日本語派生語辞書第一版の編纂*

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

言語には同じ意味を表す複数の表現がある．計算機で言語を柔軟に扱うには，異なる表現が同じ意味を持つことを認識する，ある表現から他の表現を生成するといった言い換え技術が不可欠である．

様々な言い換えを計算機で扱うためには，語に関する知識と，語間の関係に関する様々な知識が必要である．本稿では，そのような語間の関係のうち，例 (1) に示すような言い換えに必要な，派生関係を扱う．

(1) a. 部屋は十分**暖かい**⇔ 部屋は十分**暖まっている**
b. 身体**がだるいと**感じる ⇔ 身体**のだるさを**感じる
c. 円**の**レート**が下がった**⇔ 円**が**レート**を下げた**

英語に関しては，派生関係を収録した大規模な資源が存在する．WordNet [1] は様々な語間の関係の資源であり，同義，対義，上位–下位，部分–全体，論理的含意とともに派生関係を扱っている．また，CatVar [2] は，派生関係を持つ語の集合をまとめた大規模な資源である．一方，日本語に関しては，厳密に派生を扱った実用的な資源は存在しない．例えば，計算機用日本語辞書 IPAL [4, 5] に派生関係の記述があるが，基本語を対象とした小規模な資源である，比較的緩やかな基準で派生を定義して扱っているという問題がある．一部の書籍体の辞書には見出し語から派生する語についての記述があるが，一部の接辞による派生しか扱っていなかったり，そのような接辞による全ての派生語を収録していなかったりするため，辞書中の記述を書き起こしても網羅的な資源にはならない．

そこで本稿では，日本語の派生語辞書の編纂について述べる．自然言語処理での応用を考えた結果，語とその派生語の対（これを**派生語対**と呼ぶ）を収集し，これを 1 エントリとすることにした．

派生という現象は十分に整理されていない．例えば「暖かい」と「暖かさ」のような，一部の派生は良く知られているが，派生とみなすべき現象の範囲についての明確な定義は，我々が調査した限り存在しない．派生の範囲を明確にし，その範囲の語対を網羅的に収集することが課題である．

*A Japanese Lexical Derivation Database.
Naoki Kato†, Atsushi Fujita†, Satoshi Sato†
†Graduate School of Engineering, Nagoya University

本稿では，2 節で派生とその周辺現象について述べた後，3 節，4 節で日本語派生語辞書の編纂手法について述べる．最後に，5 節で日本語派生語辞書第一版の具体的編纂手順と結果物の諸元について述べる．

2 派生とその周辺現象の整理

派生という現象は，言語学的にも工学的にも十分に整理されていない．そこでまずは，派生の定義とその周辺の現象を整理し，派生語対を収集するための手順について述べる．

言語学においては，語形成に関する分析の中で，派生が取り上げられている．伊藤ら [3] 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（文献中は語基）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．以上より，次のことが言える．

- 接辞という形態的特徴によって派生を説明できる．
- 接辞ごとに，付加できる語の品詞が決まっている．
- 派生を表す形態的特徴は，複数の派生語対に共通であるが，あらゆる語に適用可能ではない．
  e.g. 「S-i:S-meru」という形態的特徴は,「高い」–「高める」,「広い」–「広める」など多くの派生語を作るが,「美しい」–「美しめる」とはならない．

これらから，ある語対が派生語対であるための必要条件として,「S-i:S-meru」のような形態的特徴（以下，**派生パターン**と呼ぶ）を持つことがあげられる．しかし，派生の根拠となるような接辞あるいは派生パターンの全体像は自明ではないため，派生語対を網羅的に収集する際の事前知識とはならない．

前述の「S-i:S-meru」という派生パターンは,「楽しい」–「楽しめる」という語対にも対応する．しかし，この語対の関係は形容詞と動詞の可能形であり，派生ではない．他にも，表 1 に示すような，様々な現象が派生と同様に形態的特徴で表される．このように，形態的特徴を参照するだけでは，派生語対とその他の関係を持つ語対を分類することはできない．

以上の問題を考慮して，以下の 2 ステップで語の集合から派生語対を収集することにした．

ステップ 1. 派生語対候補の自動収集： 派生パターンが複数の派生語対に共通であることから，なんらかの形態的特徴で表される（共通の語幹を持つ）語対を

表 1: 派生とその境界の現象

現象	例
派生	強い–強さ，暖かい–暖める
活用 サ変名詞＋「する」	早い–早く，哀れむ–哀れみ 運転–運転する
複合	早い–早すぎる．輪–輪投げ 取る–取り組む，鰹–初鰹
副詞＋「する」 動詞＋「たい」	ぐずぐず–ぐずぐずする 食べる–食べたい
受身／自発 可能 使役	悔やむ–悔やまれる 泳ぐ–泳げる 上がる–上がらせる
その他	待つ–待ち遠しい，苦労–ご苦労な

表 2: 派生語対の性質による分類

	異品詞	同品詞
語頭の接辞 による派生	解決–未解決 黒–まっ黒	弱い–か弱い 捨てる–うち捨てる
語末の接辞 による派生	強い–強さ 暖かい–暖める	広がる–広げる 黒い–黒っぽい

網羅的に収集し，より多くの語対に対応する形態的特徴を持つ語対から順に派生語対候補とする．

ステップ 2. 人手による分類： 派生とその周辺現象の分類と語対の意味的関連の有無による，派生語対の分類基準を作成しておき，それに従って派生語対候補を分類する．

各ステップについて，3 節，4 節で述べる．

3 派生語対候補の自動収集

この節では，語対の形態的特徴を手がかりに派生語対候補を自動収集する 2 つの手法について述べる．1 つは，語の集合を入力として，その中に存在する派生語対候補を網羅的に収集する手法，もう 1 つは，語の集合と，いくつかの生産性の高い派生パターンを入力として，派生語対候補を収集する手法である．ここで，語の集合には，各語に品詞，表記，読み（ローマ字表記）の情報が付与されているとする．

3.1 単語辞書からの収集

派生語対は，派生接辞の位置と，各語の品詞の同異によって表 2 のように分けることができる．接辞ごとに付加できる語の品詞が決まっていることから，品詞対ごとに語対を収集する．この収集手法は，語末の派生による異品詞の語対を対象とした手法 [6] を，表 2 の全ての部分に拡張したものである．

次の 4 ステップで派生語対候補を収集する．

1. 語幹を共有する語対の列挙： 品詞対ごとに，次の条件を全て満たす語対を網羅的に収集する．

- 語幹を共有（表記および読みが前方一致または後方一致）する．
- 各語に含まれる漢字とその順序が一致する．
  e.g. 「強い（tuyoi）」–「強める（tuyomeru）」なら語幹は「強（tuyo）」,「弱い（yowai）」–「か弱い（kayowai）」なら語幹は「弱い（yowai）」.

表 3: 品詞の証拠となる後続文字列

品詞	後続文字列
形容動詞	だろ，だっ，で，に，だ，な，なら
名詞	が，を，に，と，で，へ，から，より，まで，の
サ変名詞	し，さ，せ，する，すれ，しろ，せよ

2. 派生パターン候補の付与： 収集した全ての語対に対して，ローマ字表記を参照して差分を結合したものを，派生パターン候補として付与する．
  e.g. 「強い」–「強める」なら「S-i:S-meru」,「弱い」–「か弱い」なら「ϕ-S:ka-S」という派生パターン候補.

3. サポート数の付与： 各語対の派生パターン候補に，対応する語対の異なり数（以下，サポート数）を付与する．
  e.g. 「S-i:S-meru」という派生パターン候補が，他にも「丸い」–「丸める」,「痛い」–「痛める」などの 25 種類の語対に対応すれば，サポート数は 25.

4. サポート数による語対の選別： 2 節で述べた知見より,『多くの語対に対応する形態的特徴を持てば，派生語対である可能性が高い』と仮定し，サポート数に閾値を定め，閾値以上の語対を出力する．

3.2 コーパスを用いた収集

形容詞に付加する名詞化接辞「さ」は,「暖かい」,「嬉しい」など多くの形容詞に付加して派生語を作る．このような生産性の高い派生パターンを用いて，派生語対を次の 2 ステップで収集する．

1. 候補語対の生成： 派生元の品詞の各語に，与えられた派生パターンを適用し，候補語対を生成する．
  e.g. 「S-i:S-sa」という派生パターンを形容詞に適用して,「暖かい」–「暖かさ」,「嬉しい」–「嬉しさ」を生成.

2. 語らしさの検証： 生成した語が，その品詞の語であるかどうかを検証する．具体的には，当該品詞の証拠となるような後続文字列（表 3）と組み合わせた文字列の，コーパス中の出現頻度を調べる．頻度が閾値以上であれば，候補語対を出力する．

語の集合によっては，特定の（とくに生産的な）派生接辞による派生語を語として収録していない場合があるが，この手法によって，入力する語の集合に含まれていない語による語対も収集することができる．

4 人手による分類

3 節の手法で機械的に収集した派生語対候補は，2 節で述べたように派生以外の様々な語対を含んでいる．また,「浅い」–「浅ましい」のような，明確な意味的関連がない語対も含んでいる．

自動収集した派生語対候補から，これらの語対を除去

設問 1. 各語の意味が分かる．
　Yes: 設問 3 へ．No: 設問 2 へ．
設問 2. 少なくとも一方が語ではない．
　Yes: 派生語対でない (C1)．No: 判定保留 (B1)．
設問 3. 語幹の表記・読みともに共通である．
　Yes: 設問 4 へ．No: 派生語対でない (C2)．
設問 4. 各語の品詞は正しい．
　Yes: 設問 5 へ．No: 派生語対でない (C3)．
設問 5. 一方のみが，複合語または可能・使役の接辞を持つ．
　Yes: 派生語対でない (C4)．No: 設問 6 へ．
設問 6. 語間に明確な意味的関連がある．
　Yes: 設問 7 へ．No: 派生語でない (C5)．
設問 7. 派生元はどちらか．
　左: A1，右: A2，分からない: A3，
　共通の語からの派生語: A4．

図 1: 派生語対候補に対する分類基準

表 4: 単語辞書からの収集結果

	異品詞	同品詞
語頭の接辞による派生	22	961
語末の接辞による派生	4,769	6,355
合計		12,107

するため，7 つの質問で語対を分類する分類基準（図 1）を作成した．設問 1，2，4 が語の集合のノイズを除去するためのもの，設問 3，5，6 が派生語対であるかの判断，設問 7 が派生方向を定めるものである．判定者は，各派生語対候補をこの基準に照らして分類し，派生語対である (A)，派生語対ではない (C)，判定保留 (B) というラベルを付与する．

5　日本語派生語辞書第一版の編纂

この節では，日本語派生語辞書第一版の，編纂手順と具体的な入力，辞書の諸元について述べる．

5.1　手順

以下の手順で日本語派生語辞書を編纂した．

ステップ 1. 派生語対候補の自動収集

1. 3.1 項で述べた手法で派生語対候補を収集した．収集結果を表 4 に示す．入力となる語の集合を，形態素解析器用辞書 IPADIC[1] の 5 品詞（名詞，動詞，形容詞，形容動詞，副詞）の中の，漢字を含む 61,895 語とし，サポート数による閾値を 2 とした．ただし，次の 2 つの条件のいずれかに該当する候補は出力しなかった．
   - 完全に一致する語対（派生パターンが「S-ϕ:S-ϕ」または「ϕ-S:ϕ-S」）．
   - 語頭の派生による語対に関して，片方の語がもう一方の語の部分文字列でない語対（派生パターンが「ϕ-S:$*$-S」または「$*$-S:ϕ-S」ではない．ただし “$*$” は任意のアルファベット列）．

[1] http://chasen-legacy.sourceforge.jp/

表 5: 生産的派生パターンとコーパスを用いた収集結果

品詞対	派生パターン	語対数	例
形容詞	S-i:S-ge →	27	誇らしい–誇らしげ †
–形容動詞	S-i:S-sou →	68	難しい–難しそう †
形容詞–名詞	S-i:S-sa →	237	強い–強さ †
	S-i:S-mi →	37	痛い–痛み †
	S-i:S-me →	40	甘い–甘め †
	S-rasii:S-ϕ ←	317	男らしい–男
	S-Qpoi:S-ϕ ←	36	黒っぽい–黒
副詞–動詞	S-ϕ:S-suru →	218	騒然と–騒然とする
形容動詞	ϕ-S:hu-S →	54	透明–不透明
–形容動詞	ϕ-S:mu-S →	5	節操–無節操
	ϕ-S:hi-S →	5	公式–非公式
	ϕ-S:mi-S →	3	完成–未完成
形容動詞	S-teki:S-ϕ ←	1,103	基本的–基本
–名詞	S-ϕ:S-sei →	113	可能–可能性
	S-ϕ:S-sa →	352	大切–大切さ
	S-ϕ:S-mi →	12	温か–温かみ
	S-ϕ:S-me →	8	軟らか–軟らかめ
形容動詞	S-teki:S-ϕ ←	355	圧倒的–圧倒
–サ変名詞	S-ϕ:S-ka →	103	効率–効率化
名詞–サ変名詞	S-ϕ:S-ka →	498	概念–概念化
合計		3,501	

表 6: 異表記の統合

異表記の種類	統合した語対の例
かなと漢字	「気づかわしい」–「気づかう」 と「気遣わしい」–「気遣う」 →「気づかわしい/気遣わしい」–「気づかう/気遣う」
送り方の違い	「呼び掛ける」–「呼び掛け」 と「呼掛ける」–「呼掛け」 →「呼び掛ける/呼掛ける」–「呼び掛け/呼掛け」
「々」	「軽軽と」–「軽軽しい」 と「軽々と」–「軽々しい」 →「軽軽と/軽々と」–「軽軽しい/軽々しい」

2. 3.2 項で述べた手法で派生語対候補を収集した．得られた派生語対候補の分布と例を表 5 に示す．ここでは，語の集合として IPADIC，20 個の派生パターン，コーパスとして毎日新聞データ集 1991〜2005 年版を用いた．語の集合には，前述の 61,895 に語に加えて，サ変名詞で漢字を含む 10,070 語を用い，コーパス中の頻度についての閾値は 10 とした．
3. 2 つの手法により収集した派生語対候補の和集合を対象とし，異表記と考えられる語対をまとめた．具体的には，表 6 に示すような語対を統合した．

このようにして収集した派生語対候補の数を表 7 に示す．これが，日本語派生語辞書の収録語対数である．

ステップ 2. 人手による分類

1. 収集した派生語対候補の一部を，人手で分類した．機械的判断が難しく，人手の分類が不可欠である範囲（表 5 と表 7 の † を付けた部分）の合計 1,346 語対を，2〜4 人の判定者が図 1 の分類基準に従って分類し，ラベルを付与した．
2. 分類結果から最終判定ラベルを決定した．複数の判定者によるラベルが一致した場合はそのラベル付与し，一致しない場合は多数決でラベルを決定した．

表 7: 日本語派生語辞書の規模

	品詞対	語対数	例	
頭／異	形容動詞–名詞	19	まっ白–白	†
頭／同	形容詞–形容詞	17	若い–うら若い	†
	副詞–副詞	4	少し–もう少し	†
	形容動詞–形容動詞	69	確か–不確か	†
	名詞–名詞	530	機嫌–ご機嫌	
	動詞–動詞	357	殴る–ぶん殴る	
末／異	形容詞–副詞	32	惜しい–惜しくも	†
	形容詞–形容動詞	156	怪しい–怪しげ	†
	形容詞–名詞	914	哀しい–哀しさ	†
	形容詞–動詞	257	楽しい–楽しむ	†
	副詞–形容動詞	20	意外と–意外	
	副詞–名詞	150	時に–時	
	副詞–動詞	290	願わくは–願う	†
	形容動詞–名詞	1,598	音楽的–音楽	†
	形容動詞–サ変名詞	457	威圧的–威圧	
	形容動詞–動詞	85	哀れ–哀れがる	
	名詞–サ変名詞	497	映画–映画化	
	名詞–動詞	3,471	哀しみ–哀しむ	
末／同	形容詞–形容詞	56	安い–安っぽい	†
	副詞–副詞	55	少し–少しも	†
	形容動詞–形容動詞	11	細か–細やか	†
	名詞–名詞	1,379	哀れ–哀れみ	
	動詞–動詞	4,716	哀れむ–哀れがる	
合計		15,140		

表 8: 最終判定ラベルと語対数

最終判定ラベル	語対数	最終判定ラベル	語対数
A1	652	C1	105
A2	181	C2	2
A4	76	C3	15
A5	17	C4	174
		C5	124
A の計	926	C の計	420
合計			1,346

B1（保留）や A3（派生の方向性が分からない）を残さないために，複数の判定者で再協議して最終判定ラベルを決定した．また，例えば,「格好いい」–「格好よい」のような語対に，同義であることを示す A5 という最終判定ラベルを，これも複数の判定者で協議して決定した．

最終判定ラベルとその語対数を表 8 に示す．今回の判定で，962 語対に派生語対であることを示すラベルを付与した．

5.2 収録情報

日本語派生語辞書の 1 エントリは以下のような収録情報を持つ．エントリの例を (2) に示す．

(2) TD-AV-02283, D, 暖かい, 形容詞, 暖める, 動詞, 暖, S-い:S-める, atata, S-kai:S-meru, A1

ID : 語対・品詞タイプ，品詞対タイプ，番号からなる．

- 語対・品詞タイプ : 派生接辞の位置によって H（語頭）または T（語末），品詞の同異によって 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 : 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 : 各語の表記と品詞．ただし，語の順番は，(a) 品詞を表すアルファベット順，(b) 接辞の長さ順，(c) 接辞のアルファベット順，という条件で決定している．

表記共通・差分 : 語対の表記の共通部分と差分．

読み共通・差分 : 読み（ローマ字表記）の共通部分と差分．

最終判定ラベル : 人手での判定を経て付与された最終判定ラベル．

6 おわりに

本稿では，日本語派生語辞書を編纂するために必要な派生とその周辺現象を整理し，編纂手法を述べた．この手法に従い，我々は，これまでに 15,140 語対の派生語対候補を自動的に収集した．そのうち 1,346 語対については，派生関係にあるかどうかを人手で判断し，926 語対に派生語であるというラベルを付与した．今回編纂した日本語派生語辞書は，近日中に公開予定である．

今後の課題は，語対の意味的関係や差分についての情報を付与することである．意味的関係の分類体系，および効率的な付与手法について検討する予定である．

本研究の一部は次の研究費の支援を受けている：科研費基盤研究 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」（課題番号：16200009，代表：佐藤理史）および科研費若手研究 (B)「文法カテゴリ交替を裏付ける語彙特性の体系化と辞書記述」（課題番号：18700143，代表：藤田篤）．

参 考 文 献

[1] C. Fellbaum. A semantic network of English verbs. In C. Fellbaum, editor, *WordNet: an electronic lexical database*. The MIT Press, 1998.

[2] N. Habash and B. J. Dorr. A categorial variation database for English. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 17–23, 2003.

[3] 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.

[4] 情報処理振興事業協会技術センター. 計算機用日本語動詞辞書 IPAL (Basic Verbs) －解説編－. 情報処理振興事業協会技術センター, 1987.

[5] 情報処理振興事業協会技術センター. 計算機用日本語形容詞辞書 IPAL (Basic Adjectives) －解説編－. 情報処理振興事業協会技術センター, 1990.

[6] 加藤直樹, 藤田篤, 佐藤理史. 語末の形態的特徴に基づく日本語派生語対の収集. 言語処理学会第 13 回年次大会発表論文集, pp. 352–355, 2007.

[7] 斎藤倫明. 現代日本語の語構成論的研究. ひつじ書房, 1992.