社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (4)
2006－ＮＬ－175 (4)
2006／9／12

擬態語・擬音語に共起する語彙の感覚的分類に関する研究

石田博子† 小野木雄三††
† 東京大学医学部
†† 東京大学大学院医学系研究科
クリニカルバイオインフォマティクス研究ユニット臨床情報工学部門

一般向け医学サイトで病名を精度良く検索するには、医学用語が必要とされる。しかし、一般ユーザーは医学用語を使用しないため、検索語は日常会話内となる。そこで、日常の言葉で検索精度を向上させるため、自らの病名説明語のうち、受容器官で動機づけられた感覚における日本語での概念形成手段としての比喩表現である擬態語・擬音語に着目する。本研究では、病名解説文書を利用し自然言語処理により擬態語・擬音語と受容器・感覚との関係を得ることを目的とした。擬態語・擬音語との共起語の関連度を KL-距離(Kullback-Leibler Divergence)によって抽出し、神経生理学的感覚分類に基づき受容器・感覚を分類した。また、別の一般向け医学コーパスで本手法の分類によるカバー率を確認した。

Study of sensory classification as co-occurring tems with imitative word

Hiroko Ishida†, Yuzo Onogi††
†Faculty of Medicine, ††Graduate School of Medicine, the University of Tokyo

To retrieve a disorder name properly on medical site for general user, it still requires medical terms. While it is assumed that people use only ordinary spoken language for retrieval, user's ordinary expression as Japanese metaphor of sense such as an imitative word caused by reception in organ can directly influence effect. The purpose of this study was to obtain relation among imitative word, site and sense with disorder discription by natural language processing. The extract with relevance ratio between imitative word and co-occurring tems was made by KL-Divergence, and the classification was premised on neurophysiological clue. Then we checked how many terms could be covered with another data by this method.

1 はじめに

近年、日本においても「患者中心の医療」が注目されているが、患者の発話や診断時間の長さと患者の理解は乖離があることが報告されている[5]。また、英国においては、患者は外来受診の時やその事前に、多くの情報を必要としており、ますます増加した Web 上の医学情報量によって、個々の健康上の安心を得たり、診察前に独習で症状を認識しておくことは本人はもとより診察にも効果的であることが研究されている[4][6]。しかし、日本で最大級の病名登録数を有する医学検索サイト「Yahoo!ヘルスケア - 家庭の医学[19](以下 Yahoo)、メルクマニュアル医学百科最新家庭版[20](以下メルク)、家庭の医学-goo ヘルスケア[21](以下 goo)」を材料として用語比率を比較すると一般名詞数に対する医学用語の比率は 1.47 倍もある。ここで問題となるのは、一般ユーザーが選択する検索語は日常的に会話などで用いられる言葉である可能性が高いことである。特に疾患を生じている患者は、痛みなどの感情や感覚が優先し、身体現象を正確に客観的・論理的に説明することが一層困難になると考えられるためである。医学用語はその性質上病名を特定する確率が非常に高いのに対し、一般の言葉は病名を特定するものではないため、絞込み検索には適していない。

そこで本研究では、症状表現の補強について考える。感覚の表現根拠は受容器官由来の痛みであるとして、神経生理学的に定義されている感覚からのアプローチを行う。ここで感覚を概念化する手段としての比喩表現である擬態語・擬音語に着目する。これまでに、痛み質問表の擬態語・擬音語がどのような痛みと感じられるかを被験者 432 名への連想実験により強弱の分類を得ることが報告されている[9]。しかし、自然言語処理による感覚分類は報告されていない。そこで、本稿では病名や症状に関して臨床現場での経験が蓄積

されている解説文書という根拠を用いて、擬態語・擬音語と受容器、感覚との関係を得ることを目的とし、症状表現が擬態語・擬音語に比喩化された意味を特定することを試みる。以下、関係語抽出と分類の方法、評価実験、考察について順次述べる。

2　擬態語・擬音語と共起する語の抽出

共起語抽出用医学コーパスとして、先に述べた3サイトのうち擬態語・擬音語を含む文書のカバー率(全文書数に対する擬態語の異なり文書数)が最も高い(goo：1311/2085=62.9%、Yahoo：286/504=56.7%、メルク: 575/1202=47.8%)gooを用いた。まず、擬態語・擬音語切り出しを正確に行うことが重要であるため、本コーパスに対し日本語構文解析システム KNP(以下 KNP)[17][18]の文節出力を日本語形態素解析システム JUMAN(以降 JUMAN)[16]への入力文に用い、擬態語・擬音語を形態素へ切り出した(本稿では 2 文字以下の擬態語・擬音語は扱わない。切り出し判定には擬態語・擬音語辞書[15]を参照した。さらに 4 文字または 6 文字の繰り返し表現の単語についても擬態語・擬音語候補とした)。また、医学属性付与のために、形態素解析用の辞書として、医学用語シソーラス第5版、ICD10 対応電子カルテ用標準病名マスター、UMLS2005AA 中の日本語医学用語、過去のレポートから抽出された放射線用語辞書が追加された医学用語辞書を用いた。

次に共起語候補として上述の形態素語から、擬態語・擬音語に対する係り受け関係が外となる語や一語では意味をなさない語を除外する必要がある。しかし、例えば「背などにぶつぶつした隆起などの症状」のように擬態語・擬音語の近辺には助詞などの受容器属性分類に不要な語彙が連続して多出することや、部位や症状、病名などは並列に表現されることが多いため、数形態素語のみで外の係り受け関係を特定することは困難であった。また、これまでの研究で、関心対象と関心属性については定型単位からペアが収集されている[11]。そこで本研究では、[14]のように文意を重要視し、関心語である擬態・擬態語を中心として十分意味のとれる前後 5 形態素語を抽出の処理要素とした。

処理要素中の不要語の判定に、要素内に出現する共起語の出現確率に基づく関連度を用いる。ここで確率密度 f(X),g(X)を持つ共起語の出現確率 D(f(X),g(X))は一般的に一致度の指標によく用いられる KL-距離[13]により次の通り定義される。

$$D(f(X),g(X)) = \int f(X)\log\frac{f(X)}{g(X)}dX \qquad (1)$$

Dが0に近いほど共起語の関連度は高い。まず共起語 $w_2(i,j)$が擬態語 $w_1(i,j)$を中心として近傍位置列内(i:±1≦i≦5)の出現確率を求める。同一擬態語が例えば10行とした場合(10例,j:1≦j≦10)、w_2の図 1 中の領域 c での出現頻度を $n_c(w_1,w_2)$、領域 b での出現頻度を $n_b(w_1,w_2)$とすると出現確率は $f_{ij}(w_1,w_2)=n_c/(n_b+n_c)$、$g_{ij}(w_1,w_2)=n_c/(n_a+n_c)$である。このときのD値は次の通りである。

$$D(f_{ij},g_{ij}) = f_{ij}\times\log\frac{f_{ij}}{g_{ij}} \qquad (2)$$

これにより算出領域内外の距離による関連度が得られ、係り受け関係で外の関係が強い語彙や意味的に不要な語が含まれていても関連度を抑えることが可能となり、重要語と分離可能となる。

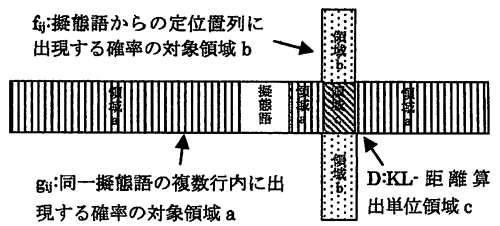

図 1　KL-距離の算出領域

次に不要語の除去を行う。閾値 t を設定して D>t の場合は除去し、残った語彙を空の除去位置に詰めるという処理を不要語がなくなるまで行う。本抽出の流れと抽出例を以下に示す。

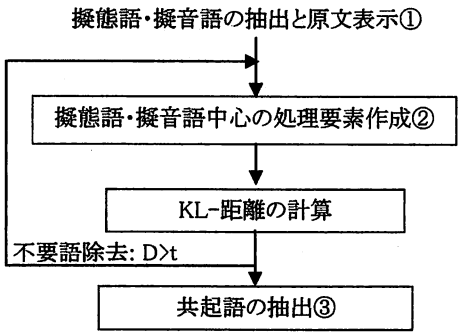

図 2　抽出の流れ

表 1　抽出例

処理	語　例
①	手指の関節の背面に1～2mm程度の表面が｜かさかさ｜して盛り上がった赤い発疹が集まり全体として
②	mm，程度，の，表面，が，かさかさ，する，盛り上がる，赤い，発疹，が
③	手指，関節，背面，mm，表面，かさかさ，盛り上がる，赤い，発疹，集まる，全体

擬音語・擬態語に後続する動詞は「する」が多い。このような語彙はD値算出領域内外ともiが正の小さい位置に出現する確率が高いため、D 値はプラスに高くなる。「痛み」、「発作」、「だ」のように出現位置を問わず使用頻度が高い語彙は、D値がマイナスに高くなる。また、特有表現であるx=-1:「表面」など、x=1:「音」などは、定位置に出現する確率が高く関心語に関連が大きいと考えられ、近傍位置で 0 に近くなる。このようにして不要語の除去と語句の再配置を2回繰り返し得た不要語を表 2 に示す。図 2 中、②から③での除去により、表 2③のように処理要素中の語彙情報は不要な語彙が減り、擬音語・擬態語に関連の深い語意の割合が増える。

表 2 抽出過程で得た不要語

不要語リスト
、,・。,ある,いる,か,から,が,こと,する,だ,で,と,ない,など,なる,に,ぬ,の,は,ます,も,もの,や,ようだ,れる,を,合併,時,症

3 抽出語に基づいた感覚の分類

分類については形態素属性が医学的部位に関する語に対し、自動的に分類する研究が行われているが[3]、本研究ではまず、症状の受容部として、神経生理学的感覚分類[7]を基に受容器属性(眼、鼻、耳、口、皮膚、関節・筋肉、臓器、神経)を設定し、文脈から各擬態語・擬音語に対し人手で属性を付与した。この属性付与ルールは、受容器が痛みの発生のおおもとの原因となるものとした。例えば、血液疾患が理由の精神的な「いらいら」の属性は、「神経」疾患ではないため、発生原因元である「臓器」とした。また、味覚や触覚によくみられる「ぼんやりした味」などの共感覚的比喩[10]は、脳疾患などではないためここでは取り扱わなかった。さらに情動感覚と前庭感覚(平衡感覚)は、それぞれの感覚が感情や周辺感覚にも依存しているため、本稿では区別せず、「神経」、「耳」とした。その結果、受容器属性とそれから逆参照される擬態語・擬音語と共起した語彙セット(擬態語・擬音語86語の共起語に受容器属性クラスを付与した1,753セット)を得た。

4 評価実験

医学コーパスgoo(異なり擬態語数86語)から求めた擬態語・擬音語と共起語の関連度と受容器属性語彙セットを用いて、擬態語・擬音語を受容器属性にクラス分類する処理を、別の医学コーパスのメルク(教科書形式:異なり擬態語数77語、内前述コーパスと一致する語数34語)に対して適用した。ここでは本来共起語(「背中」など)の共起語(「表面」、「内部」)など)によって「皮膚」や「関節・筋肉」にクラス分けする必要があるが、判定用閾値処理時には一語での判定として平均化した。また、閾値処理の対照手法として TF・IDF 法を用い、さらに提案手法として両者の論理積を適用した。

閾値は、KL-距離では、gooとメルクのKL-距離の各値(Dg,Dm)をデカルト座標に投影し、0 点からの距離(各値二乗和の平方根の平均)として0に近い側を正解とした。TF・IDF では、ある擬態語・擬音語における候補語Wの出現頻度をTF、Wを含む擬態語・擬音語数をNi、総擬態語・擬音語数をN(「しばしば」、「はっきり」のような感覚的比喩ではない擬態語・擬音語は除いた)としたときTF*log(N/Ni)として、TF・IDF 値の大きい側を正解とした。

この結果について、カバー率、正解率、一致率を比較したところ、擬態語・擬音語のカバー率(メルクの擬態語・擬音語に対して付与した受容器クラスと算出した同クラスに一致した擬態語数/2 つのコーパスで一致した擬態語数)は、28/34=82.3%であった。また、3 回の閾値設定による処理の結果、正解率(メルクの擬態語・擬音語に対して付与した受容器クラスと算出した同クラスに一致した数/算出した同クラスの数)は KL-距離で21/29=72.4%、TF・IDF で、22/31=71.0%、両者の論理積で、20/25=80.0%となった (表 3)。

表 3 正解率の閾値処理法による比較

閾値処理	正解率(%)
KL-距離	72.4
TF・IDF	71.0
KL-距離 AND TF・IDF	80.0

表 4 は、閾値処理別に同受容器クラス内で重要語の一致率(メルクで算出した重要語数と goo で重要とされた重要語の一致数比)を比較した結果である。また、閾値処理別に同受容器クラス内の重要語例を表 5、表 6 に示す。

表 4 重要語一致率の閾値処理法による比較

閾値処理	正解率(%)
KL-距離	21.6
TF・IDF	59.5
KL-距離 AND TF・IDF	70.2

表 5 KL-距離法での重要語例

擬態語・擬音語	重要語	受容器クラス
いらいら	うつ	関節・筋肉
いらいら	生活	臓器
いらいら	頭痛	神経
いらいら	不眠	神経
いらいら	落ち着き	神経
かさかさ	内側	皮膚
かさかさ	発疹	皮膚
ぜーぜー	チアノーゼ	臓器
ぜーぜー	気道	臓器
ぜーぜー	吸う	臓器
ぜーぜー	皮膚	臓器

表 6 TF・IDF 法での重要語例

擬態語・擬音語	重要語	受容器クラス
いらいら	集中	神経
いらいら	感	臓器
いらいら	不眠	神経
かさかさ	発疹	皮膚
かさかさ	かゆみ	皮膚
かさかさ	状	皮膚
ぜーぜー	呼吸	臓器
ぜーぜー	喘鳴	臓器
ぜーぜー	チアノーゼ	臓器

5 考察

評価実験の結果、閾値処理法別に比較すると、3 回の閾値設定では、正解率、一致率ともに本提案手法の 2 手法の論理積「 KL-距離 AND TF・IDF」が最も高かった。TF・IDF 法では重み付けされた語の判別根拠に何らかの他要因が必要であるのに対し、KL-距離による手法は処理領域内外の関係の優先度が判定根拠であるため設定しやすいと言える。

閾値判定の際、カバーする擬態語・擬音語数を優先してはいるが、今後正解率を向上させるためには、各語の共起関係について学習器を利用したフィードバック[12]を用いる可能性がある。また、受容器属性を付与する際、同一行内の全語彙について同一の属性としている点に問題があり、構文情報の解析など擬態語・擬音語の係り受け関係を利用することが考えられる。

本手法を適用した 2 つの医学コーパスから得られた受容器属性から特定される感覚・受容器クラス・擬態語・擬音語のセットと、同擬態語・擬音語のセットとなる感覚が唯一であるセットをまとめた結果を表 7 に示す。前述したように、発生原因によって分類したため、係り受けとしては考えにくいセットや、同一行内の全語彙について同一の属性としているため、複数の感覚に出現する擬態語・擬音語が見られた。これらについては、発生原因を絞り込んだ上、それが唯一の分類となるように係り受け関係を特定する必要がある。医学コーパスというテーマの固定された文章においても、使用する語彙表現が異なるため、擬態語・擬音語と共起語のセットが一箇所のみであったり、共起語が一致しない件数が多く見られたりした。これらは、同意味文脈の別表現と考えられるため、共感覚的比喩としての側面から解析し直す必要があるほか、同義語辞書を用いることも有効と考えられる。

実際に、今回得られた擬態語・擬音語と部位と、部位と医学用語とを検索語として、メルクのサイト内病名検索を 5 病名について行ったところ、擬態語・擬音語を用いた場合は、医学用語を用いた場合と同等またはそれ以上の検索結果を得た(表 8)。これは、症状という痛みの自覚的感覚認識が、他覚的観察結果である医学用語と、自覚的表現結果である擬態語・擬音語という異なる観察者・表現であっても、自然言語の機械的な処理によって、感覚分類の認識的な一致が可能であることが示されたと考えられる。つまり、医学用語を使用しない一般の人でも自覚症状については、擬態語・擬音語を用いることにより、医学用語と同等以上の説明語意を実現することが可能となると考えられる。さらに、症状の性質まで特定するには、時間要因(急性、慢性、時間帯など)・頻度要因・付随動作要因(動作、静止)・程度要因(強弱など)などについて、感覚以外の意味を持つ共起語から取得しなければならない。これまでに疼痛(痛み)の形容表現について研究が[8]のように報告されている。この報告では痛みが部位のほかに、性質、時間的変化、強さの 4 項目にわけられた質問表の形にまとめられている。性質については、20 群、78 語の形容表現(「flicking:ちらちらする」など)、情動的(「tiring:うんざりした」など)、評価的(「miserable:情けない」など)があげられ 5 段階評価がなされており、こうした臨床現場から得られた知見との意味的相関を求める必要もある。

本手法の結果から医学コーパス上の擬態語・擬音語という関心語を中心とした確率密度分布を基に、さらに語の受容器属性を付加することで、擬態語・擬音語側にも関連情報が投射される結果が得られた。しかし、感覚はそもそも必ず 1 部位に集約されるということはなく、複数の部位での関連情報も投射されている。つまり、表現化の際、擬態語・擬音語に限っては係り受け関係だけで意味情報を特定することが一義的に定義できないことからも、困難と思われる。関心語の周囲に分布する語彙情報を利用するというのが本手法の特徴の 1 つであり、係り受け関係は語彙の示す方向を決

定するという特徴を持っているので、双方の特徴を生かして、文法パターンの他に意味パターンも複合的に関連させて、採用要素位置の特定を行うというアプローチは必要であると考えられる。

6 おわりに

本研究では、擬態語・擬音語に共起する重要語抽出を行うことを目的とし、「擬態語・擬音語」と「感覚」との関係を抽出する KL-距離を用いた提案手法によって、それが達成された。次に、「受容器クラス」を付与し、擬態語・擬音語を「感覚クラス」に分類した。さらに擬態語・擬音語と感覚クラスが決まるとそれに付随する重要な形態素が列挙された。また、擬態語・擬音語と感覚クラスから得られた重要語セットを用いて、病名検索を実施し、症状の表現上、擬態語・擬音語は医学用語で特定することと同等またはそれ以上の検索効果があった。この結果は、擬態語・擬音語の感覚概念をオントロジー的に検索に適用すると、詳細表記である医学用語と同等相当に特定結果を得られたといえる。今後は感覚の他に関係があげられる機能語についても関係抽出を行うことにより、擬態語・擬音語の持つ性質概念が得られると考えられ、さらなる応用が期待される。

参考文献

1 今井健, 小野木雄三: 格フレームを用いた放射線読影レポートの文型分類と所見抽出. 第 24 回医療情報連合大会,pp.800-801 (2004)
2 今井健, 荒牧英治, 梶野正幸, 美代賢吾, 大江和彦: 構文情報と医学用語属性を用いた画像診断所見オントロジーの構築の試み. 医療情報学, 25(6), pp.395-403(2006)
3 荒牧英治, 今井健, 梶野正幸, 美代賢吾, 大江和彦: メタ関係を利用したテキストからの人体部位関係の抽出. 言語処理学会 第 12 回年次大会, pp.508-511 (2006)
4 Neilsen DM.,Gill K., Richketts DM.: Satisfaction levels in orthopadeic out-patients. Annals of the Royal College of Surgeons of England, 87(2), pp.106-108, Mar. (2005)
5 Takayama T., Yamazaki Y.: How breast cancer outpatients perceive mutual participation in patient-physician interactions. Patient Education & Counseling, 52(3), pp.279-289, Mar. (2004)
6 Larner AJ.: Use of internet medical websites and NHS direct by neurology outpatients before consultation. International Journal of Clinical Practice, 56(3), pp.219-221, Apr. (2002)
7 中村隆一編: リハビリテーション医学講座第 4 巻神経生理学・臨床神経学. 医歯薬出版株式会社(1985)
8 Ronald Melzack: The McGill Pain Questionnaire: Major Properties and Scoring Methods. Pain, 1, pp.277-299(1975)
9 楠見孝, 中本敬子, 子安増生: 痛みの比喩表現を支える身体感覚と形容語, 擬態語の構造. 日本認知科学会第 21 回大会発表論文集, pp.54-55(2004)
10 楠見孝: 味覚のメタファー表現への認知的アプローチ. 日本言語学会第 127 回大会, pp.9-14(2004)
11 阿辺川武, 奥村学. 形容詞を用いた対象・属性名詞対の収集および分析. 言語処理 学会第 12 回年次大会 (2006)
12 Paul Fabry, Robert Baud, Patrick Ruch, Christelle Despont-Gros, Christian Lovis: Methodology to ease the construction of a terminology of problems. International Journal of Medical Informatics, 75, pp.624-632(2006)
13 阿辺川武, 白井清昭, 徳永健伸, 田中穂積: 統計情報を利用した日本語連体修飾節の解析. 言語処理学会第 7 回年次大会, pp. 269-272, Mar. (2001)
14 清田 陽司, 黒橋 禎夫: WWW テキストの自動要約と KWIC インデックスの作成. 情報処理学会 (SIG-NL) 137-5, pp. 31-38, July (2000)
15 山口仲美: 暮らしのことば擬音・擬態語辞典. 講談社(2003)
16 日本語形態素解析システム JUMAN version 5.1 (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html)
17 日本語構文解析システム KNP version 2.0 使用説明書 (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html)
18 分類語彙表増補改訂版 (http://www.kokken.go.jp/katsudo/kanko/data/index.html)
19 Yahoo!ヘルスケア - 家庭の医学 (http://katei.health.yahoo.co.jp/katei)
20 メルクマニュアル医学百科最新家庭版 (http://mmh.banyu.co.jp/mmhe2j/index.html)
21 家庭の医学-goo ヘルスケア (http://health.goo.ne.jp/medical/index.html)

表 7 感覚と共起する擬態語との一覧[1)]

皮膚感覚	深部感覚	内臓感覚	神経系
皮 膚	関節・筋肉	臓 器	神 経
かさかさ,がさがさ, ごつごつ,ごろごろ, ざらざら,じくじく,じめじめ, じゅくじゅく,しわしわ, ずきんずきん,ちくちく, てかてか,ぱんぱん, ぴりっ,ひりひり,びりびり, ぶつぶつ,ぶよぶよ, ぷよぷよ,べとべと, ぼろぼろ,ぽろぽろ, ぼんやり,むずむず	いらいら,がんがん,くねくね,ぐらぐら,くりっ,ぐりっ, じーん,だらん,ばきっ, ばしっ,ばたばた,ぱちぱち, ぴくぴく,ひらひら, ぷちぷち,ぽこん, むにゃむにゃ,もぐもぐ, もじもじ,ゆらゆら	いらいら,ぎらぎら,きりきり, ぐったり,ぐりぐり,ごつごつ, こりこり,ごろごろ,さらさら, ざらざら,しゅっしゅっ, じんじん,ずきずき, ずきんずきん,すっきり, ぜいぜい,ぜーぜー, ちくちく,ちくり,てかてか, どきっ,どきどき,ぱちぱち, ひゅー,ひゅーひゅー, ぴりぴり,ぶーぶー, ぶつぶつ,ぼたぼた, ぽんぽん,むかむか, むくむく,もやもや	いらいら,うとうと,ぐったり, ぐらぐら,ぐるぐる,くるり, ざらざら,しっくり,すっきり, ぴりっ,ひりひり,びりびり, ふらふら,ぼーっ,ぼきっ, ぼんやり,ぽんやり
味 覚	聴 覚	嗅 覚	視 覚
口	耳	鼻	眼
つるつる,ひりひり, ぴりぴり,ぶつぶつ, ぷつぷつ	がらがら,ぐるぐる, ごーごー,さらさら,ぱちっ	さらさら,ねばねば	ころころ,さらさら, じんじん,どろっ, ぴかぴか,ぴりぴり, ぶつぶつ,ふわふわ, ぼんやり,ゆらゆら

表 8 (部位,擬態語)または(部位,対象語)を用いた病名検索の結果

擬態語	感 覚	部 位	対照語	病 名	部位－擬態語の順位	部位－対照語の順位
ごろごろ	臓器	腹部	膨張	乳糖不耐症	2	2
ぜーぜー	臓器	気道	喘鳴	喉頭蓋炎	8	15
ちくちく	皮膚	背	脳神経	慢性髄膜炎	30	36
ふらふら	神経	頭	疲れ	起立性低血圧	2	1
ぴくぴく	関節・筋肉	筋肉	部分発作	けいれん性疾患	1	1

1 網掛部は 2 つの医学コーパスで一致した擬態語と感覚のセットを示す。下線は擬態語と感覚のセットが唯一だったものである。