日本語イディオム異形規則の構築

山田 翔平[†], 矢田 竣太郎[†], 宮田 玲[†], 竹内 孔一[‡], Ulrich Apel[♮], Wolfgang Fanderl[♮], 村山 遼[†], Iris Vogel[♯], 足立 諒子[†], 影浦 峡[†]

[†] 東京大学大学院教育学研究科　[‡] 岡山大学大学院自然科学研究科
[♮] Tübingen Eberhard Karls University　[♯] Universität Hamburg

1 はじめに

イディオムは言語表現においてかなりの比重を占めているが，翻訳や言語学習においては比較的熟練した者でも困難を感じることが少なくない。さらにイディオムの異形の存在も困難を増す原因となっている。言語処理分野では異形を含めたイディオムの自動マッチング (辞書引き等) 技術が提案されている [1, 2, 3] が，対象言語範囲も適用プラットフォームも限られており，翻訳者や言語学習者などが簡単に使える，異形を含めたイディオムの辞書引き環境はあまり多くはない。これは現実的な観点からは，各言語において簡単に異形規則を作成する環境がないことも一因であると考えられる。

著者らはオンラインの翻訳支援システム「みんなの翻訳」[1] を開発・運用しており [4]，そこでは英語イディオムの異形を含む自動辞書引き機能が実装されている [3]。他の言語でも同様のイディオムの異形を含む自動辞書引き機能を実現することを目的とし，現在は日本語においての実装を目指している。これまでに，イディオムの異形のタイプにおいて最も数の多い「挿入」[5] を対象にして，異形規則の簡単な作成を可能にすることを目指したプラットフォーム QRidiom の開発 [6]，QRidiom 上で用いられる異形用例の作成 [7] を行ってきた。異形のタイプの挿入とはイディオムを構成する品詞の間に品詞が挿入されることである。異形用例とはイディオムの異形を含む文例であり，異形規則とはイディオムの異形の検出を可能にする品詞の出現規則を指す。本稿では，異形用例の作成と，それを用いた異形規則の構築について報告する。

2 日本語異形規則の構築

挿入タイプの異形用例を用い，QRidiom 上で異形規則の作成を行った。

[1] http://trans-aid.jp/

2.1 異形用例の作成

本研究では竹内ら [6] を引き継ぎ，5 つの日本語イディオムの品詞パターンを対象とする。QRidiom が参照している和独辞典 WaDokuJT [8] には 3916 個の日本語イディオムが登録されているが，それらのイディオムの品詞パターンを，含まれるイディオムの数が多い順に並べたのが表 1 である。5 つのイディオムの品詞パターンとは，このうちの 1，3，4，7，8 である。

異形用例は人間の内省により作成したものを用いた。これは，Miyata et al. [7] において，挿入タイプのイディオム異形の用例を，コーパスを用いた方法と人間の内省を用いた方法の 2 通りで作成し，その結果を比較したところ，コーパスを用いるより，実際にありうる異形を考えるという人間の内省を用いた方法の方がより多くの異形用例を作成可能であることが明らかになったことに基づく。各イディオムの品詞パターンの異形用例の数については表 2 の通りである。

表 1: WaDokuJT の日本語イディオムの品詞パターン

品詞パターン	イディオム数
1. Noun-Particle-Verb	1553
2. Noun	299
3. Noun-Particle-Noun	267
4. Noun-Particle-Adjective	160
5. Noun-Particle-Noun-Particle-Verb	148
6. Noun-Noun	100
7. Noun-Particle-Verb-Auxiliary	94
8. Noun-Particle-Verb-Verb	47
9. Noun-Noun-Particle-Verb	41
10. Noun-Particle-Noun-Particle	38

ここで，正例とは挿入が行われてもイディオムの意味が失われない文例，負例とは挿入により本来のイ

表 2: 各品詞パターンの異形用例数

品詞パターン	正例	負例
Noun-Particle-Verb	208	110
Noun-Particle-Noun	118	89
Noun-Particle-Adjective	181	85
Noun-Particle-Verb-Auxiliary	238	123
Noun-Particle-Verb-Verb	205	90

ディオムの意味が失われている文例である。例えば,「頭を冷やす」という Noun-Particle-Verb 型の品詞パターンのイディオムであれば，正例として「頭をまず先に冷やしてから考え直す。」，負例として「冷蔵庫がない時代は，頭を使って冷やす方法を考えていました。」が含まれる。

2.2 方法

岡山大学と東京大学の 2 つの作業グループでそれぞれ QRidiom による異形規則の作成を行った。Tübingen University の日本学専攻の学生にも行ってもらう予定であったが，日本語学習者にとって，異形規則の識別は困難であり，QRidiom も使いこなせないことがわかったため，作業は日本語母語話者にて行うこととした。また，竹内ら [6] においては各品詞パターンにおいて，挿入が Particle（助詞）の前か後かの 2 通りで区別していたが，これら 2 通りを包括的に扱うために助詞は異形規則を構成する品詞としては利用しないこととした。

両作業グループに共有された異形規則の作成の手順は以下の通りである。

1. 各品詞パターンにおいて，正例を検出する異形規則を作成する
2. 1 で作成した異形規則に対して負例を当てはめ，負例を検出しないよう規則の調整を行う。ただしあくまで正例の検出を優先し，負例へのオーバーマッチングは許容する

2.3 結果

両作業グループの異形規則数は表 3 の通りである。両作業グループの異形規則は，いずれも正例を全て検出するものとなった。ただし，形式上 2 つの点で異なりがあった。

1 点目は，異形規則の複合度に関してである。岡山大学側は，挿入される品詞の数に 0 個以上を用い，1 個で複数の挿入可能な品詞のパターンを検出できる複合的な異形規則を作成した。一方，東京大学側は挿入される品詞は 1 個以上を条件とし，1 つの挿入可能な品詞のパターンに対して 1 個の異形規則を作成した。具体的には，東京大学側が Noun-Particle-Verb 型の品詞パターンの異形規則に,「名詞（1 個以上）」の挿入を許容する異形規則,「動詞（1 個以上）-名詞（1 個以上）」の挿入を許容する 2 個の異形規則を設けたのに対して，岡山大学側は「動詞（0 個以上 1 個以下）-名詞（1 個）」の挿入を許容するという 1 個の異形規則で対処している。

2 点目は，挿入される品詞の数の上限に関してである。上記の例にもあるように，岡山大学側は異形用例の正例を検出する範囲で挿入される品詞の個数に上限を設けたのに対し，東京大学側は，異形用例中の正例以外の文法上存在可能な挿入を考慮し，挿入される品詞の数に上限を設けなかった。

表 3: 各大学作成の異形規則数

品詞パターン	岡山	東京
Noun-Particle-Verb	19	26
Noun-Particle-Noun	20	34
Noun-Particle-Adjective	19	27
Noun-Particle-Verb-Auxiliary	19	25
Noun-Particle-Verb-Verb	30	42

3 日本語異形規則の整備

両作業グループで構築した異形規則を，協議の上で一貫した規則として整備した。

3.1 方針

2.3 の通り，構築作業者間での相違は次の 2 点にまとめられる。

- 品詞数 0 以上の設定を用いて複合的な異形規則を作るかどうか
- 異形用例で実際に出現した品詞数を挿入規則の上限とするかどうか

異形規則の整備にあたって，異形検出機能が最終的には「みんなの翻訳」へ実装されることを考慮し，以下のような方針を設定した。

(i) 異形規則の品詞数は 1 以上を基本とし，異形規則の複合は行わない
(ii) 挿入可能品詞について上限は設けない

(i) が意図するところは，異形規則を管理する際の保守性である。あえて異形規則の複合を行わないことで，人間にとっての可読性を高め，将来的な異形規則の追加・削除における人的ミスを予防する。これは異形規則の数の大小が QRidiom におけるイディオム異形検出性能にほとんど影響を与えなかったことにも立脚している。例えば Noun-Particle-Adjective 型の品詞パターンにおいて,「副詞（0 個以上 1 個以下）-形容詞（1 個）」という異形規則が岡山大学側には見られたが，これを「副詞（1 個以上）-形容詞（1 個以上）」と「形容詞（1 個以上）」に分割する，ということである。

(ii) は，異形規則として，異形用例に厳密に準じて品詞数に上限を加えると，今回準備した異形用例に含まれていなかったものの人間が内省的にありうると判断できる異形を除外してしまうおそれがあったからである。例えば,「運と金の尽きだよ。」（運の尽き：Noun-Particle-Noun 型）という異形用例の挿入パターン（挿入される品詞列）は「名詞（1 個）」であるが,「運と金と時間の尽きだよ。」といったように名詞をさらに並列して挿入できる。このような場合を考慮し，少なくとも同じ品詞の連続は際限なく許容しておくこととした。これによりオーバーマッチングの可能性は増すと考えられるが，イディオム異形検出の再現率を高められる。ただし負例を除外可能な場合は，正例の検出率を下げない限り上限の設定を許容した。

3.2 結果

前節の方針に沿って両作業グループの異形規則を整備した結果，各品詞パターンについて表 4 に示す数の異形規則を得た。整備を経てもこれらの異形規則は異形用例中の正例をすべて検出する。結局，各品詞パターンについて 20-30 の異形規則で十分であることがわかった。また，この異形規則においてオーバーマッチングする負例の数は表 5 の通りである。

表 4: 整備後の異形規則の数

品詞パターン	規則数
Noun-Particle-Verb	25
Noun-Particle-Noun	30
Noun-Particle-Adjective	27
Noun-Particle-Verb-Auxiliary	25
Noun-Particle-Verb-Verb	23

表 5: オーバーマッチングする負例数

品詞パターン	負例数（内訳）
Noun-Particle-Verb	24 (22%)
Noun-Particle-Noun	18 (20%)
Noun-Particle-Adjective	8 (9%)
Noun-Particle-Verb-Auxiliary	26 (21%)
Noun-Particle-Verb-Verb	16 (18%)

4 考察

本研究の異形規則がオーバーマッチングする負例については，大きく 2 種類の傾向が認められた。

1. 挿入された格助詞や接続助詞によってイディオムの意味が失われる
2. 先頭（または末尾）の品詞と意味的に強いつながりを持つ他の語がイディオムの外側から影響する

1 は例えば,「意地が汚い君が悪いに決まっているだろう」（意地が悪い：Noun-Particle-Noun 型）による主述の切り替わりや「敷居が金無垢で高いけどいかにも趣味の悪い代物でした」（敷居が高い：Noun-Particle-Adjective 型）にみられるような補足の挿入である。一方 2 は,「腹には灸を据えかねる」（腹に据えかねる：Noun-Particle-Verb-Verb 型）のように,「灸を据える」という別のイディオムが優先しているものや,「青菜にちょっと塩を振りかけてみましょう」（青菜に塩：Noun-Particle-Noun 型）のように，通常の用法として「塩」と関連が深い「振る」という語が接続してそちらに意味が奪われているものである。

1 については，今回の異形規則作成作業につき，2.2 で述べたとおり各品詞パターン型における Particle (助詞) の前後で挿入可能な品詞パターンを区別しないこととしたため，挿入パターン内の助詞を活用できていない。的確な識別のためには助詞の種類（格助詞など）を含めて異形規則を作成できるとよいだろう。また 2

については本研究のように挿入された品詞のパターンから識別するのは困難であり，文全体の係り受け解析等の手法を利用する必要があるだろう。一方，オーバーマッチングの許容範囲と意義も，人間による翻訳の観点からは検討する余地がある。

5 おわりに

QRidiom 開発当初は日本語非母語話者による日本語イディオム学習の用途を想定していたが [5]，Tübingen University の日本学専攻の学生を対象に試験導入した際，QRidiom を用いた異形規則構築及びそれを通じたイディオムの理解は，日本語非母語話者にとって難易度が高いことがわかった。日本語非母語話者に対して有効なイディオム教育プラットフォームのあり方についてはさらなる研究が必要である。

竹内ら [6] の課題の一つである「複数の作業者が同じデータをもとに作業したときどのようなパターンを定義するのか」について，岡山大学と東京大学で異形規則には形式上の相違（複合度と品詞数上限）はあったものの，可逆的な変換が可能で，基本的にはほぼ同じ規則が作成されていたことがわかった。2.2 で共有した作成手順を守ることでこの結果を得ていることと，3.1 で策定した整備の方針とを考慮すれば，言語学の専門的な知識がなくても母語話者なら同程度の品質で異形規則を作成できることが示唆される。

そして異形規則を複合しないのであれば，イディオムに挿入可能な品詞パターンは異形用例に形態素解析を行うだけで得ることが可能であり，異形規則の作成は自動化できる。したがって，他のイディオムについて異形規則を実装する際には，人間の内省に基づく異形用例の作成だけで十分であると考えられ，比較的小規模な作業者への依頼によって低予算かつ容易に，異形検出できるイディオムの網羅性を拡張できる見通しが立った。

本研究で策定した日本語の異形規則とイディオム検出機能は，今年度内に「みんなの翻訳」及び「みんなの翻訳実習」[2] [9] の翻訳エディタ QRedit 上で利用可能になる予定である。

謝辞

本研究は 2013-2014 年度 JSPS-DAAD 二国間共同研究「日本語を起点言語とする翻訳環境における日本語熟語・慣用句の柔軟なマッチング」(JSPS: 13035821-000302; DAAD: 56455743) の支援を受けている。

[2] https://edu.ecom.trans-aid.jp/

参考文献

[1] Michael Carl and Ecaterina Rascu, 2006, A dictionary lookup strategy for translating discontinuous phrases, *Proceedings of the European Association for Machine Translation*, pp. 49-58.

[2] Gábor Prószéky and Balázs Kis, 2002, Context-sensitive electronic dictionaries, *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pp. 1-5.

[3] Koichi Takeuchi, et al., 2007, Flexible automatic look-up of English idiom entries in dictionaries, *Proceedings of the MT Summit 2007*, pp. 451-458.

[4] Masao Utiyama, et al., 2009, Minna no Hon'yaku: a website for hosting, archiving and promoting translations, *Proceedings of the Translating and the Computer*, pp. 19-20.

[5] Ryoko Adachi, et al., 2013, Development and use of a platform for defining idiom variation rules, *Proceedings of the 5th International Language Learning Conference*, pp. 1-19.

[6] 竹内 孔一 他，2014，簡単なイディオム異形規則の作成: プラットフォームと日本語の異形規則，言語処理学会第 20 回研究大会発表要綱，pp. 488-491.

[7] Rei Miyata, et al., 2014, The use of corpus evidence and human introspection to create idiom variations, *Proceedings of the Second Asia Pacific Corpus Linguistics Conference*, pp. 201-202.

[8] Ulrch Apel, 2006, Neueste Informationen zum elektronischen japanisch-deutschen Wörterbuch WaDokuJT, *Deutschsprachigen Japanologentages, Band III – Sprache, Sprachwissenschaft, Sprachlehrforschung*, pp. 141-159.

[9] Anthony Hartley, et al., 2014, 共同翻訳を考慮した「翻訳教育用みんなの翻訳」システム: みんなの翻訳第 4 報，言語処理学会第 20 回研究大会発表要綱，pp. 254-257.