統合データベース


EtoJ 逐語訳
視認性の良い日本語に置換
LIFE SCIENCE DICTIONARY PROJECT
E to J
2008年1月22日 統合データベース

解析に使用してきた生命科学コーパス

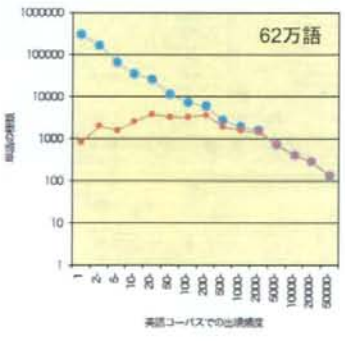
辞書の内容を高めるには，元のコーパスの質と量が重要

- 英語
  - PubMed収録のインパクトファクターの高い学術誌
  - アメリカ・イギリスの研究機関から発表された論文
    - 1994-2006 年の 100 誌の抄録
  - 一部，Bookshelf公開の教科書テキスト全文等も使用
  - 合計 368 Mbyte（6000万語）
- 日本語
  - 出版社の協力により提供された総説誌
    - 1996-2002年の全文
  - 臨床医学テキスト，医薬品データも収集
  - 合計 64 MByte（4000万文字）
- 現在は，それぞれ数倍規模に集積中

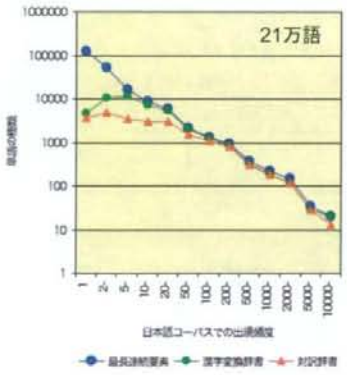
2008年1月22日 統合データベース

コーパスを用いた頻度解析

英語コーパス単語とLSD収録語（2006）



日本語コーパス語句とLSD収録語（2006）



2008年1月22日 統合データベース