

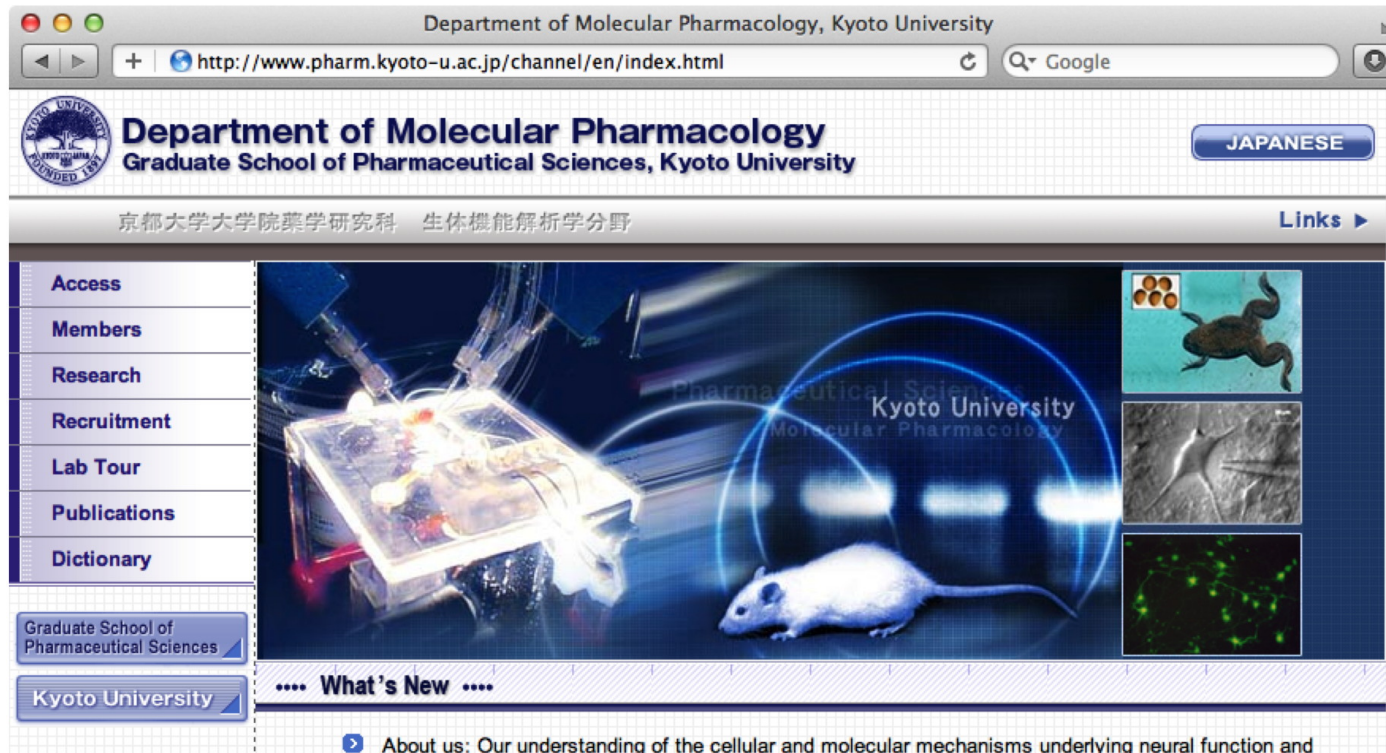
Finding Every Medical Term by Life Science Dictionary for MedNLP

Shuji Kaneko¹, Nobuyuki Fujita², Hiroshi Ohtake³

¹Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan

²National Institute of Technology and Evaluation, Tokyo, Japan

³Center for Arts and Sciences, Fukui Prefectural University, Fukui, Japan



Department of Molecular Pharmacology, Kyoto University

http://www.pharm.kyoto-u.ac.jp/channel/en/index.html

Department of Molecular Pharmacology
Graduate School of Pharmaceutical Sciences, Kyoto University

JAPANESE

京都大学大学院薬学研究科 生体機能解析学分野

Links ▶

Access
Members
Research
Recruitment
Lab Tour
Publications
Dictionary

Graduate School of
Pharmaceutical Sciences

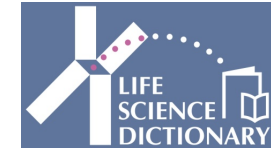
Kyoto University

.... What's New

About us: Our understanding of the cellular and molecular mechanisms underlying neural function and

The screenshot shows a web browser window with the URL <http://www.pharm.kyoto-u.ac.jp/channel/en/index.html>. The page header includes the department name and a 'JAPANESE' button. A navigation menu on the left lists 'Access', 'Members', 'Research', 'Recruitment', 'Lab Tour', 'Publications', and 'Dictionary'. The main content area features a large image of a laboratory mouse and a petri dish, with smaller images of a frog and a microscope. A 'What's New' section at the bottom contains a link to 'About us: Our understanding of the cellular and molecular mechanisms underlying neural function and'.

ABOUT US



- The **Life Science Dictionary (LSD) project**, founded in 1993, is a research project to develop a systematic database for life science (of course, including medical) terms and tools for the convenience of life scientists.
- Our services are designed to provide and encourage access within the scientific community to the most **up-to-date and comprehensive** information on English-Japanese translation dictionary of life science terms.
- In keeping with the users' expectations, we have been enriching and refining the database records to a **medical thesaurus** compatible with MeSH (Medical Subject Headings developed by National Library of Medicine, USA) thesaurus.
- Recent 2013 version of LSD contains approximately **30 thousand headings** with 200 thousand English and Japanese synonyms, consisting of the names of anatomical concepts, biological organisms chemical compounds, methods, disease and symptoms.

A screenshot of a web browser displaying the Life Science Dictionary website. The browser's address bar shows the URL "lsd.pharm.kyoto-u.ac.jp/webbsd/c/tree/D002289". The website header includes the title "ONLINE LIFE SCIENCE DICTIONARY" and navigation tabs for "Project", "WebLSD", "Reading", "EtoJ Vocab", "EtoJ", and "WebSpell". A search bar contains the query "Non-Small-Cell Lung Carcinoma". Below the search bar, the results for "LSD Thesaurus: 非小細胞肺癌 Non-Small-Cell Lung Carcinoma" are displayed, including a list of synonyms in both English and Japanese, and a hierarchical concept tree.

AIM OF THIS STUDY

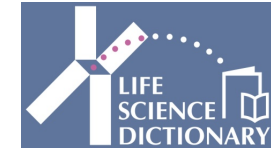


- One of the practical applications of thesaurus is **text mining**.
- For example, **adverse drug events** can be rapidly extracted by finding the causal relationship of drug treatment and related symptoms recorded in medical records.
Ex. Allergy + An antihistamine → arrhythmia = a side effect of the antihistamine
- Favorably, we have previously developed a series of **gloss-embedding Perl scripts** for medical English texts.
- In this study, therefore we aimed to find every medical term (English or Japanese) as many as possible.
- As the source text, an NTCIR10 test set (1,121 sentences) was used.

A screenshot of a text editor window titled 'ntcir10.txt'. The window displays a list of Japanese medical sentences, each preceded by a line number from 681 to 707. The text describes a patient's medical history, including symptoms like arrhythmia and hypertension, and treatments such as etanercept and various antibiotics. The text is in Japanese with some English medical terms and abbreviations.

```
681 術後経過は順調で5月6日胸腔ドレーン抜去し、5月9日退院となった。
682 今回感染症をおこしており、入院後からetanercept一時中止とした。
683 感染については、PAPM/BPの1週間の投与でも改善なく一般細菌感染は否定的であった。
684 【入院後経過と考察】
685 心エコーで壁運動障害なく、左心不全症状、肺高血圧などを示す所見はなかった。
686 ③高脂血症
687 その他表在リンパ節腫脹なし、
688 #4. 感染性塞栓症
689 胸部X線写真：両側下肺野にスリガラス影を認める。
690 幼少；てんかん、
691 食道（剖検時）：Enterococcus fecalis3+ Candida albica
692 ns1+ E. coli 少数 Klebsiella pneumoniae少数。
693 側関節屈曲進展は視認可能。
694 #1. 胸部異常陰影、咳嗽→特発性器質化肺炎。
695 筋萎縮なし、
696 尿検査：異常所見なし、
697 圧痛なし。
698 ロセフィン®0.3g×2回/日の点滴投与とクラリス®の内服も併用した。
699 【主な入院時現症】
700 Microangiopathyとしては、
701 #1に対してTAXUSφ2.75×32#2にたいしてTAXUSφ3.5×24留置した。
702 【血液培養（2セット）】陰性。
703 【感染症】β-D-グルカン 6.0pg/ml, CMV (-)
704 長期コントロールとして、吸入ステロイド薬の導入が望ましいと考えたが、関節リウマチによる手指
705 変形があり、吸入器を使用できないことから、近医より処方されていたtheophylline (2
706 00mg) 1T1×内服継続に加え、tulobuterol hydrochloride 2mg
707 /日貼付、predonisolone (5mg) 1T1×内服を開始した。
708 Barre徴候：右で陽性（回内）左は陰性、
709 2月13日入院。
710 43歳ごろ健康診断で高血糖（詳細不明）指摘され、44歳時近医受診し薬物療法導入となった。
711 腫瘍マーカーNSE8.2ng/ml SYFRA1.0ng/ml ProGRP24.6pg
712 /ml CEA2.9ng/ml
```

METHOD – DICTIONARY



- **A tagger dictionary** was made from LSD database as an EUC text file, which contains approximately 200,000 rows and 4 columns:
 - (1) synonym strings
 - (2) subject heading strings (converged to 30,000 descriptors)
 - (3) category of term
 - (4) subject heading ID (from MeSH) for external reference link
- For the category of terms, all terms were classified and marked by one of the following categories according to the MeSH tree:
anatomy, biological, disease, molecule, method, and knowledge.

Index	Japanese Term	English Translation	Category	MeSH ID
173898	肝疾患	肝疾患	disease	D008107
173899	肝実質細胞	肝細胞	anatomy	D022781
173900	肝腫	肝腫大	disease	D006529
173901	肝腫大	肝腫大	disease	D006529
173902	肝腫瘍	肝腫瘍	disease	D008113
173903	肝循環	肝循環	knowledge	D008102
173904	肝傷害	肝疾患	disease	D008107
173905	肝障害	肝疾患	disease	D008107
173906	肝新生物	肝腫瘍	disease	D008113
173907	肝腎症候群	肝腎症候群	disease	D006530
173908	肝腎障害	肝腎症候群	disease	D006530
173909	肝腎不全	肝腎症候群	disease	D006530
173910	肝性ポルフィリン症	肝性ポルフィリン症	disease	D017094
173911	肝性昏睡	肝性脳症	disease	D006501
173912	肝性脳症	肝性脳症	disease	D006501
173913	肝星細胞	肝星細胞	anatomy	D055166
173914	肝静脈	肝静脈	anatomy	D006503
173915	肝静脈血栓症	バッド・キアリ症候群	disease	D006502
173916	肝静脈閉塞症	肝静脈閉塞症	disease	D006504
173917	肝静脈流出路閉塞	バッド・キアリ症候群	disease	D006502
173918	肝切除	肝切除術	method	D006498
173919	肝切除術	肝切除術	method	D006498
173920	肝線維症	肝硬変	disease	D008103
173921	肝臓腫	肝細胞腺腫	disease	D018248
173922	肝臓	肝臓	anatomy	D008099
173923	肝臓Xレセプター	肝臓X受容体	molecule	C469720
173924	肝臓X受容体	肝臓X受容体	molecule	C469720
173925	肝臓X受容体α	肝臓X受容体	molecule	C469720
173926	肝臓X受容体β	肝臓X受容体	molecule	C469720

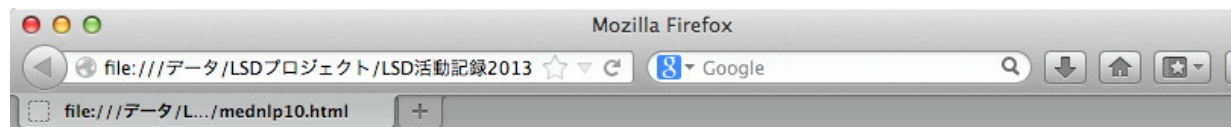
METHOD - PERL SCRIPT



- To take full advantage of the LSD in which many phrases have been registered, “**the longest matches first**” principle was adopted in the matching process.
- For this purpose, the tagger dictionary was sorted in the descending order of byte lengths, and text matching was performed for each of the dictionary entries in this order.
- Both the text and the dictionary were first converted to EUC, and they were treated as byte strings in the matching process.
- All two-byte roman characters were converted to corresponding ASCII characters.
- For better readability of result, **a standard HTML format** was used, in which “class” attribute was assigned to each of the category.

A screenshot of a web browser window displaying the source code of an HTML document. The title bar reads "tagged (130129提出バージョン) .html". The code includes a DOCTYPE declaration, meta tags for content language (ja) and charset (utf-8), and a CSS style block with various classes like .anatomy, .biology, .disease, .knowledge, .method, and .molecule. A JavaScript function showThes(meshID) is defined, which opens a new window to a URL containing the meshID. The body of the document contains HTML elements with class attributes and onclick events, such as "入院" (hospitalization), "心音" (heart sound), "鼻粘膜" (nasal mucosa), and "焼灼術" (cauterization).

THE OUTPUT



術後経過は順調で5月6日胸腔ドレーン抜去し、5月9日退院となった。
今回感染症をおこしており、入院後からetanercept一時中止とした。
感染については、PAPM/BPの1週間の投与でも改善なく一般細菌感染は否定的であった。

【入院後経過と考察】

心エコーで壁運動障害なく、左心不全症状、肺高血圧などを示す所見はなかった。

(3)高脂血症

その他表在リンパ節腫脹なし、

#4.感染性塞栓症

胸部X線写真:両側下肺野にスリガラス影を認める。

幼少;てんかん、

食道(剖検時):Enterococcus fecalis3+ Candida albicans 1+ E.coli 少数 Klebsiella pneumoniae少数

側関節屈曲進展は視認可能。

#1.胸部異常陰影、咳嗽→特発性器質化肺炎

筋萎縮なし、

尿検査:異常所見なし、

圧痛なし。

ロセフィン(R)0.3g×2回/日の点滴投与とクラリス(R)の内服も併用した。

【主な入院時現症】

Microangiopathyとしては、

#1に対してTAXUSφ2.75×32#2にたいしてTAXUSφ3.5×24留置した。

[血液培養(2セット)]陰性

[感染症]β-D-グルカン 6.0pg/ml,CMV(-)

長期コントロールとして、吸入ステロイド薬の導入が望ましいと考えたが、関節リウマチによる手指変形があり、

吸入器を使用できないことから、近医より処方されていたtheophylline(200mg)1T1×内服継続に加え、

tulobuterol hydrochloride2mg/日貼付、prednisolone(5mg)1T1×内服を開始した。

Barre徴候:右で陽性(回内)左は陰性、

2月13日入院。

43歳ごろ健康診断で高血糖(詳細不明)指摘され、44歳時近医受診し薬物療法導入となった。

腫瘍マーカーNSE8.2ng/ml SYFRA1..0ng/ml ProGRP24.6pg/ml CEA2.9ng/ml

- From the 0.1 MB test document, **2,569 terms** (including English spellings) were tagged and isolated in 2 min by personal PC.

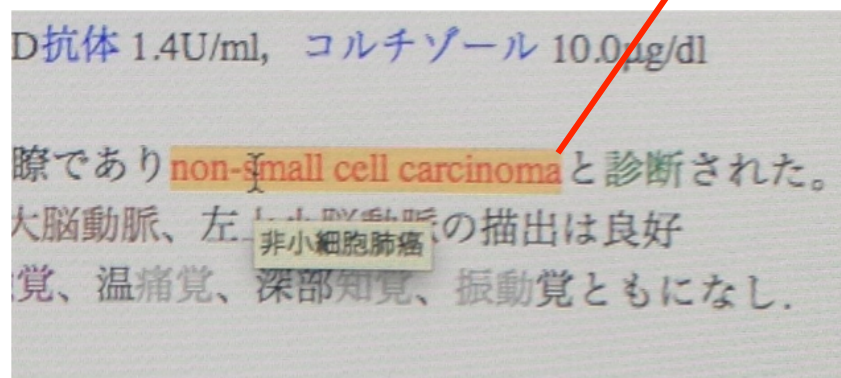
Table 1 Number of tagged terms

Category	Tagged
Anatomy	439
Biological	35
Disease (or Symptom)	893
Molecule (or Drug)	395
Method (or Index)	622
Other knowledge	185
Total	2,569

EASY REFERENCE



- We also added a ‘**mouse-over heading**’ feature, in which the embedded subject heading of the term will be displayed when the cursor was placed over the tagged term.
- In addition, by clicking the tagged part, the user can confirm the thesaurus entry in our WebLSD online
- This allows users to judge the justness of tagged terms.



click

LSD Thesaurus: 非小細胞肺癌 Non-Small-Cell Lung Carcinoma

lsd.pharm.kyoto-u.ac.jp/webstd/c/tree/D002289

Mirror [Kyoto 1 | Kyoto 2 | Tokyo] Font [S | M | L | XL] Width [Fixed | Variable] About cookies JAPANESE

ONLINE LIFE SCIENCE DICTIONARY
ライフサイエンス辞書オンラインサービス

Project WebLSD Reading EtoJ Vocab EtoJ WebSpell

Eng/Jpn Thesaurus Concordance Options ON/OFF

Query: **Non-Small-Cell Lung Carcinoma** search clear

LSD Thesaurus: 非小細胞肺癌 Non-Small-Cell Lung Carcinoma

非小細胞肺癌 (Non-Small-Cell Lung Carcinoma) を Google Scholar, Entrez, Google, Wikipedia で検索

同義語 (異表記) :

- Non-Small Cell Lung Cancer
- Non-Small Cell Lung Carcinoma
- non-small-cell cancer
- non-small-cell carcinoma
- non-small-cell lung cancer
- Nonsmall Cell Lung Cancer
- nonsmall-cell cancer
- nonsmall-cell lung cancer
- nonsmall-cell lung carcinoma
- NSCLC
- 非小細胞癌
- 肺非小細胞癌
- 非小細胞がん
- 非小細胞性肺癌
- 非小細胞肺癌

概念ツリー :

- 腫瘍 Tumor
 - 発生部位別腫瘍分類 Neoplasms by Site
 - 胸部腫瘍 Thoracic Tumor
 - 気道腫瘍 Respiratory Tract Neoplasm
 - 肺癌 Lung Cancer
 - 気管支腫瘍 Bronchial Neoplasm
 - 気管支原性肺癌 Bronchogenic Carcinoma
 - 非小細胞肺癌 Non-Small-Cell Lung Carcinoma

- 呼吸器疾患 Respiratory Tract Disease
- 肺疾患 Lung Disease
 - 肺癌 Lung Cancer
 - 気管支原性肺癌 Bronchogenic Carcinoma
 - 非小細胞肺癌 Non-Small-Cell Lung Carcinoma
- 呼吸器疾患 Respiratory Tract Disease
- 気道腫瘍 Respiratory Tract Neoplasm
 - 肺癌 Lung Cancer
 - 気管支腫瘍 Bronchial Neoplasm
 - 気管支原性肺癌 Bronchogenic Carcinoma
 - 非小細胞肺癌 Non-Small-Cell Lung Carcinoma

連想検索 (共起語との組み合わせでウェブを検索) :

以下のリストで

日本語をクリック → 非小細胞肺癌 との組み合わせでGoogleを検索

英語をクリック → Non-Small-Cell Lung Carcinoma との組み合わせでEntrezを検索

上皮増殖因子受容体 (Epidermal Growth Factor Receptor); 小細胞肺癌 (Small Cell Lung Carcinoma); 株化細胞 (Cell Line); 肺癌 (Lung Cancer); 腫瘍 (Tumor); ゲフィチニブ (gefitinib); 薬物治療 (Drug Therapy); カルボプラチン (Carboplatin); エルロチニブ (erlotinib); 治療 (Therapeutics); パクリタキセル (Paclitaxel); タンパク質チロシンキナーゼ (Protein-Tyrosine Kinase); 肺 (Lung); シスプラチン (Cisplatin); 第II相臨床試験 (Phase II Clinical Trial); 腫瘍転移 (Tumor Metastasis); 放射線療法 (Radiotherapy); ras遺伝子 (ras Gene); 腺癌 (Adenocarcinoma); 薬物毒性 (Drug Toxicity); 異種移植 (Heterologous Transplantation); ドセタキセル (docetaxel); 扁平上皮癌 (Squamous Cell Carcinoma); 白金 (Platinum); アポトーシス (Apoptosis); シクロオキシゲナーゼ-2 (Cyclooxygenase-2); ゲムシタビン (gemcitabine); 抗癌薬 (Anticarcinogenic Agent); ヘテロ接合性消失 (Loss of Heterozygosity); 細胞増殖 (Cell Proliferation); 無増悪生存期間 (Disease-Free Survival); メチル化 (Methylation); 乳癌 (Breast Cancer); 腫瘍抑制因子p53 (Tumor Suppressor Protein p53); 予後 (Prognosis); リスク (Risk);

MISSED TERMS

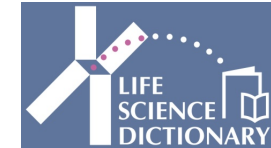


- In addition to many correctly-tagged terms, several patterns of missed or incorrect tags were found.
- The mostly missed terms were **English abbreviations**.
- Especially, in the description of clinical test data, a variety of abbreviations and acronyms were used, which cannot be marked.
- Since the meanings of 2- or 3-word abbreviations are ambiguous, we had omitted most of the abbreviations from tagger dictionary.
- However, if we know the part of document is apparently indicating clinical data, we can make a specific tagger dictionary for clinical tests.

Table 2 List of missed abbreviations

Subcategory	Examples
Clinical test	T-Chol, Hb, Plt, eosino, BP, MPO, PaCO ₂ , ALT, Cre, T-Bil, ZTT, APTT, etc.
Drug name	DIC (ダカルバジン) CLDM (クリンダマイシン) PIPC (ピペラシリン) PAPM/BP (パニペネム・ベタミプロン合剤)

WRONG-TAGGED TERMS



- The most typical pattern of incorrect tag was ‘partly-tagged’ term.
- In these cases, part of unit concepts were registered in the dictionary.
- However, the combination of two or more concepts is common particularly in the names of disease and symptom, which were not completely covered in our thesaurus.
- In these cases, we have to expand our dictionary.

Table 3 Examples of partly-tagged words

Partial

温痛覚
顔面紅斑
日光過敏
剥離爪

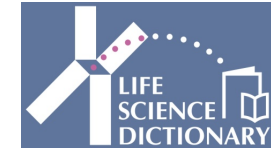
Compounded

Murphy徵候
心音不整
眼球結膜黄染
肺MAC症

More complex case

眼球の黄染
前頸部の腫脹
胆嚢軽度腫大
下肺にはhoney comb

MISSPELLING AND TYPO



- To our surprise, there were many misspellings and typographical errors, even in Japanese terms, in the test document.
- Precise text matching did not tag incorrect spellings that medical doctor can recognize their meanings.

Table 4 List of misspellings

In the text (**Wrong**)

predonisolone
theophyline
Mycobacterium abcessus
Enterococcus fecalis
Klebsiella pneumonoae
コルトコフ音
グルドパ
クオンテェンフェロン

Correct

prednisolone
theophylline
Mycobacterium abscessus
Enterococcus faecalis
Klebsiella pneumoniae
コロトコフ音
グルトパ (Grtpa)
クオンティフェロン

SUMMARY



- With our tagging dictionary and Perl scripts, most of medical terms were easily marked and visualized as an HTML document.
- From the 0.1 MB test document, 2,569 terms (including English spellings) were tagged and visualized in a color HTML format.
- Additional ‘mouse-over heading’ and web reference enables easy reviewing of the tagged terms.
- Through this task, we have learnt the potential and limitation of our thesaurus and scripts in finding medical terms from given Japanese texts.
- This process has a limitation in assigning **ambiguous abbreviations** and **misspelled words**.
- Moreover, there is an insurmountable difficulty to accomplish a ‘perfect matching’ with a fixed text dictionary, since **improvement of thesaurus is a laborious work**.
- The simple tagging strategy might be useful **in preprocessing of medical reports**.
- Combination of natural text processing with this tool will be convenient for the practical use.