派生語間の意味の差の記述法：記述形式と自動付与

加藤 直樹† 藤田 篤† 佐藤 理史†
†名古屋大学大学院工学研究科
naoki@sslab.nuee.nagoya-u.ac.jp, {fujita,ssato}@nuee.nagoya-u.ac.jp

1 はじめに

流通するテキストデータの大規模化と言語解析技術の成熟を背景に，意味を処理する単位は，近年，語から述語項構造のような複雑な構造へとシフトしつつある．そして，情報検索やテキストマイニング，評判情報処理などの応用における効果も報告されつつある．しかしながら，語のレベルにおいても，個々の語を異なる記号とみなすだけではなく，同義語や反義語，上位・下位語のような語間の関係をとらえることで，複雑化とは違う側面で高度化の余地が残されている．

我々は,「丸い」と「丸める」,「楽しい」と「楽しみ」のように，語幹を共有し，意味的に明確な関連のある語の対（**派生語対**）を収集し，日本語派生語辞書を編纂した [5]．この辞書は，派生語対に見られる「〜い」と「〜める」のような形態的特徴に基づいて収集した 15,126 語対を収録している．収録情報を図 1 に，エントリの例を (1) に示す．

(1) TD-AV-02160, D, 丸い, 形容詞, 丸める, 動詞, 丸, S-い:S-める, maru, S-i:S-meru, A1

個々の派生語対の間には，上述の「〜める」のような接辞付加 [4] によってもたらされる，同義，反義という関係よりも多様性に富む**意味の差**がある．本稿では，派生語辞書の各エントリに，そのような意味の差の情報を記述する方法について述べる．具体的には，

記述形式： 派生語対間の意味の差を，派生先の語と派生元の語を同義とするような言語表現パターンによって表すこととし，

自動付与： そのようなパターンを個々の派生語対に対して自動的に付与する方法を提案する．

それぞれについて，2 節，3 節で説明する．そして，実際に我々の派生語辞書中のエントリに対して意味の差の情報を付与した結果を 4 節で示し，5 節では，その評価について述べる．最後に，6 節でまとめる．

2 意味の差の記述形式

派生は，言語学の語形成に関する分析の中で扱われている．伊藤ら [4] は，派生を『接辞付加による語形成』としており，斎藤 [7] は，形容詞「高い」を例に，「高」を語幹（本文中は語基）とする派生語の形態的特徴を派生プロセスと呼んでいる．

ID (1): 語対タイプ，品詞対タイプ，番号からなる．
- 語対タイプ : 派生接辞の位置．H（語頭）または T（語末）．品詞の同異．D（異品詞）または S（同品詞）．
- 品詞対タイプ : A（形容詞），D（副詞），G（形容動詞），N（名詞），S（サ変名詞），V（動詞）．
- 番号 : 全ての語対に対する通し番号．

収集手法 (2): D（単語辞書からの収集手法），C（コーパスを用いた収集手法），DC（2 つの手法で重複して収集）．
語対の表記と品詞 (3-6): 各語の表記と品詞．
表記共通・差分 (7-8): 語対の表記の共通部分と差分．
読み共通・差分 (9-10): 読み（ローマ字表記）の共通部分と差分．読み差分は派生パターンである．
判定ラベル (11): 人間による判定を経て付与された判定ラベル．

図 1: 日本語派生語辞書の収録情報

我々はこれらをふまえ，派生を，接辞付加などによる語形成と考える．そして，派生先の語が持つ意味は，派生元の語の意味が，接辞の働きによって変化したものであると考える．

語間の意味の差を記述する形式としては，大きく分けて次の 3 つが考えられる．

意味ラベルで記述する： 各語の意味を何らかの意味ラベルで表せば，その差を計算することで，意味の差を記述できる．たとえば，Edmonds [1] や Inkpen [3] は，類義語（near-synonym）の意味の違いを表すモデルを構築し，機械翻訳の訳語選択に用いる方法を示している．しかし，派生によって生じる意味の差の全体像は明らかになっていないため，どのような範囲で，どのくらいの細かさでラベルの集合を定義すれば良いかは自明ではない．

言語表現で記述する： 意味の差を言語表現で記述する場合，国語辞典などの既存の資源やコーパスから，求める表現を収集できる可能性がある．藤田ら [2] は，名詞の類義語間の意味差分を，各語の国語辞典の語釈文の差で表現している．

各語の共起情報で記述する： 言い換え知識獲得の先行研究の成果から，語をそれと共起する表現の束で表し，2 つの語の意味の差を，それらの共起表現の差によって間接的に表すことが考えられる．しかし，この方法には，次に示す 3 つの問題がある．

- 派生語対には異品詞の語対があるため，共起情報を比較することは適切ではない．
- 共起表現の差と，各語の細かな意味の差が対応するとは限らない．
- 統計量によって表された意味の差は，人間にとっ

て難読なものである．

以上をふまえ，我々は，日本語派生語辞書における意味の差を，言語表現で記述することにした．具体的には，**派生先の語を派生元の語と同義とするような言語表現パターン**で記述する．例を (2) に示す．

(2)　派生語対：丸い → 丸める
言語表現パターン：w_D=w_S（連用テ接続）する

これは，派生元の語を含み，派生先の語と同義である言語表現中の，派生元の語の部分を一般化したものに相当する．言語表現パターンでは，派生元の語を w_S，派生先の語を w_D を表し，w_S が活用語の場合は括弧で活用形を指定する．

ただし，派生によって生じる意味の差の理解や，言語処理への応用のためには，意味ラベルによる記述も検討する必要がある．本稿で提案する言語表現パターンは，そのような意味ラベルの設計と記述の足がかりになると考えている．

3　意味の差の自動付与手法

この節では，派生語対に対して言語表現パターンを自動的に付与する手法について述べる．提案手法の基本方針は次の通りである．

1. 国語辞典の語釈文から言語表現パターンを作成し，同じ派生パターンを持つ派生語対に付与する．
2. 付与した言語表現パターンが派生語対の意味の差を表すか否かを，派生語対と言語表現パターンから生成した表現対が同義表現か否かによって検証する．

以下，3.1 節で，言語表現パターンの作成と派生語対への自動付与について，3.2 節で，検証のための同義表現対生成と判定について説明する．

3.1　言語表現パターンの作成と自動付与

国語辞典において，派生先の語が，派生元の語を含む表現で説明される場合がある．したがって，言語表現パターンを獲得できると期待できる．例を (3) に示す．

(3)　見出し語：高める
語釈文：高くする。「公徳心を高める」。

我々は，派生語対の意味の差を接辞の働きによる意味の変化と考え，同じ派生パターンを持つ派生語対の意味の差が，同じ言語表現パターンで表せると仮定する．実際に，2 節で例示した「w_D=w_S（連用テ接続）する」という言語表現パターンは，同じ派生パターン「S-i:S-meru」を持つ派生語対「丸い → 丸める」,「広い → 広める」の意味の差も表す．

これをふまえ，次に示す手順で，国語辞典の語釈文から言語表現パターンを抽出し，派生語対に付与する．

表 1: 言語表現パターン末尾の統一

派生先の語の品詞	言語表現パターンの末尾
形容詞	連体形，名詞が後続する形
形容動詞	連体形，名詞が後続する形
副詞	連用形，動詞が後続する形
名詞	名詞
サ変名詞	動詞の基本形
動詞	動詞の基本形

手順 1. 語釈文の抽出： 派生語対と国語辞典から，⟨ 派生元の語, 派生先の語, 語釈文 ⟩を抽出する．抽出条件を次に示す．

- 見出し語が派生先の語である．
- 語釈文中に派生元の語が存在する．

手順 2. 同義表現の抽出： 手順 1 で抽出した語釈文から，人間が同義表現を抽出する．すなわち，⟨ 派生元の語, 派生先の語, 同義表現 ⟩を得る．抽出する同義表現は，派生先の語とほぼ同じ意味になる，可能な限り短い表現とし，表 1 に示すように，表現の末尾を派生先の語の品詞によって統一する．一つの語釈文から複数の同義表現が得られることもあれば，同義表現と考えられる表現が得られないこともある．

手順 3. 同義表現のパターン化： 手順 2 の出力から，言語表現パターンと，その付与条件をまとめて，⟨ 言語表現パターン, 付与条件 ⟩を出力したもの．

- 言語表現パターン：同義表現中の派生元の語の部分を一般化し，基本的に漢字を使う方針で表記のゆれを除去する．
- 付与条件：⟨ 派生タイプ, 品詞対タイプ, 派生パターン, 派生方向 ⟩である．この条件は，日本語派生語辞書における当該派生語対の情報に基づいて定める．

手順 4. 言語表現パターンの付与： 派生語対に対して，付与条件に合致する全ての言語表現パターンを付与する．

以上の手順における，出力の例を (4) に示す．

(4)　派生語対：高い → 高める
語釈文：高くする。「公徳心を高める」。
同義表現：高くする
言語表現パターン：w_D=w_S（連用テ接続）する
付与条件：⟨TD, AV, S-i:S-meru,→⟩

3.2　言語表現パターンの適切さの判定

ただし，次の例に示すように，同じ派生パターンを持つ派生語対とその意味を表す言語表現パターンは必ずしも 1 対 1 に対応するものではない．

異なる派生パターンで同じ言語表現パターン： 異なる派生パターンを持つ「安らか → 安らぐ」と「静か

表 2: 派生語対候補の数と例（図 1 参照）

語対タイプ	語対数	例
HD	22	黒（名詞）-まっ黒（形容動詞）
HS	1,024	若い（形容詞）-うら若い（形容詞）
TD	7,747	丸い（形容詞）-丸める（動詞）
TS	6,333	可愛い（形容詞）-可愛らしい（形容詞）
合計	15,126	-

→ 静まる」の意味の差が，同じ「w_D=w_S になる」という言語表現パターンで表せる．

同じ派生パターンで異なる言語表現パターン：「印象 → 印象的」の意味を表す言語表現パターン「w_D= 強い w_S をあたえられる」は，同じ派生パターンを持つ「営利 → 営利的」の意味の差を表さない．

3.1 節の手順で言語表現パターンを付与する場合，後者のような，派生語対の意味の差を表さないものを除外する必要がある．

ある言語表現パターンがある派生語対の意味の差を表すか否かは，言語表現パターン中の，w_D，w_S を，派生語対に置き換えたもの（同義表現対候補）が同じ意味になるか否かによって判定できる．例を (5) に示す．

(5) a. 派生語対：安らか → 安らぐ
言語表現パターン：w_D=w_S になる
同義表現対候補：安らぐ ⇔ 安らかになる
⇒ **言語表現パターンは意味の差を表す**

b. 派生語対：営利 → 営利的
言語表現パターン：
w_D= 強い w_S をあたえられる
同義表現対候補：
営利的 ⇏ 強い営利をあたえられる
⇒ **言語表現パターンは意味の差を表さない**

2 つの表現間の同義性を，各表現と共起する表現の分布の比較に基づいて計算することが考えられる．ただし今回は，そのような同義性判定は辞書を仕上げる際の人間の判断に委ねることにし，例 (6) のような，明らかに不適格な表現を生成してしまう言語表現パターンのみを自動的に除外することにした．

(6) a. 派生語対：少ない → 少なくとも
言語表現パターン：
w_D=w_S（連用テ接続）なったとしても
同義表現対候補：
少ない，少な**くっ**なったとしても

すなわち，自動的に付与した言語表現パターンの適否を，次の手順で判定する．

手順 1. 同義表現対候補の自動生成： 各派生語対と，付与した言語表現パターンから，同義表現対候補を自動生成する．

表 3: 対象派生語対の品詞対（図 1 参照）

語対タイプ	品詞対タイプ	語対数	品詞対タイプ	語対数
HD	GN	3	-	
TD	AD	18	DV	71
	AG	129	GN	1,070
	AN	499	GS	283
	AV	154	GV	28
	DG	7	NS	310
	DN	16	NV	1,115
合計				3,703

手順 2. コーパスによるフィルタリング： 各同義表現対候補の，w_S を含む表現のコーパスにおける出現頻度を調べ，閾値以上の表現を生成するような言語表現パターンのみを出力する．

4 意味の差の自動付与結果

提案手法によって，日本語派生語辞書の各語対に意味の差の情報を付与した．4.1 節で，対象とする派生語対について，4.2 節で，自動付与の結果について述べる．

4.1 意味の差を付与する対象の派生語対

我々が作成した日本語派生語辞書は，15,126 語対の派生語対候補を収録している [5]．派生語対候補は，派生接辞の位置が語頭であるか語末であるか，また，各語の品詞の同異によって分類できる．それぞれの語対数と例を表 2 に示す．

ここから，意味の差を付与する対象の派生語対を，次の手順で選定した．

1. 異品詞の派生語対候補から，派生語対でないという判定ラベルを持つ語対を除去する．
2. 同じ語幹を持つ派生語対候補をグループ化し，各グループ内で派生元の語を 1 つ決定する．
3. 「派生元の語 →∗」となる派生語対候補を抽出する．
4. 複数の派生の組み合わせで表される派生語対候補を除去する．
5. 前稿 [5] で示した基準に沿って，残った派生語対候補の適否を判定する．

この結果，派生方向を付与した 3,703 語対の派生語対を得た．品詞対の内訳を表 3 に示す．

4.2 結果

4.1 節で得た 3,703 語対に対して，意味の差を表す言語表現パターンを付与した結果について述べる．

4.2.1 言語表現パターンの作成と付与

国語辞典として，岩波国語辞典第五版（電子化データ）[6] を用いた．

対象派生語対に対応した ⟨ 派生元の語, 派生先の語, 語釈文 ⟩ 439 個から，292 個の ⟨ 言語表現パターン, 付与条件 ⟩ を作成した．付与条件に合致する派生語対に，言語表現パターンを付与した結果，派生語対 2,987 語

「同義表現対候補」に対して，
設問 1. そのまま置き換えて同義になる文が思い付く．
Yes: 【o1】．No : 設問 2 へ．
設問 2. 格の変化をともなって同義になる文が思い付く．
Yes: 【o2】．No : 【x】．

図 2: 同義表現対候補の適否判定基準

対に対して，92,658 個の言語表現パターンを付与した．

4.2.2 言語表現パターンの適切さ判定

派生語対と言語表現パターンから，92,658 件の同義表現対候補を生成した．これらのうち，新聞コーパス（毎日新聞 1991〜2005 年版）によるフィルタリングを通過したのは，派生語対 2,215 語対に対応する 9,019 表現（平均 4.1 表現）であった．

5 サンプリング評価

自動付与した言語表現パターンのカバレージを，サンプリング評価した．

同義表現対候補が生成された派生語対をランダムサンプリングし，その派生語対に付与されている全ての言語表現パターンの適否を，そこから生成される同義表現対候補が同義か否かによって人間が判定した．具体的には，1 人の判定者が，図 2 の判定基準に従って，同義表現である (“o1”，“o2”)，もしくは，同義表現でない (“x”) というラベルを付与する．例を (7) に示す．

(7) a. 派生語対：喜劇 → 喜劇的
同義表現対候補：喜劇的，喜劇のような
思い付いた文：**喜劇的な**結末を向かえる，**喜劇のような**結末を向かえる ⇒ **o1**

b. 派生語対： 弱い → 弱さ
同義表現対候補：弱さ，弱いこと
思い付いた文：部品 の **弱さ**が問題だ，部品 が **弱いこと** が問題だ ⇒ **o2**

図 2 の設問 2. において，格の変化があっても同義表現であるとした．これは，同義表現対候補を置き換えたとき，格の交替が起こる場合が事前に想定されたためである．同義表現対候補が “o1”，“o2” と判定された場合，それを生成した言語表現パターンは，派生語対の意味の差を表していると考える．

同義表現対候補を付与した派生語対 2,215 語対の約 10%にあたる 220 語対をランダムサンプリングし，それらに対する 916 件の同義表現対候補の適否を判定した．その結果を表 4 に示す．1 つ以上の同義表現が生成されている語対は 164 語対であった．同義表現候補を自動付与した派生語対の 75%に，1 つ以上の同義表現を，つまり，派生語対の意味を表す言語表現パターンを付与できたことになる．

同義表現でないと判定されたものの大部分は，3 節で述べた，派生パターンと言語表現パターンの不一致の影響によるものだった．派生先の語の語義によって付与すべき言語表現パターンは様々であり，これを区別して生成ルールを作ることは難しい．同義表現対候補のフィルタリングのために，同義性を機械的に計算する技術が必要である．

表 4: 同義表現対候補の判定結果

判定ラベル	候補数	同義表現対候補の例	
o1	183	喜劇的（形容動詞）	喜劇のような
o2	30	弱さ（名詞）	弱いこと
x	703	試験的（形容動詞）	試験しているような
合計	916	-	

対象派生語対のうち，1,488 語対に対しては言語表現パターンを 1 つも付与できなかった．これについては次の 2 つの原因がある．

1. 付与条件がカバーしていたが，得られなかった．
2. 付与条件がカバーしていなかった．

新たな ⟨ 言語表現パターン, 付与条件 ⟩ を作成するだけでなく，言語表現パターンを付与する条件についての検討が必要である．

6 おわりに

本稿では，派生語対の意味の差の記述法を提案した．具体的には，派生先の語と派生元の語を同義とするような言語表現パターンによって意味の差を記述することにした．そして，そのようなパターンを国語辞典の語釈文から抽出，派生語対に自動付与し，これまでに，日本語派生語辞書から選定した異品詞の派生語対 3,703 語対のうち，2,215 語対に対して 9,019 個の言語表現パターンを付与できた．サンプリング評価の結果，同義表現対候補が得られた場合は，その 75% (164/220) に対して，1 つ以上同義表現が得られていた．

参 考 文 献

[1] P. Edmonds and G. Hirst. Near-synonymy and lexical choice. *Computational Linguistics*, Vol. 28, No. 2, pp. 105–144, 2002.

[2] 藤田篤, 乾健太郎. 語釈文を利用した普通名詞の同概念語への言い換え. 言語処理学会第 7 回年次大会発表論文集, pp. 331–334, 2001.

[3] D. Inkpen and G. Hirst. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, Vol. 32, No. 2, pp. 223–262, 2006.

[4] 伊藤たかね（編）. 文法理論:レキシコンと統語. 東京大学出版会, 2002.

[5] 加藤直樹, 藤田篤, 佐藤理史. 日本語派生語辞書第一版の編纂. 言語処理学会第 14 回年次大会発表論文集, pp. 1053–1056, 2008.

[6] Real World Computing Project. RWC テキストデータベース第 2 版, 岩波国語辞典タグ付き/形態素解析データ第 5 版, 1998.

[7] 斎藤倫明. 現代日本語の語構成論的研究. ひつじ書房, 1992.