英語イディオムの異形を整理する

金平昂 † 豊島実和 ‡ 竹内孔一 † 影浦峡 ♯

† 岡山大学大学院自然科学研究科
‡ 東京大学大学院総合文化研究科
♯ 東京大学大学院教育学研究科

1 はじめに

翻訳者は，既存の高品質辞書に掲載されているイディオムにほぼ満足しているが，その検索機能には満足していない [1]．紙の辞書を用いる場合でも，核となる単語を推測して辞書を引かなくてはならず，知らないイディオムを見つけることは大変困難な作業である．電子辞書／オンライン辞書の中には構成単語の AND 検索ができるものもある．しかし，ある範囲にイディオムが存在すると予測されながら，正確にどこかわからないという，頻発する状況で AND 検索は不十分であり，電子テキストに対する自動的な辞書引きとしても不十分である．

イディオム処理は機械翻訳システムでも不十分である．たとえば，(1) He said that with his tongue in his cheek. と (2) He said that with his big fat tongue in his big fat cheek. の文中には “with one’s tongue in one’s cheek”（意味：あざけって）というイディオムが存在する．しかし，(2) を正しく処理できるシステムは，我々が調べた範囲では存在しなかった．

本研究は，テキスト中に現れる多様な異形を含むイディオムを，辞書に登録された標準形と自動的にマッチさせる手法を開発するために，イディオムの異形を整理し分析する．言語としては英語を扱う．計算言語学分野ではイディオムや連語の抽出研究はあるが [6, 8, 9]，見出しとのマッチングの研究はない．また，イディオムの異形について言語学的研究や記述は見られるが [3, 4, 5, 7]，自動処理に十分な定式化はなされていない．

本研究の最終的な目的は，イディオムの自動マッチングである．したがって，既存のリソースや手法を考慮し，解析や記述のレベルについて，次のような基準を設けた：

- POS タガーや形態素解析器は使うが，パーザーは使わない．パーザーは，今のところ，我々の目的には十分でないと判断したためである．つまり，品詞情報と有限オートマトン以上の計算クラスは想定しない．
- シソーラスは利用できるが，詳細な概念分類や談話情報をもつ語彙資源は利用できないと考える．

なお，問題がマッチングであるため，異形の過剰生成は一定程度まで許容できるし，一定程度ならば，翻訳者にとって積極的に有用でもある．

2 異形データの準備

イディオムの異形を集める効果的手法はないため，人手で異形データを作成した．広く使われている英日対訳イディオム辞書 [2] からイディオムを抽出し，３人の英語母語話者（うち２人は編集の専門家）に，異形の作成を依頼した．表 1 に，作成したデータの基本的な量を示す（h, j, s は作業者のラベル）．

表 1: イディオムと異形の基本量

作業者	(a) イディオム	(b) 異形	(b)/(a)
h	475	469	0.99
j	661	890	1.35
s	777	822	1.06
合計	1913	2181	1.14

このデータには負例がなく，可能な異形の限界もわからないため，言語的性質を考慮した分析により補完する必要があるが，異形パターンを見る出発点としては十分有用である．ここで考慮できなかったものは，イディオム検索を試作して，使いながら補強していく予定である．

3 異形の分析結果

3.1 イディオムのパターン

イディオム自体は，文法的な性格により，11 の基本タイプに分類された（表 2）．イディオムは，定義上，単純な文法パターンによって類型化できるものではないが，検索メカニズムを定義するためにイディオムの異形を記述するという観点からは有用である．検索メ

カニズムを実装するときには，品詞情報や語彙情報しか使えないからである．ただし，これらの類型は，異形のパターンを整理した後，さらに細かくパターン化する（3.3 参照）．

表 2: イディオムの基本タイプ

タイプ	例	数量
名詞句 (前置)	poker face	215
名詞句 (後置)	babe in arms	79
名詞句 (混合)	a rotten apple in the barrel	55
動詞句 (一般)	take the plunge	905
動詞句 (be)	be all balled up	64
形容詞句	straight as an arrow	112
副詞句	all long	26
前置詞句	with open arms	201
非従属節	it’s your baby	195
従属節	if you prefer	40
その他	popsicle	21

表 3: イディオムの異形の分類

タイプ（タグ）	例	数量
置換 (x)	the boiling point → the burning point	759
挿入 (y)	take off → take right off	1203
削除 (s)	not get to first base → got to first base	3
置換＋置換 (xx)（相関関係有）	more alive than dead → more dead than alive	39
挿入＋挿入 (yy)（相関関係有）	can swing it → can swing it no problem	101
置換＋挿入 (xy)	weak as a baby → strong as a baby ox	91
削除＋置換 (sx)	go back to the basics → plunge into the basics	20
削除＋挿入 (sy)	people will talk → people happily talk	8
その他 (z)	take it from me → rely on me	95

3.2 異形のパターン

異形は，大きく，(1) １つ以上の構成要素の系列的な置換，(2) 統語的関係をもつ語句の挿入，(3) 削除，の三種類に分類できた．主要なパターンは前２者である．加えて，(4) これらの異形が依存して起こる場合，(5) 少数の複雑な異形，も認められた．異形の基本分類を表 3 に示す．

置換と挿入の区別が微妙な場合もある．例えば，“carry one’s point” の異形である “cash and carry one’s point” は，“carry” が “cash and carry” に置換されたとも，“carry” の前に “cash and” が挿入されたとも解釈できる．このような場合，言語学的な理由付けにこだわらず,「はじめに」で述べた基準により判断した．例えば，この場合，“cash and carry” は語彙化されていないため，“carry” と “cash and carry” の関係を示すシソーラスはありそうにない．したがって，この異形は，統語的関係をもつ語句の挿入とする．

置換　置換 (x) は，置換する語とされる語の関係のタイプおよび置換される語の品詞によって細分類される．表 4 は，異形データにおける置換の数を品詞と意味別に示したものである（n：名詞，v：動詞，ad：形容詞，av：副詞，p：前置詞，det：冠詞，cj：接続詞，aux：助動詞，dg：冠詞所有格交換）．データ中の個別の異形は複雑で，１つ以上の異形パターンが組み合わさってできた異形が多数存在するため，異形パターンの数は，表 1 の異形数よりも多い．

反対の概念，同義・類義，同列・同等，上下関係の置換の割合は，名詞置換の約 75 % (245/327)，動詞置換の約 69 % (133/192)，形容詞置換の約 80 % (113/141)，副詞置換の約 65 % (26/40) である．このことは，高品質のシソーラスを用いることで，主要なタイプの置換による異形を扱うことが可能であることを示している．「その他」には，“back and forth”→“buck and forth” のように，音の類似などを利用する，創造的で柔軟な置換が含まれる．人間にとってこれらの関係を理解するのは簡単であるが，機械で体系的に扱うことは難しい．

表 4: 置換の内訳

	n	v	ad	av	p	det	cj	aux	dg
反対の概念	20	16	22	12	8	-	-	-	-
同義・類義	133	79	61	12	6	-	-	-	-
同列・同等	88	38	30	2	-	-	-	-	-
付加置換	4	0	13	0	-	-	-	-	-
idiom 内の他語との関連語	2	3	3	0	-	-	-	-	-
別の文脈での関連語	3	2	1	1	-	-	-	-	-
大きな単位での入れ替え	23	5	1	3	-	-	-	-	-
上下関係	4	-	-	-	-	-	-	-	-
単数形／複数形での入れ替え	3	-	-	-	-	-	-	-	-
その他	47	49	10	10	21	8	5	4	7
合計	327	192	141	40	35	8	5	4	7

挿入　挿入 (y) は，挿入語の位置及び，品詞と形式的／意味的役割によって下位分類される．表 5 に下位分類ごとの挿入の数を示す．表中，“pp” は前置詞句，“cl” は節，“p” は前置詞，他は表 4 と同様の意味を表す．位置情報は省略した．

表 5: 挿入の内訳

	n	v	ad	av	pp	cl	det	p	aux
否定・反対	0	0	6	13	-	-	-	-	-
強め	2	0	163	265	-	-	-	-	-
並列・追加 (and/or 有)	9	3	7	6	-	-	-	-	-
並列・追加 (and/or 無)	0	1	15	0	-	-	-	-	-
修飾一般	44	-	402	167	-	-	-	-	-
所有格	8	-	-	-	-	-	-	-	-
その他	20	2	16	21	25	2	2	0	4
合計	83	6	609	472	25	2	2	0	4

挿入の操作は，基本的に通常の文法規則に従っており，主要な 3 つのタイプを認めることができる．(a) “very”, “fucking”, “really” のような，少数の顕著な語彙によりどこにでも起こりうる挿入，(b) 通常の挿入，(c) “blind as a bat” → “blind as a baseball bat” のような音の類似等による挿入である．挿入が標準的な文法規則に従っているということは，イディオムの自動マッチングにとって，第一段階としては有益である．これによって，挿入パターンに体系的な文法上の制限を組み込むことで，どの程度まで探索空間を狭くできるかが，次のステップで観察すべきポイントとなる．

表 6: 削除およびその他の異形の内訳

パターン	s-n	s-av	s-p	s-det	s-gen
数	0	2	0	1	0

xx	xy	yy	sx	sy	z
39	91	101	20	8	95

削除／その他　削除 (s) は，削除される語の品詞によって分類した．表 6 では削除の数に加えて，依存関係にある異形の組合せによる異形についても示す．表から，異形を組合せた異形の数がかなり多いことがわかる．これらの多くは，依存関係を無視し，基本的な異形の組み合わせとして扱うことができると考えられるが，さらなる分析が必要である．

3.3　挿入の異形パターンの定式化

これまで述べてきた主な異形のうち，置換はシソーラスを準備することでそのままある程度対処できる．一方，挿入は，パターンの記述をさらに定式化する必要がある．そのために，イディオムを構成する単語の品詞の並びによってさらに識別し，分類した各イディオムのパターンを挿入語の品詞の種類で分類した．例えば，“take the plunge” というイディオムの異形 “take the wild plunge” について分類する場合，“動詞 冠詞 名詞” というパターンに分類され，挿入語は “形容詞” である．このパターンのイディオムの異形として，挿入語に “名詞” もある場合は，挿入語の品詞の種類は “形容詞” と “名詞” の二つである．このパターンを用いて，異形の検索を精度よく行うことが可能となる．分類結果を表 7 に示す．6 語以上から構成されるイディオムは，1〜5 語から構成されるイディオムのパターンを組合わせとみなすことができる．従って，1〜5 語から構成されるイディオムに対して分類を行なう．6 種類以上の挿入が起こるパターンはなかった．

表 7: 挿入語の品詞とイディオムのパターン数

挿入語の品詞の種類数	1	2	3	4	5
イディオムのパターン数	82	23	8	1	1

表 7 より，1 種類の品詞の単語の挿入を異形としてもつイディオムのパターンが多く（80 %強）見られる．イディオムのパターン毎に，異形のパターンが制限されることが多いのである．よって，挿入の異形はイディオムを構成する語の品詞と挿入語の品詞の関係を分類することで，自動マッチングに利用できる．

分類の方法としては 2 つ考えられる．挿入語と，(1) イディオム全体のパターン，(2) イディオムの挿入箇所の前後のパターン，である．例えば，“take the plunge” というイディオムの異形 “take the wild plunge” について，(1) では，“動詞 冠詞 ＊ 名詞”（＊の位置に形容詞が挿入）とする．(2) では，“冠詞 ＊ 名詞”（＊の位置に形容詞が挿入）とする．ここでは，拡張性と一般性を考慮し，(2) の方法をとった．分類結果を表 8 に示す．なお，分類の際，名詞を以下の 3 つに分類した．(1) 所有代名詞（my, his 等の one’s 型），(2) 人称代名詞（I, he 等の someone 型），(3) その他の名詞，である．

表 7 で，挿入の異形をもつイディオムは 115 パターンに分類された．表 8 より，挿入場所の前後の 2 単語の品詞と挿入語の品詞との関係で分類すると，25 パターンに分類できた．この分類は，後に自動処理を行なう上で適した体系である．ただし，挿入句の場合は句のパターンについて考える必要があるなど，完全ではない．

表 8: 前後の品詞と挿入語の品詞との関係

挿入箇所の前後の 2 単語の品詞（前，後）	挿入語の品詞
（名詞，名詞）	名詞，形容詞
（名詞，前置詞）	形容詞，名詞，副詞
（名詞，形容詞）	副詞，形容詞
（名詞，動詞）	副詞，助動詞
（形容詞，名詞）	名詞，形容詞，副詞
（形容詞，前置詞）	名詞
（副詞，形容詞）	副詞
（副詞，副詞）	副詞
（副詞，前置詞）	副詞
（副詞，動詞）	副詞
（接続詞，動詞）	副詞
（接続詞，前置詞）	副詞
（接続詞，名詞）	形容詞
（動詞，副詞）	副詞，形容詞
（動詞，名詞）	名詞，形容詞
（動詞，形容詞）	副詞，形容詞
（動詞，前置詞）	副詞，形容詞
（前置詞，名詞）	名詞，形容詞
（前置詞，形容詞）	副詞，形容詞，名詞
（冠詞，形容詞）	形容詞，副詞，名詞
（冠詞，名詞）	名詞，形容詞，副詞
（所有代名詞，名詞）	形容詞，名詞
（人称代名詞，名詞）	形容詞
（人称代名詞，副詞）	副詞
（人称代名詞，前置詞）	副詞，形容詞

3.4 考察

結局，異形の大部分が系列的な置換（置換）と統語的関係をもつ語句の追加（挿入）でカバーされていることがわかった．どちらの場合も，自動処理の観点から明確に定式化可能な範囲がかなりあり，標準的なシソーラスの意味や基本的な文法パターンによって，主要なタイプの異形を適切に扱えると期待される．さらに解析や自動検索システムの性能の実験を行なう必要はあるが，ここでの解析によるとイディオムの自動検索が発展する現実的な可能性があることがわかる．

4 まとめ

本稿では，英語イディオムの異形パターンを分析整理した．それらの多くは次の 2 つのレベルで記述できる．
(1) 統語的な異形が関係している時は，文法パターン．
(2) 系列的な異形が関係している時は，シソーラス関係．

一般にイディオムが文脈や談話内で一つに統合されたものであると考えると，この研究で我々が採用した記述レベルは，必ずしもイディオムの異形の本質的な範囲を十分説明的に特徴づけるものではない．しかしながら，我々が採用した記述のレベルは，テキスト中に現れるイディオムと辞書に登録されているイディオムを自動的にマッチングするという，計算機で扱える手法を実現するためには重要かつ有用である．現在，我々は，ここで報告したパターンを考慮し，イディオムの自動検索マッチング・システムを構築中である．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金基盤 (A)「翻訳者を支援するオンライン多言語レファレンス・ツールの構築」（課題番号 17200018）および NICT 受託研究「翻訳者を支援する多言語レファレンス・ツールの研究」の支援を得て行われた．

参考文献

[1] 専門翻訳者・ボランティア翻訳者 6 名への最終著者による聞き取り．

[2] ジャン・マケーレブ，岩垣守彦. 2003.『英和イディオム完全対訳辞典』東京：朝日出版者．

[3] Benson, M. 1985. “Collocations and idioms,” In Ilson, R. ed. *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon Press. p. 61–68.

[4] Biber, D. et. al. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

[5] Moon, R. 1998. *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

[6] Piao, S. S. L. et. al. 2003. “Extracting multiword expressions with a semantic tagger,” *ACL2003 Workshop on Multiword Expressions*, p. 48–53.

[7] Quirk, R. et. al. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

[8] Smadja, F. 1993. “Retrieving collocations from text: Xtract,” *Computational Linguistics*, 19(1), p. 143–177.

[9] Widdows, D. and Dorow, B. 2005. “Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns,” *ACL2005 SIGLEX Workshop on Deep Lexical Acquisitions*, p. 48–56.