情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2009-NL-191 No.20
Vol.2009-SLP-76 No.20
2009/5/22

文字係り受けに基づく 専門用語の内部構造表現と解析

山田恵美子† 松本裕治††

本研究の目的は複合名詞である専門用語の内部構造解析である．本稿では内部構造を表現するためのタグセットと，これまで扱われていなかった言語現象を処理する方法としての文字単位の係り受け表現を提案する．また，既存の係り受け解析アルゴリズムによる解析を試みた結果，語に対して 88.2%の精度であった．

Internal Structure Representation and Analysis of Japanese Technical Terms based on Character-wise Dependency Relation

Emiko Yamada† and Yuji Matsumoto††

The aim of this paper is to present a method for structural analysis of technical terms, which are often in the form of compound nouns. We propose a tag set and an annotation scheme for representing internal structure of technical terms as dependency structure. Character-based dependency relations are introduced and enable to describe peculiar linguistic phenomena found in compound technical terms that involve coordination and contraction. Shift-reduce style dependency analysis achieved 88.2% word-level accuracy.

1. はじめに

特定分野の文章において専門用語は特有の意味を保持しており，それを知ることはその文章を解析する際に有用である．専門用語は複合名詞であるものが多く，したがって図１のような内部構造を持っている．内部構造とは語の構成要素とその結合の順序であり，これを知ることは意味を推測するうえで役に立つ．しかし専門用語は数多く存在しており，また動的に作られうるものであるため，予め全ての語に内部構造を記述しておくのは現実的ではない．本研究では，専門用語の内部構造で見られる複雑な現象を調べ，内部構造を表現する方法を提案し，それに基づいて専門用語を解析する．対象としたのは疾患名を中心とした生命科学分野の用語である．本稿では後で説明するように用語の内部構造を係り受け構造によって表現するが，一般の文で使われる係り受け構造だけでは表現しきれない構造として，後ろから前への係り受けや，「大腿骨折＝大腿**骨**＋**骨**折」「角結膜＝角**膜**＋結**膜**」のような縮退が起こる場合がある．このような構造を表現するのに必要な表現方法を提案し，これを文字単位の係り受けと捉えることが可能であることを示す．更に既存の係り受け解析アルゴリズムを用いた解析を試みたのでその結果を報告する．

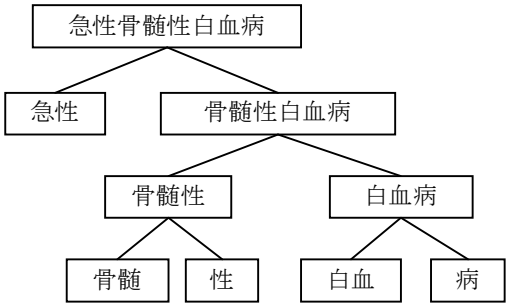

図 1 複合語の内部構造

† 東京大学医学系研究科
Graduate School of Medicine, the University of Tokyo
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

Vol.2009-NL-191 No.20
Vol.2009-SLP-76 No.20
2009/5/22

2. 専門用語

本稿では生命科学分野の専門用語，特に疾患名や解剖部位の用語の内部構造に主眼を置く．ここで内部構造とはその語を構成する構成要素の間の係り受け関係とする(図2).

専門用語が一般の複合名詞と異なる点として，以下のようなことが挙げられる．まず専門用語とは特定の領域内で使われるものであり，その分野の文化に依存して独自の内部構造を持ちうる．例えば「糖原病 I 型」は I 型の糖原病という意味であり，I 型が糖原病に係っていて係り受けの方向が通常と逆になっているし，「角結膜炎」の「角結膜」とは「角**膜**」と「結**膜**」が並列に結びついたうえで文字の縮退が起こっているものである．このような構造は，従来の日本語の形態素の係り受け構造をそのまま適用しても表現しきれない.

複合名詞の内部構造に関する研究として，漢字熟語内の品詞列・係り受けの調査[2]，医学用語を対象とした用語構造解析[5]，意味クラスの共起情報を用いた構造解析[4]，相互情報量を用いた構造解析[6]などが挙げられる．しかしいずれも従来の形態素解析・係り受け解析の域を出ておらず，本研究とは異なる．竹内らの提案する LCS は同じく複合語の解析を試みているが[7]，サ変動詞を含んだ語を扱っており，本研究とは扱う対象が異なる.

また，ある程度複雑な概念を限られた文字数で表すため，構成性が弱いと考えられる.従って内部構造を決定するのが知識を持つ人間であっても困難である場合がある．例えば「全前脳症」の「全」はどこに係るだろうか.「前脳」とはヒトの発生段階で脳の一部として存在し，複数の組織に分化する部位である.「全前脳症」とはこれが分化せずそのまま残ってしまったために奇形が生じるという疾患である．この場合，「前脳が全部そのまま残っている」ということから「全」は「前脳」に係るのが正解であろう.

3. 内部構造の表現方法

3.1 語単位の係り受けによる表現

既に述べたように複合語の内部構造とは構成要素とその間の関係から成る．分割された構成要素間の関係を以下の 4 種類に分類した.

D：前から後ろへの係り受け
　例：急性 => 肺炎（急性肺炎）
R：後ろから前への係り受け
　例：糖原病 <= I 型（糖原病 I 型）
P：並列
　例：脊髄 + 小脳（脊髄小脳変性症）

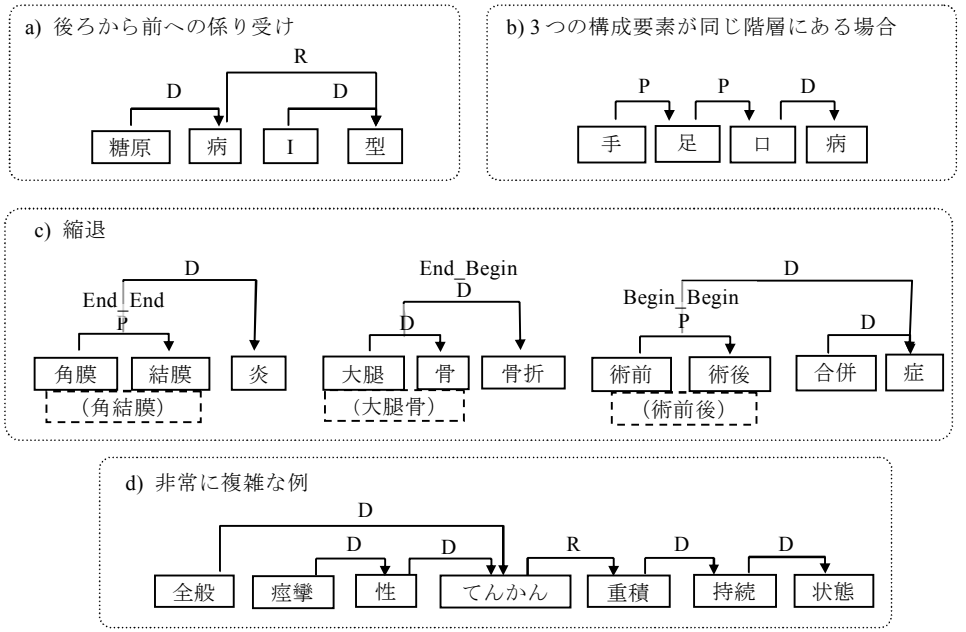

図 2　語単位のラベル付き係り受け

U：上記以外の結びつき
　例：B + 1 + 6（B16 メラノーマ細胞）

図 1 のような内部構造は，枝に上記 4 種のラベルのいずれかを付与した係り受け木で表現できる（図 2）．日本語の係り受け解析では一般に「自身よりも後の形態素に係る」「1 つの形態素が複数の係り先を持たない」「交差が起きない」の 3 つの制約を利用する．1 つ目の「自身よりも後の形態素に係る」に関して，「糖原病 I 型」は従わないことになるが，図 2a のように係り受けの方向は前から後ろに固定してラベル R を付与することで，制約を守ったまま逆向きの係り受けを表現することが出来るようになる．2 つ目の「1 つの形態素が複数の係り先を持たない」を有効にするため，1 つの構成要素に対して R で係る構成要素は 1 つであると仮定した．現在これに反する事例は観測されていない.

また，構文木は原則 2 分木であるが，ラベルが P または U である時には子ノードは 2 個以上ある場合がある．この場合は同じ係り受け関係の連続によって表現する．上

Vol.2009-NL-191 No.20
Vol.2009-SLP-76 No.20
2009/5/22

記の「B+1+6」のほか，図 2b「手足口病」の「手+足+口」のような 3 語が並列関係にある場合がこれにあたる．

上記の表現方法は，今回対象としたデータに対して十分な表現力を持っていた．ラベルを導入したことによって，内部構造解析はラベル付きの係り受け解析として捉えられる．

ここで構成要素への分割について述べる．語を分割する際には文字の縮退を考慮する必要がある．縮退とは「大腿骨折（=大腿**骨**＋**骨**折）」「角結膜（=角**膜**＋結**膜**）」のように，複数の構成要素が結合する際，オーバーラップしている部分が 1 つに纏められる現象を指す．縮退は「大腿骨折」のように構成要素（大腿**骨**，**骨**折）のつなぎ目の文字が縮退する場合以外にも，角結膜のように末尾を同じくする構成要素（角**膜**，結**膜**）の結合や，「術前後」のように先頭を同じくする構成要素（**術**前，**術**後）の結合もある．図 2 に示すとおり，本研究ではそれぞれの現象ごとにラベルを用意した．「End_End」は構成要素の最後の文字が縮退していることを意味し，「End_Begin」は構成要素の 1 番目の最後と 2 番目の最初の文字が縮退していることを意味する．また「Begin_Begin」は構成要素の最初の文字が縮退していることを意味する．縮退が起こっていない時にはこのラベルは付与されない．

3.2 文字単位の係り受けによる表現

形態素の単位が与えられていれば前節の表現で係り受け解析が可能であるが，本研究で扱う形態素は縮退などを扱うために従来の形態素の概念とは異なる．そこで文字単位でラベルを付与し，内部構造を表現することとした（図３）．これは 3.1 で述べた 4 種の係り受けの他，構成要素そのものを作る係り受けを二種類（WB，WI）追加することで可能になる．ここで WB は構成要素の先頭部分，WI はそれ以外の部分での係り受けである．これらを用いて縮退を表現する．例えば End_End 型の縮退は「角-膜」と「結-膜」が WB で結ばれて構成要素を成すと表現する．Begin_Begin 型の場合は「前-後」を P で結んだうえで「術」が WB で係るとした．「術-前」「術-後」を WB で結ぶこともできるが，「1 つの形態素が複数の係り先を持たない」という制約に違反することになるためにこの方法は避けた．End_Begin 型は図中の「大腿骨折」のように表現できる．この時，もし「大腿」が「骨折」に係るのであれば「腿」は「骨」ではなく「折」に係るということに注意して欲しい．

ここで縮退が起きているうちの 2 語について，「角結膜炎」では「角膜」と「結膜」が P で結ばれることが，「大腿骨折」では「大腿骨」が「骨折」に D で係ることが明示されていないが，縮退が起きている時には以下のようにラベルを決める．前者のように縮退文字の位置が構成要素内で同じであった場合には P，後者のように一つ目の構成要素の最後と二つ目の構成要素の最初の文字が縮退する時には D とする．これはそれぞれ「End_End」「End_Begin」に対応するものである．

また，この表現方法では 1 回の縮退で消える文字は必ず 1 文字でなければならない．

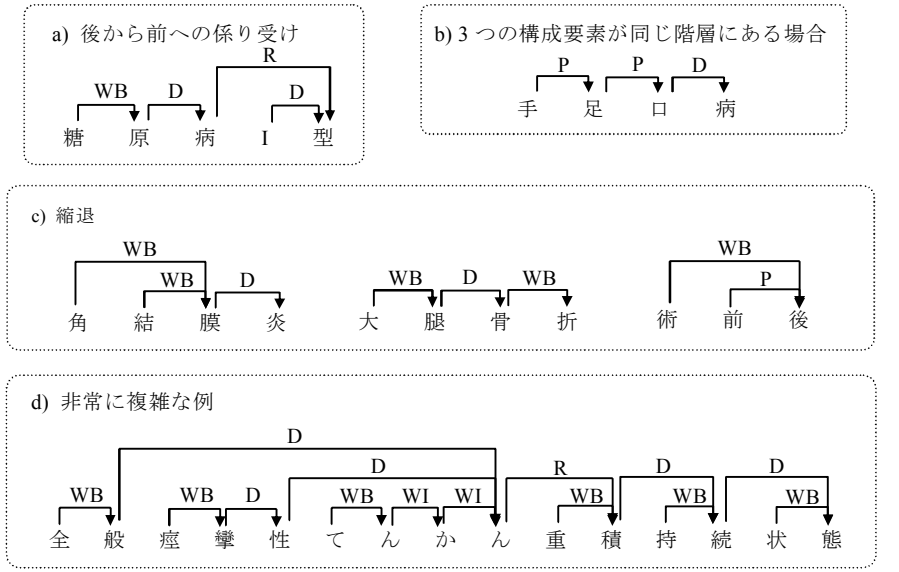

図 3　文字単位のラベル付き係り受け

仮に 2 文字が縮退すると，縮退する 2 文字の間に 2 つの関係が同時に成立し，これを表現することができない．現実には 2 文字以上の縮退は起こらないと思われるため，この制約が問題となることはない．

この表現方法を用いると文字の縮退を自然に表現することができる．また従来の形態素解析と係り受け解析に相当する処理を一つの処理に纏めることができる．

Vol.2009-NL-191 No.20
Vol.2009-SLP-76 No.20
2009/5/22

4. 内部構造解析

4.1 学習データ

ライフサイエンス辞書（以下 LSD）2008 年 4 月版を利用した．LSD は生命科学分野の専門用語辞書で，2008 年 4 月版では日本語 94,707 語，英語 83,956 語を掲載している．掲載語の一部には文献管理用に開発されたシソーラスである MeSH（Medical Subject Headings）のコードが付与されている．このうち疾患を表すコードの付与された八文字以上の日本語から 273 語に対してタグ付けを行った．この時，タグ付けした語の構成要素の内部構造も同時に保存される．例えば「急性骨髄性白血病」のタグ付け作業は図 1 の一番上から順に行われるが，この時 LSD に掲載されている「骨髄性白血病」「骨髄性」「白血病」も同時に内部構造が定義される．従って，タグ付けされた語は疾患名の他に解剖部位名や物質名を含む．これらを合わせるとタグ付けされた語は合計 804 語であり，このうち係り受け関係ラベルとして P（並列），R（逆向き），U（方向無し）を含むのはそれぞれ 21 語，8 語，44 語であった．

4.2 アルゴリズム

[Nivre 2004]の Shift-Reduce モデルによる係り受け解析アルゴリズムを使った．パーサの状態は{S,I,A}の三つ組で表される．解析中のノードを Stack S に，未解析ノードを Input I に，成立した係り受け関係をリスト A に格納し，初期状態{nil, W, nil}（W は入力列）が{S, nil, A}となったら解析終了となる．詳細は[Nivre 2004]を参照されたい．今回は実装として http://maltparser.org/に公開されている MaltParser を用いた．

4.3 素性

使用した素性を表 1 に示す．大きく分けて，文字の持つ内部素性，辞書による外部素性，解析済部分の係り受け情報の三種類を使った．

4.4 結果

上記の素性全てを用いて，5 分割交差検定を行った結果，文字対での精度は 96.9%（ラベル無しでは 97.8%），語での精度は 87.6%（ラベル無しでは 89.9%）であった．ラベルごとの誤り率を表 2 に示す．

表 1. 使用した素性

文字素性	1	文字	STRING
	2	文字種	STRING
	3	文字の位置	STRING
辞書素性	4	この語の部分文字列でこの文字を最後とする辞書掲載語が存在する	BOOLEAN
	5	4の辞書掲載語の長さ	INTEGER
	6	4の辞書掲載語のMeSHカテゴリ（1桁+3桁+全て）	STRING
	7	この語の部分文字列でこの文字を最後とする接尾辞が存在する	BOOLEAN
	8	この語の部分文字列でこの文字を含む辞書掲載語が存在する	BOOLEAN
	9	この文字とn文字先の文字を繋げた二文字語が辞書に存在する	BOOLEAN
	10	9でのn	INTEGER
係り受けラベル	11	Stack[0]，Stack[1]，Input[0]，Input[1]の係り受けラベル	
	12	Stack[0]の最左/右dependentの係り受けラベル	
	13	Input[0]の最左dependentの係り受けラベル	
	14	Stack[0]の左隣の文字の係り受けラベル	

4～6 は辞書掲載語が複数あった場合，短い方から 2 語を使った．9 と 10 は 10 文字先まで見ている．

素性の例：「急性骨髄性白血病」
「骨」の文字・辞書素性（‘骨’，漢字, 3, True, 1, ’A10.165.265’, ’A’, False, nil, nil, nil, True, True, True, 1, True, 2, nil, …）
「病」の文字・辞書素性（‘病’，漢字，7，True，1，’C23.550.288’，’C’，True，3，’C04.557.337’，’C’，True，True，null，null，…）

Vol.2009-NL-191 No.20
Vol.2009-SLP-76 No.20
2009/5/22

出現回数が多く比較的簡単だと思われる WB，WI，D は間違いが少ない．また，一度 WB を間違えるとその後の WI や D に間違いが伝播する場合がある．

表 3 に D を含む語数を示す．誤答のうち係り先を誤っているものがほとんどであることが分かる．エラーのパターンとしては係り先を隣の語にする，または最後の語にするというものが多い．前者については決定的な手法を使っているのが影響している可能性がある．後者については素性や学習データ量の不足によるものであると考えられる．

R は出現回数が少ないためにこの結果だけで精度の評価をするのは妥当ではない．使用したデータ中で R が入っている語は，「<病名> / <アルファベット・数字>型」，「オリーブ橋 / 小脳」「てんかん / 重積」の三種である．エラーが起きたのは 1 つ目と 3 つ目であった．前者について，学習データには「<アルファベット>型」が入っていたのに対し，テストデータでは「<数字>型」となっていた．これは文字種の素性を工夫することで解決可能と考えられる．一方，後者の「てんかん重積」は「一度のてんかん発作が治まらないうちに次の発作が繰り返し起こる」ことを意味しており，R を付与するためには，てんかん重積がてんかんの下位語である，というようなカテゴリ素性が必要であろう．

P は上記の例と同様に難しいタスクである．例えば「神経 / 筋」が並列であるということは知識が無ければ答えられない．一方で U は片仮名語やアルファベットの連続の場合に付与されるラベルであるので，精度が良いと思われる．例えば「Dandy / - / Walker」「I / I / I (/ 型)」のような場合に使われるために間違いが起こりにくい．

前述したとおり，LSD 掲載語には MeSH のカテゴリが付与されているものが含まれる．カテゴリは階層になっており，トップレベル（コード 1 桁）では解剖部位，組織，疾患名，物質・薬品名，…のように分類されている．ここに数字を加えていくことで階層が深くなっていく．今回，素性としてトップレベル（1 桁），次の階層（3 桁），最深階層（全て）の 3 つを利用したが（表 1 の 6），これらがどの程度の寄与があるのかを調べた（表 4）．結果，3 桁と全ての 2 つを使った場合が 88.2%と最も精度が高かった．

5. 考察

提案手法のコーパスは小規模なものであるが，専門用語の内部構造解析精度として 88.2%という数値を得ることが出来た．この精度が他の言語処理アプリケーションにどの程度貢献するのかは今後の課題である．また次のような課題も残されている．

専門用語は必ずしも体系的に命名されているわけではない[3]．構成要素間の係り受けが複数考えられることもある．専門家であっても内部構造の定義は簡単ではない．「慢性肉芽腫性疾患」で「慢性」の係り先は「肉芽腫」「疾患」のどちらも正解と受け

表 2 ラベルごとの誤り率

ラベル	誤答	総数	誤答率
WB	35	1496	2.30%
WI	14	1114	1.20%
D	41	1598	2.00%
R	2	8	25%
P	11	22	50%
U	4	85	4.70%

表 3 ラベル D を含む語数

頻度	総数	誤答	誤ラベル	誤係り先
0	23	0	0	0
1	372	10	8	7
2	195	20	7	17
3	99	13	6	10
4	65	19	3	18
5	31	8	0	8
6	12	4	1	4
7	5	5	2	4
8	1	1	0	1
9	1	1	0	1

表 4　カテゴリ素性の精度への影響

1桁	3桁	全て	文字対	語
			96.9% (98.1%)	86.8% (90.3%)
✓			97.0% (98.1%)	87.6% (90.1%)
	✓		97.1% (98.1%)	87.2% (90.2%)
		✓	97.0% (98.0%)	87.3% (89.2%)
✓	✓		97.1% (98.1%)	87.8% (90.2%)
	✓	✓	**97.2% (98.2%)**	**88.2% (90.3%)**
✓		✓	97.1% (**98.2%**)	87.4% (90.1%)
✓	✓	✓	96.9% (97.8%)	87.6% (89.9%)

※括弧内: ラベル無での精度

入れられるものであろう．同様に「静脈洞血栓症」は「静脈洞における血栓症（静脈洞が症に係る)」とも「静脈洞に血栓が出来る症（静脈洞が血栓に係る)」とも捉えられる．今回の実験結果で誤りとされた例の中にもこれに該当するものがあった．こういった難しい事例をどう捉えるべきかという議論が今後必要であると考える．

タグ付け作業は当該分野の専門知識をある程度持つ人間が行う必要があるが，作業者は構文木等の言語学的な考え方を知っていなければならない．作業者が考えるその言葉の意味を係り受け関係へ変換する際には構文木や係り受け関係の考え方を理解していることが必要であるからである．今回トップダウンに語を分割する手続きをとったこともあり，言語学に馴染みの無い人には直観的でない場合がしばしば見られた．例えば「甲状腺機能亢進」を分割する場合，「甲状腺」が「機能」に，「機能」が「亢進」に係るので，最初に「甲状腺機能+亢進」と分割すべきだが，自然言語では「甲状腺の機能亢進」としても妥当であるために，このように分割してしまうことがあった．トップダウンな方法の利点は，構成要素として出現した語が既に内部構造を定義されていた場合に同じ定義を繰り返さなくて良いという点である．この利点を生かしたままボトムアップ的に作業できる環境が望ましいと言えよう．

今回使用した単語素性は MaltParser の実装に合わせて選んだために，非常に粗いものとなっている．具体的には，入力文字列と辞書のみから文字に対して単語素性を入れている．しかしこの「単語」は必ずしも解析対象としている複合語の構成要素ではない（例：角膜の角は動物の角ではないが LSD には動物の角として「角」が掲載されている)．使用した手法は決定的な手法であるので，解析済み部分の係り受け情報を利用すればよりきめ細やかな素性が使えるはずである．例えばその文字に WI で係っている文字があったら WB で係っているところまで辿っていくことで確実な単語素性が使える．

6. おわりに

専門用語の内部構造の表現方法としてラベル付きの構文木，文字単位の係り受けを提案した．また MaltParser を用いて文字単位の係り受け解析を試みた．対象としたのは疾患名を始めとした生命科学分野の語である．文字対に対する係り受け関係の判定精度は 97.2%，内部構造解析の精度は 88.2%であった．今後は非決定的な手法の採用や，よりきめ細やかな素性を入れることによって精度の向上を図りたい．

謝辞　本研究の一部は，文科省統合データベースプロジェクト「ライフサイエンス分野の統合データベース整備事業」の支援を得て行われました．また，ライフサイエンス辞書を提供していただいた京都大学金子周司教授に感謝します．また多大なご助言を頂いた東京大学知の構造化センター荒牧英治氏に心より感謝いたします．

参考文献

1)　Nivre, J., Hall, J. and Nilsson, J. (2004) Memory-Based Dependency Parsing. In Ng, H. T. and Riloff, E. (eds.) Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL), pp. 49-56.
2)　梅木定博，後藤智範．辞書見出し語の 7 文字漢字熟語を対象とした語基構成の解析．情報処理学会研究報告 自然言語処理 研究報告 No.184 pp.113-118 (2008).
3)　大島智夫．日本の医学用語についてのおぼえ書き．専門用語研究. No.1, pp.18-21 (1990).
4)　小林義行，徳永健伸，田中穂積．名詞間の意味的共起情報を用いた複合名詞の解析．自然言語処理. Vol.3. No.1. pp.29-43 (1996).
5)　小山照夫，大江和彦．医学専門用語の構造解析．学術情報センター紀要. No.6, pp.115-124 (1994).
6)　韓東力，伊藤毅志，古郡廷治．要素間の依存関係に基づく複合語の構造分析．電子情報通信学会論文誌 D Vol.J86-D2 No.5 pp.706-714 (2003).
7)　竹内孔一，乾健太郎，藤田篤，竹内奈央，阿部修也:分類の根拠を明示した 動詞語彙概念構造の構築，自然言語処理研究会 2005-NL-169 (2005).