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

竹内孔一 †, 白石貴大 ♯, Ulrich Apel‡, 宮田玲 ♮, 足立諒子 ♮,
Wolfgang Fanderl‡, 村山遼 ♮, Iris Vogel‡, 影浦峡 ♮

† 岡山大学大学院自然科学研究科
‡ Tübingen Eberhard Karls University
♯ 岡山大学工学部
♮ 東京大学大学院教育学研究科

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、言語学でも重要な研究対象となっている (Benson, 1985; Čermák, 2001; Everaert, 1995; Fraser, 1970; Moon, 1998; Numberg et al., 1994)。翻訳や言語学習においては、比較的熟練した者でもイディオムを苦手とする場合が少なくないことから、イディオムの扱いは重要な課題である。異形の存在もこの困難を増している。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が (Breidt et al., 1996; Breidt and Feldweg, 1997; Carl and Rascu, 2006; Michiels, 2000; Poznanski et al., 1998; Proszeky and Kis, 2002; Takeuchi et al., 2007)、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柔軟な辞書引き環境は現在もあまり多くはない。著者らはオンラインの翻訳システムを開発・運用しており[1]、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組み込まれている (Takeuchi et al., 2007)。翻訳支援環境において、他の言語でも同様の辞書引き機能を実現するために、イディオム異形規則作成プラットフォームを開発し日本語イディオム異形規則を予備的に作成した。

2　基本的な考え方

2.1　異形のタイプ

異形は以下に分類できる (Adachi et al., 2013)：

1. タイプレベルの異形（“to/till the last”）、
2. 受動化や話題化等による変形（“the breeze was shot”, “it is those strings that he pulled”）、
3. 挿入（“go exact halves”）、
4. 要素の置換（“head screwed on wrong”）、
5. 創造的異形（“ball point pen of view”）。

ここでは最も数の多い挿入を対象とする。タイプレベルの異形は辞書見出しで扱われ、受動化や話題化による変形は一般的な文法規則で処理できる。要素の置換を扱うためにはシソーラスのような言語資源が必要であり、創造的異形は今のところ自動的には扱えない。

2.2　異形の範囲

異形の範囲の定義は様々ありうる。例えば、“kick the bucket”を受動化した “the bucket is kicked”はイディオム的な意味を失うので、「言語学」的な観点からは、異形ではないことになる。しかしながら、言語使用の実態としては、“the bucket is kicked”という表現が使われる際に “kick the bucket”というイディオムの存在を前提としていることは少なからずあるので、翻訳や言語学習の場においては、受動形が文字通りの意味で使われているとしても、能動形におけるイディオムの意味を知ることは重要となる。翻訳支援を想定する場合、最終的な意思決定は翻訳者が行うことが前提となるため、オーバーマッチングは、過度なものでなければ情報の欠落と比べて問題とはならない。従って、異形の範囲は広く取ることが有効な方針となる。そのため、イディオムの構成品詞パターンごとに、挿入可能（あるいは禁止）品詞パターンを定義すれば、細かい個別イディオムの性格は考慮しなくても、現実的に有用な異形マッチング規則となる[2]。

2.3　規則の作成手続き

イディオムの異形を含めた自動辞書引きをめぐる問題は、言語処理的な問題でも言語学的な問題でもなく、

[1]「みんなの翻訳」(http://trans-aid.jp)

[2]実際には可能な異形の範囲は同じ品詞パターンのイディオムでも個別に異なる。例えば「糸目を付けない」は「金に糸目をほとんど付けない人だ」と言いうるが、同じ品詞パターンである「跡を断たない」では挿入は難しい。

異形規則作成をめぐる現実的な問題である。この現実的な問題を解決するための一つの方法は、異形マッチングにより利益を得る人が規則を簡単に作れる枠組みを構築することである。翻訳支援の観点からは異形規則は粗くて（粗い方が）よいことを考慮すると、言語学的な訓練を受けていなくても、要求される規則の構築は可能であると考えられるので、この方法は現実的である。そこで我々は、言語学習者や翻訳者が規則を作ることを想定し、イディオムの異形データを参考にしながら簡単に規則を作ることができるプラットフォームを構築した。このプラットフォームには、イディオム構築の対象となる言語のイディオムが組み込まれる。規則の作成手続きは次のようになる。

1. 例となるイディオムと異形データ（オプション）を準備する。
2. イディオムや異形、排除したい異形等を含む例文をシステムに入力し、結果を見る。
3. 結果を参考に、挿入可能（あるいは不可能）な品詞パターンをプルダウンメニューで指定する。挿入品詞パターンは正規表現のクラスを指定できる。
4. 例文をシステムに入力して検証する。

言語学習者や翻訳者は、品詞等の言語的なカテゴリーや規則についても、形式的な規則の記述についても慣れていない場合が多い。また、複数の作業者で調整しながら作業をすることを想定すると好きな時間に適当に作業することができなくなる。このことから、プラットフォームの機能もインタフェースも単純にし、上記の手続きは個々人が勝手に行うことを想定する。規則の統合と整理は、別途、異形を含めた辞書引きに規則を組み込むときに（研究者や利用者代表が）行う。

3　規則構築プラットフォーム

規則構築プラットフォーム（QRidiom）はブラウザでアクセスできる[3]。初期画面には例文を入力し検索する窓がある。現在、日本語辞書としては三省堂グランドコンサイス和英辞典 (三省堂, 2003) のデータとオンライン和独辞書 WaDoku(Apel, 2008) のデータが入っている。システムでは、主要素が 2 個のイディオムに対して挿入パターンが品詞レベルで定義できる。

例えば「足が出る」というイディオムの異形範囲を記述するとする。「足が痛くて血が出る」という、妥当でない異形をシステムで検索した結果を図 1 に示す。システムはイディオム主要素間に最大 8 要素の挿

[3]暫定的な利用アドレスは http://qredit-dev.silklabs.jp/qridiom/Service.action

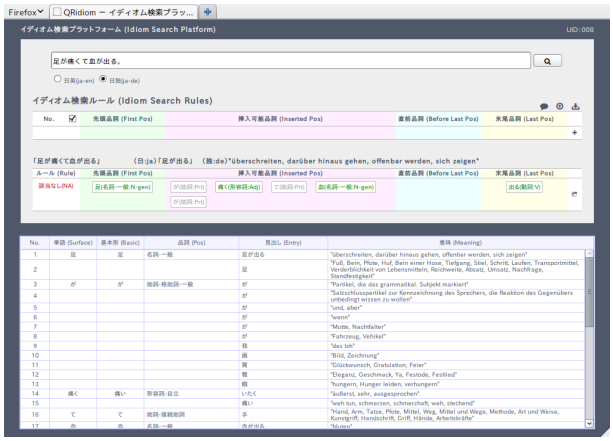

図 1. 異形（負例）の検索結果

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

表 1. WaDoku イディオムの品詞パターン

入があるものまではマッチし結果を出力する。利用者は例文の検索結果から、挿入された品詞列を参照しながら、規則を作成する（図 2)。

定義された規則には番号が付けられ、その後の検索では、異形に対してどの規則が用いられたかが検索結果として表示される（図 3)。

4　日本語異形規則の試験的構築

プラットフォームの実用性を検証するために、予備的に日本語イディオムの挿入異形規則を構築した。

4.1　対象イディオム選択と用例構築

まず WaDoku に登録されイディオムと認定できる日本語イディオム 3916 の品詞パターンを調査した。総異なりパターンは 451 であった。頻度順上位 10 パターンを表 1 に示す。今回は、このうち、1、3、4、7、8 を対象とした。

各パターンから無作為に 20 のイディオムを抽出し、それらを対象にイディオムの異形用例（正例・負例）データを構築した。用例はコーパス（BCCWJ と Web

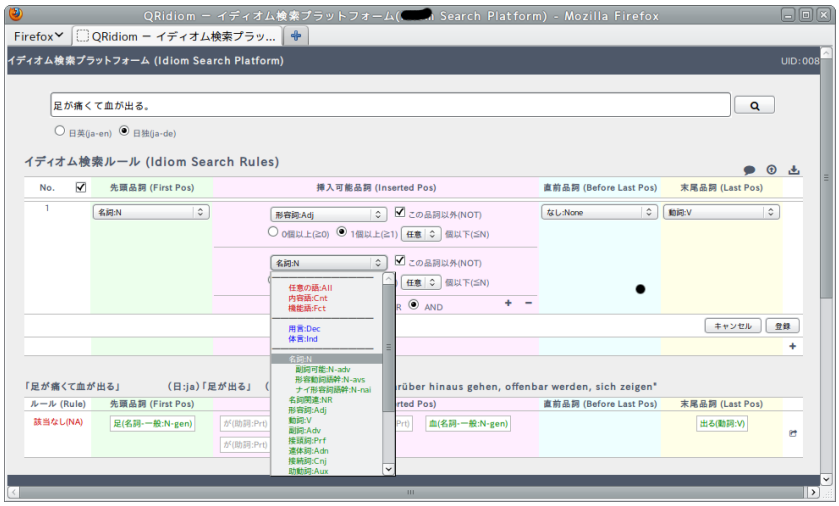

図 2. プルダウンメニューによる規則の定義

図 3. 規則追加後の検索結果

検索）からの収集と人手による作成を試みたが、コーパスからの収集は十分できなかったため、異形用例データは、人手で内省により作成した異形が中心となった。このプロセスの詳細については別途報告している (Miyata et al., 2014)。

4.2 異形規則の構築

これらのイディオム用例を参考に、まず Noun-Particle-Noun のパターンについて、日本語を母語とする理科系の学部学生 1 名が、挿入パターンを記述した。Noun-Particle-Noun を対象にしたのは、用例データを見ると、このパターンにおける挿入がとりわけ変則的だったためである。現在までに作成した挿入パターンを整理すると表 2 のようになる。

予備的な構築から、言語学的な知識がなくても、品詞列を参照しながらパターンを構築することが可能であることは確認された。しかしながら、異形パターンは、可能パターンとしても禁止パターンとしても定義可能であり、複数の作業者が同じデータをもとに作業

品詞パターン	事例
N-P 間の挿入パターン	
Particle-Noun	運 (と金) の尽き
P-N 間の挿入パターン	
Noun-Particle*	命の (真の) 恩人
Noun-Auxiliary	青菜に (累々たる) 塩
Noun(adj)-Particle	運の (完全な) 尽き
Adjective-Auxiliary*	犬猿の (とげとげした) 仲
Adjective-Noun(nai)	宝の (とんでもない) 持ち腐れ
Verb-Particle*	売り言葉に (続き) 買い言葉
Adverb-Particle*	汗の (まさに) 結晶
Adnominal	命の (ちょっとした) 恩人
Prefix	命の (大) 恩人
Particle	論より (も) 証拠

表 2. N-P-N 型慣用の挿入パターン
(*印は 0 回以上の出現を意味する)

したときどのようなパターンを定義するのか、最終的に異形マッチングの規則として組み込むときにどのような整理が必要なのかは、今後の課題として対応が必要である。

5 おわりに

本プラットフォームの構成は意図的に極めて簡単なものとしてあり、また構築されるイディオムの規則も非常に単純なものを想定している。これは、とりわけ翻訳者を（ただし言語学習者も同時に）想定した言語処理による支援として我々が想定している枠組みの中で、実際に強く必要とされながら欠けているものが比較的単純であること、単純であるにもかかわらず広く実現していない状況を変えるために必要なのは、ニーズのある層（言語実務家）が簡単に貢献できる単純な機構であると考えたことによる。

現在、試験的に構築した異形規則だけでなく、同じデータに基づいて日本語を母語としない言語学習者に異形規則の構築を依頼中である。翻訳においては起点言語が母語でない場合が標準であるため、非母語話者によるプラットフォームを用いた規則作成が十分可能であることがわかれば、本プラットフォームはより現実的に有効であることになる。日本語の規則は、統合し整理した上で、「みんなの翻訳」の辞書引きメカニズムに組み込んでいく予定である。

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。株式会社三省堂には高品質の辞書『グランドコンサイス和英辞典』データの利用許可をいただいた。

参考文献

Adachi R. et al. 2013. Development and use of a platform for defining idiom variation rules. *The 5th International Language Learning Conference*, 1–19.

Apel U. Neueste Informationen zum elektronischen japanisch-deutschen Wörterbuch WaDokuJT. Genenz, K. and Unkel, M. eds. *Deutschsprachigen Japanologentages, Band III —— Sprache, Sprachwissenschaft, Sprachlehrforschung*. Bier'sche Verlagsanstalt, Bonn, 141–159.

Benson M. 1985. Collocations and idioms. Ilson, R. ed. *Dictionaries, Lexicography and Language Learning*. Pergamon, Oxford, 61–68

Breidt E., Segond F. and Valetto G. 1996. Formal description of multi-word lexemes with the finite-state formalism IDAREX. *COLING 1996*, 1036–1040.

Breidt E. and Feldweg H. 1997. Accessing foreign languages with COMPASS. *Machine Translation*, 12:153–174.

Carl M. and Rascu E. 2006. A dictionary lookup strategy for translating discontinuous phrases. *EAMT 2006*.

Čermàk F. 2001. Substance of idioms: perennial problems, lack of data, or theory? *International Journal of Lexicography*, 14(1):1–20.

Everaert, M. et al. eds. 1995. *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum, Hillsdale.

Fraser B. 1970. Idioms within a transformational grammar. *Foundations of Language*, 6:22–42.

Jacquemin C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, Cambridge, Mass.

Kageura K. and Toyoshima M. 2006. Analysis of idiom variations in English for the enhanced automatic look-up of idiom entries in dictionaries. *Euralex 2006*, 989–995.

Michiels A. 2000. New developments in the DEFI matcher. *International Journal of Lexicography*, 13(3):151–167.

Miyata R. et al. 2014. Corpus evidence and human introspection for idiom variations. *The 2nd Asia Pacific Corpus Linguistic Conference*.

Moon R. 1998. *Fixed expressions and idioms in English*. Clarendon Press, Oxford.

Numberg G. et al. 1994. Idioms. *Language*, 70(3):491–538.

Poznanski V. et al. 1998. Practical grossing by prioritized tiling. *COLING-ACL 1998*, 1060–1066.

Proszeky G. and Kis B. 2002. Context-sensitive electronic dictionaries. *COLING 2002*, 1–5.

三省堂. 2003. グランドコンサイス和英辞典, 三省堂.

Takeuchi K. et al. 2007. Flexible automatic look-up of English idiom entries in dictionaries. *MT Summit 2007*, 451–458.