平成 22 年度研究進捗状況報告：日本語機能表現班
大規模階層辞書を用いた日本語機能表現解析体系の研究

宇津呂武仁 (班長：筑波大学大学院システム情報工学研究科)*
鈴木敬文 (協力者：筑波大学大学院システム情報工学研究科)
島内蘭 (協力者：筑波大学大学院システム情報工学研究科)
松吉俊 (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 (協力者：豊橋技術科学大学情報メディア基盤センター)

Progress Report of the Year 2010: 'Japanese Functional Expressions' Group

Takehito Utsuro (University of Tsukuba)
Takafumi Suzuki (University of Tsukuba)
Ran Shimanouchi (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．

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

*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

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．

本研究では，[長坂 08] の提案の実現に向けて，代表的表現・派生的表現の間の用法の派生関係について分析を進めてきた [長坂 09b, 長坂 09a, 長坂 10b]．

[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．

[長坂 10b] 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．

以上の分析結果のうち，本稿では，特に，[長坂 10b] において明らかになった特性として，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，代表的表現と派生的表現の間で用法の傾向に相関がある，という点に注目する．そして，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する代表的表現，およびその派生的表現を対象として，この特徴をより詳細に分析する．

具体的には，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する代表的表現に対する派生的表現の表記の出現箇所を対象として，まず，機能的用法・内容的用法の判定を人手で行う．次に，

> 派生的表現の表記の各出現箇所と同様の用法となる用例が，対応する代表的表現の表記の用例中に存在するか否か，あるいは，そのような用例が作例可能か否か，

を人手で調査する．実際に，用法判定済みの用例数が 10 例以上となる派生的表現 56 表現，2,195 用例を対象とした調査を行ったところ，代表的表現の表記の用法，および，その前後の形態素の品詞が，派生的表現の表記の前後の形態素の品詞と同一となる用例が，すでに用法判定済みの 50,000 用

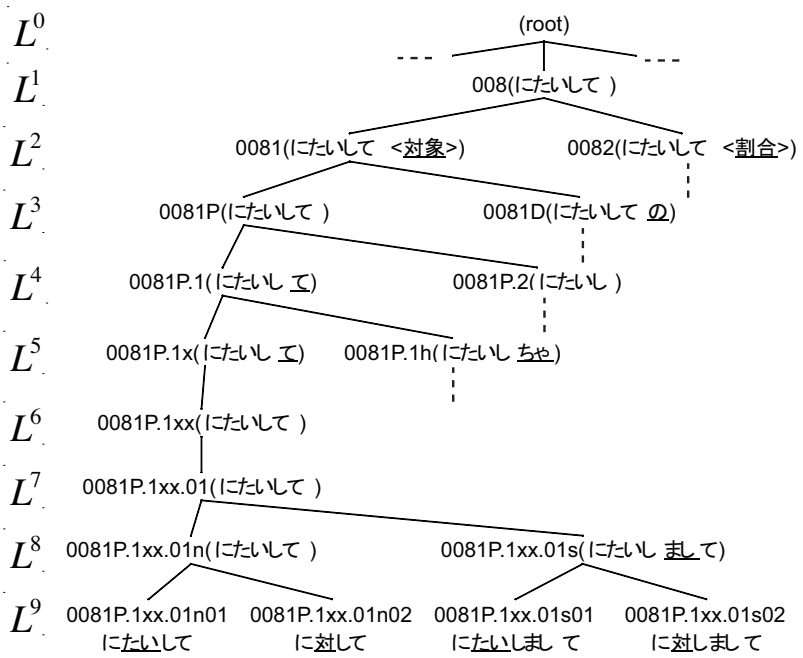

図 1: 機能表現辞書階層構造の一部

例中に存在する場合が，約 66% となった．また，同様に，代表的表現の表記の用法，および，その前後の形態素の品詞が，派生的表現の表記の前後の形態素の品詞と同一となる用例が容易に作例可能である場合が，約 18%となった．このように，調査対象となった派生的表現の表記の用例のうちの約 84%については，[長坂 10b] で示された特性を一般化した,「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，代表的表現と派生的表現の間で用法の傾向に相関がある」という仮説に従うことが分かった．一方，残りの 16%のうちの 11%については，

> 代表的表現から派生的表現へと表記上の派生が生じた結果，偶然，別の機能表現や内容的表現，あるいは，それらが複合した表現 (の一部) と同一の表記となってしまった．　(1)
> 一方，代表的表現の表記においては，派生的表現の表記と類似の現象は起こらない．

という用例であった．

以上の調査結果から，派生的表現 (の表記) を代表的表現 (の表記) に言い換えた後，用法判定を行う方式 [長坂 08] の実現に向けて，以下の重要な指針が得られた．

(p) まず，準備として，機能表現表記に対して用法判定のされていない大規模なコーパスを対象として,「機能表現一覧」 [松吉 07] 中の全ての代表的表現・派生的表現の間で,「派生的表現 (の表記) と代表的表現 (の表記) の間で，前後の形態素の品詞が共通となる用例が存在するか否か」の調査を，人手を介さず自動解析により，事前に行っておく．以下の (q-1)，あるいは，(q-2) における判定は，この結果を参照することにより行う．

(q-1) **「派生的表現 (の表記) と代表的表現 (の表記) の間で，前後の形態素の品詞が共通となる用例が存在する」**ならば，その代表的表現 (の表記) に対する用法判定結果を適用することによって，当該の派生的表現 (の表記) に対して適切な用法判定結果が得られる可能性が高い．

(q-2) **「派生的表現 (の表記) と代表的表現 (の表記) の間で，前後の形態素の品詞が共通となる用例が存在しない」**ならば，当該の派生的表現 (の表記) の用例は，上述 (1) に該当する．ここで，(p) の結果を参照することによって，このクラス (q-2) に該当する用例を網羅的に収集することができる．よって，それらの例外的用例の用法判定は，別途，事前に行っておけばよい．

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

表 3: 派生的表現の表記の用法の類型化

類型		箇所数 (割合 (%))	表現数
代表・派生間で共通の用法	同一の用法の代表的表現の用例が新聞記事に存在 (前後の形態素の品詞及び判定ラベルが一致)	1,451 (66.1)	52
	同一の用法の代表的表現の用例が 作例可能	394 (17.9)	41
	合計	1,845 (84.0)	54
派生的表現の表記に特有の用法		242 (11.0)	10
文字列照合における過検出 (形態素解析における形態素境界と派生的表現の表記の境界が不一致)		108 (5.0)	9
合計		2,195 (100)	56

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．

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．

4. 派生的表現の表記の用法の類型化

4.1 機能表現表記の前後の形態素の品詞を用いた分析の手順

まず，毎日新聞 1995 年の 1 年分を対象として，機能表現表記 637 表記の用例を収集し，人手で用法判定を行った用例，50,064 用例中から，用法判定済の箇所が 50 箇所以上となり，機能的用法の割合が 1~9 割となる代表的表現の表記を選定したところ，94 表記となった．次に，この代表的表現の派生的表現となる表記の用例を，毎日新聞 1995 年の 1 年分より収集し，人手で用法判定を行った後，用法判定箇所数が 10 以上となる派生的表現の表記を収集したところ，56 表記，2,195 用例となった．一方，これらの派生的表現に対する代表的表現の表記数は 46 表記であり，用法判定済み用例は 1,167 用例であった．

本節では，この，

- 派生的表現の表記，56 表記，2,195 用例

を対象として，

- 代表的表現の表記，46 表記，1,167 用例

との間で，

> 派生的表現の表記の各出現箇所と同様の用法となる用例が，対応する代表的表現の表記の用例中に存在するか否か，あるいは，そのような用例が作例可能か否か，

の調査を行い，その結果を分析した．

4.2 分析結果

表 3 に，分析結果を類型化したものを示す．

この結果から分かるように,「対応する代表的表現の表記を対象として，派生的表現の表記の各出現箇所と同様の用法となる用例」が存在する場合が約 66%となった．この場合に該当する用例を表 4(a) に示す．一方，用法判定済の用例中に,「対応する代表的表現の表記を対象として，派生的表現の表記の各出現箇所と同様の用法となる用例」は存在しなかったが，そのような用例が容易に作例可能な場合が約 18%となった．この場合に該当する用例を表 4(b) に示す．また，これらの合計約 84%の用例中で，機能的用法の割合は約 64%であった．

一方，残りの 16%のうちの 11%(10 表記，242 用例) については，表 5 に例を示すように，1 節の (1) で述べた,「派生的表現の表記に特有の用法」であった．これらの全用例は，表 5 に示すように，

(a) L^1 表現「ていく」「てくる」から派生する派生的表現の表記の用例

(b) L^1 表現「ておく」から派生する派生的表現の表記の用例

(c) L^1 表現「てよい」から派生する派生的表現の表記の用例

の三種類に大別できる．これらの用例においては，いずれも，当該表記そのものが「機能表現一覧」[松吉 07] において登録されている機能的用法の意味として用いられているものはない．例外として，(a) L^1 表現「ていく」「てくる」系において，派生的表現の表記「できます」，および,「できまし」が,「可能」の意味の表現「できます」や「ことができまし」の機能的用法と誤って照合したという用例 (全体の 19%) があるが，それ以外の用例においては，いずれも，派生的表現の表記が他の内容的表現と誤って照合したというものである．

表 4: 派生的表現の表記の用法の類型: 「代表・派生間で共通の用法」の例

(a) 同一の用法の代表的表現の用例が 新聞記事に存在 (前後の形態素の品詞及び判定ラベルが一致)

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) (新聞記事から収集)	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) (新聞記事から収集)	意味的等価クラス	用法
名詞 (法律)- でよい- 記号 (」)	名詞 ((ガラス) ふき)- でもよい- 記号 (,)	許可-許可-テヨイ	機能的用法
動詞 ((精神を) 踏まえ)- てよく- 動詞 (できて (いる))	動詞 (非自立)((歩いて) い)- てもよく- 動詞 (わかる)	(「てもよく」が 機能的用法の場合は 「許可-許可-テヨイ」)	内容的用法
動詞 ((覚悟で) 臨んで)- でいく- 記号 (」)	動詞 (膨らん)- でいきます- 記号 (.)	進行-継続-テイク	機能的用法
名詞 (マンツーマン)- でいく- 記号 (.)	名詞 ((ゲリラ) 作戦)- でいきます- 記号 (」)	(「でいきます」が 機能的用法の場合は 「進行-継続-テイク」)	内容的用法

(b) 同一の用法の代表的表現の用例が 作例可能

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) (作例)	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) (新聞記事から収集)	意味的等価クラス	用法
動詞 (進ん)- でよい- 名詞 (こと (になった))	動詞 (休ん)- でもよい- 名詞 (こと (にしている))	許可-許可-テヨイ	機能的用法
副詞 (少し)- でよい- 名詞 (香り (がする)))	副詞 (少し)- でもよい- 名詞 (成績 (を上げよう))	(「でもよい」が 機能的用法の場合は 「許可-許可-テヨイ」)	内容的用法
動詞 (飛ん)- でいけ- 記号 (.)	動詞 (飛ん)- でけ- 記号 (,)	進行-継続-テイク	機能的用法

5. 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, Nagasaka10a, 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．

6. まとめ

本稿では，[長坂 10b] において明らかになった特性として，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，代表的表現と派生的表現の間で用法の傾向に相関がある，という点に注目した．そして，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する代表的表現，およびその派生的表現を対象として，この特徴をより詳細に分析した．その結果，1 節で述べたように，派生

表 5: 派生的表現の表記の用法の類型: 「派生的表現の表記に特有の用法」の例

派生的表現の表記に特有の用法	参考情報: 代表・派生間で共通の「機能的用法」		
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) (「機能的用法」/「内容的用法」)	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) (「新聞記事から収集」/「作例」の別)	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) (「新聞記事から収集」/「作例」の別)	意味的等価クラス

(a) L^1 表現「ていく」「てくる」系，152 箇所/3 表現
代表/派生組: 「でいく，でくる」/「できます」,「でいか，でこ」/「できませ」,
「でいき，でき」/「できまし」

名詞 (持続)-できます-記号 (.) (「可能」の意味の「できます」の「機能的用法」)	動詞 (稼い)-でくる-記号 (」) (新聞記事から漢字表記「で来る」を収集)	動詞 (運ん)-できます-記号 (.) (新聞記事から収集)	進行-継続-テクル
助詞 ((迎えること) が)-できまし-助詞 (て) (「可能」の意味の「ことができまし」の「機能的用法」)	動詞 (呼ん)-でき-助詞 (て) (新聞記事から漢字表記「で来」を収集)	動詞 (乗り込ん)-できまし-助動詞 (た) (新聞記事から収集)	進行-継続-テクル

(b) L^1 表現「ておく」系，85 箇所/5 表現
代表/派生組: 「ておく」/「とく」,「ておい」/「とい」,「ておき」/「とき」,
「ておけ」/「とけ」,「ておこ」/「とこ」

助詞 ((誤解) を)-とく-助詞 (と) (「内容的用法」)	動詞 (つけ)-ておく-記号 (.) (新聞記事から収集)	動詞 (上げ)-とく-助詞 (と) (新聞記事から収集)	過去-完了-タ
助詞 ((パズル) を)-とい-助詞 (て) (「内容的用法」)	動詞 (浸し)-ておい-助動詞 (た) (新聞記事から収集)	動詞 (言っ)-とい-助動詞 (た) (新聞記事から収集)	過去-完了-タ

(c) L^1 表現「てよい」系，5 箇所/2 表現
代表/派生組: 「でいい」/「でもいい」,「でよく」/「でもよく」

文頭-でもいい-助詞 (じゃ(ない)) (「内容的用法」)	名詞 (先)-でいい-記号 (」) (新聞記事から収集)	名詞 (カラオケ)-でもいい-助詞 (から) (新聞記事から収集)	許可-許可-テヨイ

的表現 (の表記) を代表的表現 (の表記) に言い換えた後，用法判定を行う方式 [長坂 08] の実現に向けて，重要な指針を得ることができた．

今後は，1 節で述べた指針にしたがって，まず，準備手順 (p) を行うことにより,「派生的表現 (の表記) と代表的表現 (の表記) の間で，前後の形態素の品詞が共通となる用例が存在しない」という，例外的用例の網羅的収集，および，用法判定を行う．そして，その結果をふまえて，派生的表現 (の表記) を代表的表現 (の表記) に言い換えた後，用法判定を行う方式 [長坂 08] を実現する．

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, pp. 194–197 (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).

[Nagasaka10a] 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*, pp. 1778–1785 (2010).

[長坂 10b] 長坂泰治, 宇津呂武仁, 松吉俊, 土屋雅稔：階層的機能表現辞書に基づく日本語機能表現の分析と検出, 言語処理学会第 16 回年次大会論文集, pp. 970–973 (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, pp. 611–614 (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).