大規模階層辞書を用いた日本語機能表現の代表・派生関係の分析

宇津呂武仁 (日本語機能表現班班長：筑波大学大学院システム情報工学研究科)*
長坂泰治 (日本語機能表現班協力者：筑波大学大学院システム情報工学研究科)
坂本明子 (日本語機能表現班協力者：筑波大学大学院システム情報工学研究科)
島内蘭 (日本語機能表現班協力者：筑波大学大学院システム情報工学研究科)
劉 颯 (日本語機能表現班協力者：筑波大学第三学群工学システム学類)
松吉俊 (日本語機能表現班協力者：奈良先端科学技術大学院大学情報科学研究科)
土屋雅稔 (日本語機能表現班協力者：豊橋技術科学大学情報メディア基盤センター)

Analyzing Canonical/Derivational Relation of Japanese Functional Expressions Based on a Large Scale Hierarchical Lexicon

Takehito Utsuro (University of Tsukuba)
Taiji Nagasaka (University of Tsukuba)
Akiko Sakamoto (University of Tsukuba)
Ran Shimanouchi (University of Tsukuba)
Sa Liu (University of Tsukuba)
Suguru Matsuyoshi (Nara Institute of Science and Technology)
Masatoshi Tsuchiya (Toyohashi University of Technology)

1. はじめに

機能表現[1]とは,「にあたって」や「をめぐって」のように，2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現である．一方，この機能表現に対して，それと同一表記をとり，内容的な意味をもつ表現が存在することがある．例えば，文 (1) と文 (2) には「にあたって」という表記の表現が共通して現れている．

(1) 出発する にあたって，荷物をチェックした．

(2) ボールは壁 にあたって，跳ね返った．

文 (1) では，下線部はひとかたまりとなって,「機会が来たのに当面して」という機能的な意味で用いられている．それに対して，文 (2) では，下線部に含まれている動詞「あたる」は，動詞「あたる」本来の内容的な意味で用いられている．このような表現においては，機能的な意味で用いられている場合と，内容的な意味で用いられている場合とを識別する必要がある．

我々はこれまでに，現代語複合辞用例集 [国研 01](以下，用例集) 中の代表的複合辞一覧に基づいて，それらの派生形である 337 種類の機能表現を規定し，その用例データベース (日本語複合辞用例データベース [土屋 06]，以下，用例データベース) を作成した．また，それらの用例データベースを訓練事例として，機械学習により機能表現の検出・係り受け解析を行う方式を提案した [土屋 07, 注連 07]．また，機能表現の異形の語構成パターンを網羅することにより，日本語機能表現一覧 [松吉 07](以下，機能表現一覧[2]) を作成した．

*utsuro @ iit.tsukuba.ac.jp

[1]機能表現は，複数形態素からなる複合辞と一つの形態素からなる機能語から構成されるが，本稿では，複合辞と同等の意味で機能表現という用語を用いる．

[2]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 機能表現辞書の 9 つの階層

階層		分類数	表現数		
			合計 (L^9 表現数)	助動詞型以外	助動詞型
L^1	見出し語	—	341 (488)	281	207
L^2	意味	45/128/199	435 (488)	281	207
L^3	派生 (格助詞型，接続助詞型，連体助詞型，接続詞型，助動詞型，形式名詞型，とりたて詞型，提題助詞型)	8	555	348	207
L^4	機能語の交替	—	774	492	282
L^5	音韻的変化	38	1,187	633	554
L^6	とりたて詞の挿入	18	1,810	659	1151
L^7	活用	—	6,870	659	6211
L^8	「です/ます」の有無	2	9,722	895	8827
L^9	表記のゆれ	—	16,801	1360	15411

ここで，[土屋 07,注連 07] の機械学習による機能表現検出においては，一つの表現あたり 50 例程度の訓練用例に対して，人手で機能的・内容的等の用法判定を行う必要がある．しかし，機能表現一覧の全機能表現 16,801 種類に対して，それだけの規模の作業を行うことは容易ではない．そこで，[長坂 08] では，機能表現一覧の階層性を利用し，階層において下位に位置する機能表現 (以下，派生的表現) について，用法が類似するより上位の表現 (以下，代表的表現) に言い換えた後，用法判定を行う方式を提案した．

一方，[長坂 09b] では，[長坂 08] の提案をふまえて，機能表現一覧中の情報のうち，特に文体の情報に注目し，代表的表現および派生的表現の区別を整理した．さらに，毎日新聞 1995 年分のテキストデータ中において，機能表現一覧の機能表現の出現頻度調査を行い，[土屋 07,注連 07] の機械学習による機能表現検出において必要となる訓練事例 (出現頻度 50 以上) が存在する機能表現の規模を推定した．また，[長坂 09a] では，毎日新聞 1995 年分のテキストデータ中に 50 回以上出現する代表的表現の表記を対象として，人手で機能的用法・内容的用法の判定作業を行った．さらに，各機能表現表記に対して，機能的用法・内容的用法の両方が適度な割合で混合して出現し，機械学習によって機能表現検出を行う必要のある機能表現表記の割合を求めた結果について報告した．

本稿では，[長坂 09a] における人手用法判定作業の範囲を拡大し，毎日新聞 1995 年分のテキストデータ中に 50 回以上出現する代表的表現の表記のうち，[長坂 09a] では未判定であった表記を対象として，新たに人手で機能的用法・内容的用法の判定作業を行った．さらに，それらの派生的表現の各出現箇所の前後の形態素の品詞の組み合わせ，および，代表的表現の出現箇所の前後の形態素の品詞の組み合わせについて，代表・派生間の傾向の差異を分析する．分析の結果より，前後の形態素の品詞が代表・派生間において不変の場合には，派生的表現が機能的用法である割合が高く保たれることを示す．

2. 階層的機能表現辞書

機能表現一覧 [松吉 07] は，9 つの階層構造をなしており，各階層は，表 1 に示されるような観点によって分類されている．同表に，各階層における機能表現数が示されており，図 1 に階層構造の一部をそれぞれ示す．

また，機能表現の文体に着目し，文体ごとの機能表現の振る舞いについて述べる．文体とは，機能表現一覧中の表現に付与されている情報であり，常体，堅い文体，口語体，敬体の 4 種類がある．表 2 にそれぞれの文体における表現例を示す．

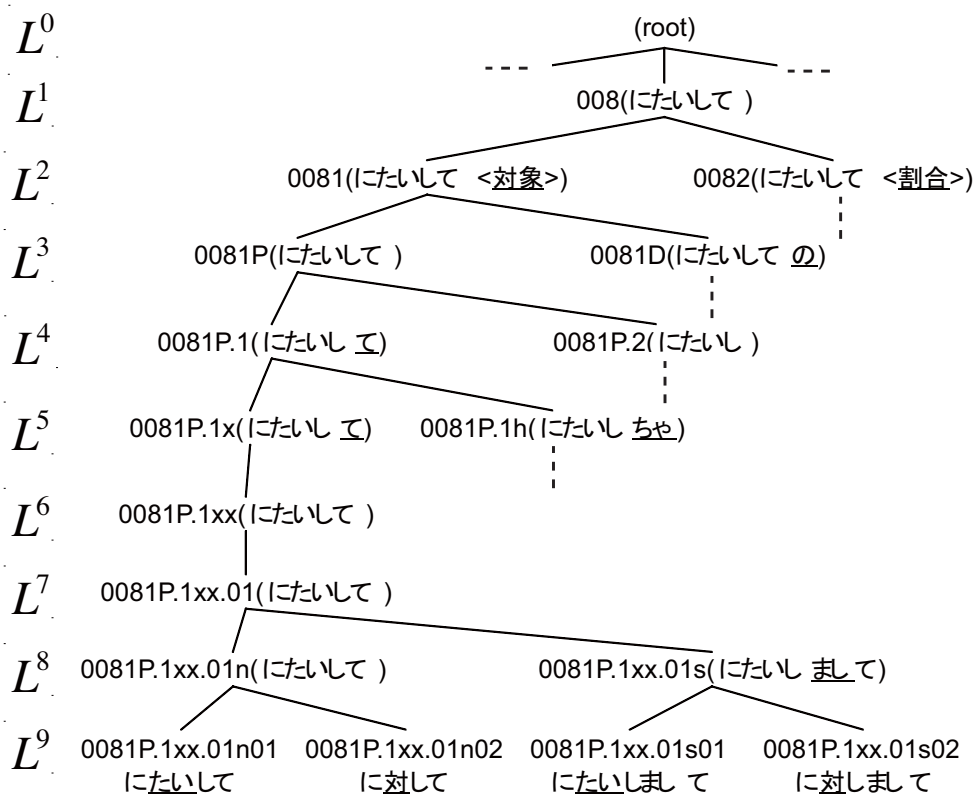

図 1: 機能表現辞書階層構造の一部

表 2: 文体の種類

文体	表現例
常体	について
堅い文体	につき
口語体	についちゃ
敬語体	につきまして

3. 代表的表現への集約

[長坂 08] で提案した代表的表現への集約方式においては，階層の上位に位置する代表的表現は，L^4 階層相当の 1,000 表現程度の規模とする．そして，機能表現一覧において，代表的表現を除く表現はすべて，言い換えの対象の表現となる．本研究では，これらの表現を派生的表現と定義する．派生的表現を代表的表現に言い換える際には，以下の制約を課す．

- 機能表現の語頭の無声・有声の制約により前接する活用語の活用型が制限される場合は，この制限を保持する．
- 機能表現の仮名表記・漢字表記の違いを保持する．
- 助動詞型の機能表現の場合には，言い換え前後で活用形を保持する．

表 3: 毎日新聞 1995 年において，50 回以上出現する機能表現数の分布

	助動詞型 (基本形)		助動詞型以外			
	代表的表現	派生的表現	代表的表現	派生的表現	その他	合計
常体	164	38	87	0	178	467
堅い文体	8	3	0	0	9	20
敬体	14	42	0	1	0	57
口語体	7	37	1	13	0	58
合計	193	120	88	14	187	602

表 4: 機能的用法・内容的用法の分布

	助動詞型 (基本形)		助動詞型以外			
	代表的表現	派生的表現	代表的表現	派生的表現	その他	合計
常体	63.9/29.9/6.1	89.7/3.4/6.9	53.5/34.9/11.6	—	46.0/44.8/9.2	56.7/34.8/8.5
堅い文体	75.0/12.5/12.5	100/0/0	—	—	11.1/55.6/33.3	44.4/33.3/22.2
敬体	92.9/7.1/0	88.9/8.3/2.8	—	—	—	90.2/7.8/2.0
口語体	16.7/66.7/16.7	42.1/36.8/21.1	0/0/100	50.0/50.0/0	—	35.7/42.9/21.4
合計	65.1/28.6/6.3	78.8/12.9/8.2	52.9/34.5/12.6	50.0/50.0/0	44.2/45.3/10.5	58.4/32.6/9.0

x：機能的用法の割合
$90\% \leq x \leq 100\%$ となる表現の割合 (%) / $10\% < x < 90\%$ となる表現の割合 (%)
/ $0\% \leq x \leq 10\%$ となる表現の割合 (%)

4. 新聞記事における機能表現の分布

4.1 機能表現表記数の分布

毎日新聞 1995 年の一年分の中で，50 回以上出現する機能表現表記数の調査を行った結果を表 3 に示す．機能表現一覧の中には，同一の表記に対して複数の ID が存在する表現があることをふまえて，表 3 には，機能表現表記単位で集計した結果を示す．表 3 より，全 602 表現のうち，常体の機能表現が 467 表現と大きい割合を占めていることが分かる．

4.2 機能的用法・内容的用法の分布

これまでに，[土屋 06, 土屋 07, 注連 07] で述べたように，全機能表現表記のうち，特に，機能的用法・内容的用法の両方が適度な割合で混合して出現する機能表現表記に対してのみ，機械学習によって機能表現検出を行う必要がある．また，[土屋 06] で報告したように，[土屋 06] の用例データベースの範囲においては，毎日新聞 1995 年の一年分の中で，50 回以上出現する機能表現表記 187 表記のうち，機能的用法・内容的用法の両方が適度な割合で混合して出現する機能表現表記の割合は約 3 分の 1 程度であった．

一方，本節では，表 3 中の代表的表現，派生的表現，および,「その他」の表現 (助動詞型以外の場合のみ) を対象として，機能的用法・内容的用法の用法判定作業を行った結果を表 4 に示す (判定箇所数 50,064 箇所)．この結果から，機能的用法・内容的用法の両方が適度な割合で混合して出現する (機能的用法の割合 x が，$10\% < x < 90\%$ となる) 機能表現表記の割合は，代表的表現，派生的表現および「その他」の全体では約 3 分の 1 程度であることがわかる．しかし,「その他」のみでは，約 45%と多くなっている．

表 5: 毎日新聞 1995 年に 50 回以上出現する代表的表現に対する派生的表現のうち，毎日新聞 1995 年に 50 回以上出現するものを除いた表現の数

文体	助動詞型 (基本形)	助動詞型以外
常体	347	8
堅い文体	52	2
敬体	537	78
口語体	164	25
合計	1100	113

表 6: 代表的表現・派生的表現の前後の形態素の品詞を用いた機能的用法の判定: 評価結果

前後の形態素の 品詞に対する条件	箇所数 (割合 (%))	表現数	判定精度 (%)
前後とも代表・派生間で不変	759 (80.3)	132	**97.2**
前のみ代表・派生間で不変	71 (7.5)	27	63.4
後のみ代表・派生間で不変	97 (10.3)	25	39.2
前のみ不変の代表出現箇所・後のみ 不変の代表出現箇所の両方が存在	5 (0.5)	5	100.0
その他	13 (1.4)	6	23.1
合計	945 (100)	158	87.7

5. 代表的表現・派生的表現の前後の形態素の品詞を用いた機能的用法の判定

機能的用法として偏って出現する代表的表現に対して，毎日新聞 1995 年分のテキストデータ中に 50 回以上出現しない派生的表現 (表 5 の機能表現の一部) の出現箇所 945 箇所 (機能表現数 158) を対象として，機能的用法・内容的用法の判定作業を行った．その結果，派生的表現が機能的用法である割合は約 88%であり，代表的表現における出現割合よりも高い割合で内容的用法が出現することが分かった．さらに，それらの派生的表現の各出現箇所の前後の形態素の品詞の組み合わせ，および，代表的表現の出現箇所の前後の形態素の品詞の組み合わせについて，代表・派生間の傾向の差異を分析した[3]．代表・派生間で前後の形態素の品詞が不変か否かに応じて，派生的表現が機能的用法である割合が変動する様子を表 6 に示す．この結果から，前後の形態素の品詞が代表・派生間において不変の場合には，派生的表現が機能的用法である割合は 97%と非常に高いことが分かった．逆に，前後の形態素の品詞のいずれか，もしくは，両方の文脈が代表・派生間で保存されない場合には，内容的用法の割合が高くなった．特に，前接する形態素の品詞が保存されない場合には，内容的用法の割合が極端に高くなった．

表 7 には，代表・派生間で前後の形態素の品詞が不変か否かの各条件のもとで，派生的表現が機能的用法となる事例，および内容的用法となり誤りとなる事例をそれぞれ示す．これらの事例から，実際に，前後の形態素の品詞のいずれか，もしくは，両方が代表・派生間で保存されない場合に，助詞の挿入の有無等が異なる，等，代表・派生間で特性が大きく異なることが分かる．以下，各事例について説明する．

[3]ただし，助動詞型機能表現の場合には，代表的表現を活用させて，派生的表現と同じ活用形を代表的表現とした．また，本節における分析対象箇所については，その大半が助動詞型機能表現であった．

表 7: 代表的表現・派生的表現の前後の形態素の品詞を用いた機能的用法の判定: 正解例・誤り例

前後の形態素の品詞に対する条件	前形態素の品詞 (表記)-代表的表現表記-後形態素の品詞 (表記)	前形態素の品詞 (表記)-派生的表現表記-後形態素の品詞 (表記)	正解・誤り	派生的表現表記の用法
(1) 前後とも代表・派生間で不変	動詞 (示す)-ことができ-助動詞 (ない)	動詞 (歩く)-ことさえでき-助動詞 (なく)	正解	機能的用法
	動詞 (控え)-てほしい-記号 (」)	動詞 ((上昇) し)-てもほしい-記号 (」)	誤り	内容的用法
(2) 前のみ代表・派生間で不変	動詞 ((するなど) し)-ておれ-助詞 (ば)	動詞 (立っ)-てもおれ-助動詞 (ぬ)	正解	機能的用法
	動詞 (張っ)-てあり-記号 (,)	動詞 (探し)-てもあり-助動詞 (ませ (ん))	誤り	内容的用法
(3) 後のみ代表・派生間で不変	動詞 ((気が) し)-てならない-記号 (.)	助動詞 ((申し訳) なく)-てなりません-記号 (.)	正解	機能的用法
	動詞 (望ん)-でいる-名詞 (人)	名詞 (今)-でもいる-名詞 (そう)	誤り	内容的用法

(1) 前後とも代表・派生間で不変

- 代表的表現の表記，派生的表現の表記とも機能的用法となる．代表的表現の表記「ことができ」は「示すことができない」という文脈で出現し，派生的表現の表記「ことさえでき」は「歩くことさえできなく」という文脈出現しており，前後の形態素の品詞はともに「動詞」と「助動詞」で一致している．
- 代表的表現の表記は機能的用法となるが，派生的表現の表記は内容的用法となる．代表的表現の表記「てほしい」は『控えてほしい」』という文脈で出現し，派生的表現の表記「てもほしい」は『上昇してもほしい」』という文脈で出現しており，前後の形態素の品詞はともに「動詞」と「記号」で一致している．

(2) 前のみ代表・派生間で不変

- 代表的表現の表記，派生的表現の表記とも機能的用法となる．代表的表現の表記「ておれ」は「するなどしておれば」という文脈で出現し，派生的表現の表記「てもおれ」は「立ってもおれぬ」という文脈で出現しており，前接の形態素の品詞は「動詞」で一致している．後接の形態素の品詞では，代表的表現では「助詞」，派生的表現では「助動詞」で一致していない．
- 代表的表現の表記は機能的用法となるが，派生的表現の表記は内容的用法となる．代表的表現の表記「てあり」は「張ってあり,」という文脈で出現し，派生的表現の表記「てもあり」は『探してもありません』という文脈で出現しており，前接の形態素の品詞は「動詞」で一致している．後接の形態素の品詞では，代表的表現では「記号」，派生的表現では「助動詞」で一致していない．

(3) 後のみ代表・派生間で不変

- 代表的表現の表記，派生的表現の表記とも機能的用法となる．代表的表現の表記「てならない」は「気がしてならない.」という文脈で出現し，派生的表現の表記「てなりません.」

は「申し訳なくてなりません.」という文脈で出現しており，後接の形態素の品詞は「記号」で一致している．前接の形態素の品詞では，代表的表現では「動詞」，派生的表現では「助動詞」で一致していない．

- 代表的表現の表記は機能的用法となるが，派生的表現の表記は内容的用法となる．代表的表現の表記「でいる」は「望んでいる人」という文脈で出現し，派生的表現の表記「でもいる」は『今でもいるそう』という文脈で出現しており，後接の形態素の品詞は「名詞」で一致している．前接の形態素の品詞では，代表的表現では「動詞」，派生的表現では「名詞」で一致していない．

6. 関連研究

[松吉 08] においては，機能表現一覧 [松吉 07] 中の機能表現を対象として，意味を保存する言い換えが可能な機能表現の分類を規定している．一方，本論文では，機能表現の用法判定の性能を保ったまま，代表的表現への言い換えを行うという，より緩い制約のもとでの機能表現の言い換えが目的である．また，代表的表現への言い換えを介した機械翻訳の研究としては，内容語と口語的な機能表現を扱った [山本 01]，機能表現一覧 [松吉 07] の機能表現を対象とした [坂本 09, 島内 10, Nagasaka10, 劉 10] が，機能表現の検出・係り受け解析等の解析を対象とした研究としては，[土屋 07, 注連 07, 小早川 09] がある．

7. まとめ

本稿では，機能的用法として偏って出現する代表的表現に対して，毎日新聞 1995 年分のテキストデータ中に 50 回以上出現しない派生的表現の出現箇所を対象として，機能的用法・内容的用法の判定作業を行った．さらに，それらの派生的表現の各出現箇所の前後の形態素の品詞の組み合わせ，および，代表的表現の出現箇所の前後の形態素の品詞の組み合わせについて，代表・派生間の傾向の差異を分析した．分析の結果，前後の形態素の品詞が代表・派生間において不変の場合には，派生的表現が機能的用法である割合が高く保たれることが分かった．

一方，これまで，[長坂 08] においては，機能的用法・内容的用法の両方が適度な割合で混合して出現し，機械学習によって機能表現検出を行う必要のある代表的表現に対しては，その派生的表現についても，代表的表現に言い換えた後，代表的表現を対象として訓練した機能表現検出器を適用する方式を提案している．しかし，本稿の結果より，機能的用法・内容的用法の両方が適度な割合で混合して出現する代表的表現の場合においても，代表的表現を対象として訓練した機能表現検出器を安定して適用できる場合と，前後の形態素の品詞が代表・派生間において不変であるか否かとの間の相関を分析する必要があると考えられる．

参考文献

[小早川 09] 小早川健, 関場治朗, 木下明徳, 熊野正, 加藤直人, 田中英輝：単語格子とマルコフモデルによる日本語機能表現の解析 — 日本語機能表現辞書「つつじ」を用いて —, 電子情報通信学会技術研究報告, NLC2009-1, pp. 15–20 (2009).

[国研 01] 国立国語研究所: 現代語複合辞用例集 (2001).

[劉 10] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊：意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析, 言語処理学会第 16 回年次大会論文集 (2010).

[松吉 07] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[長坂 08] 長坂泰治, 宇津呂武仁, 土屋雅稔：大規模日本語機能表現辞書の階層性を利用した機能表現検出, 言語処理学会第 14 回年次大会論文集, pp. 837–840 (2008).

[長坂 09a] 長坂泰治, 坂本明子, 宇津呂武仁, 森下洋平, 松吉俊, 土屋雅稔：階層的機能表現辞書に基づく新聞記事中の機能表現の調査・分析, NLP 若手の会 第 4 回シンポジウム (2009).

[長坂 09b] 長坂泰治, 宇津呂武仁, 松吉俊, 土屋雅稔：大規模階層辞書を利用した日本語機能表現の集約と解析, 言語処理学会第 15 回年次大会論文集, pp. 328–331 (2009).

[Nagasaka10] Nagasaka, T., Shimanouchi, R., Sakamoto, A., Suzuki, T., Morishita, Y., Utsuro, T. and Matsuyoshi, S.: Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Translation Rule Acquisition from Parallel Patent Sentences, *Proc. 7th LREC* (2010).

[坂本 09] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集, pp. 654–657 (2009).

[島内 10] 島内蘭, 長坂泰治, 坂本明子, 宇津呂武仁, 松吉俊：日英特許翻訳における日本語機能表現の集約的英訳可能性の調査, 言語処理学会第 16 回年次大会論文集 (2010).

[注連 07] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史：日本語機能表現の自動検出と統計的係り受け解析への応用, 自然言語処理, Vol. 14, No. 5, pp. 167–197 (2007).

[土屋 06] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一：日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).

[土屋 07] 土屋雅稔, 注連隆夫, 高木俊宏, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一：機械学習を用いた日本語機能表現のチャンキング, 自然言語処理, Vol. 14, No. 1, pp. 111–138 (2007).

[山本 01] 山本和英, 白井諭, 坂本仁, 張玉潔：SANDGLASS: 両言語換言機構を基軸とする音声翻訳, 言語処理学会第 7 回年次大会発表論文集, pp. 221–224 (2001).