置換・挿入を考慮した異形イディオム検索システムの構築

竹内 孔一，金平昂，平尾 一樹，
岡山大学大学院自然科学研究科
koichi@it.okayama-u.ac.jp

阿辺川武，影浦峡
東京大学大学院教育学研究科
{abekawa, kyo}@p.u-tokyo.ac.jp

1 はじめに

人間の翻訳作業において文中に現れるイディオムに気が付くことは重要であるが，イディオムの数は多く（三省堂『グランドコンサイス英和辞典』[10] では約 2 万語が登録されている），また翻訳者が相対的に苦手とするところであるため，電子テキスト中でイディオムを自動検索するシステムが有効である．ところが，一部の例外を除いて [3]，これまで異形を含めた柔軟なイディオム自動検索システムの研究はあまり行われてこなかった．そのため，人手による検索でも自動検索でもイディオム検索はボトルネックとなっている．本研究では翻訳者支援ツール構築の一環として，イディオム自動検索システムの構築を目指しており，本稿では，すでに提案した挿入による異形を扱う手法 [5] に加え，置換による異形 [4] を扱う手法を提案する．

以下ではまずイディオムの異形に対する我々の考え方を簡単に整理する．次に，置換による異形の種類を明らかにしたのち，本稿で扱う置換のタイプについて考察する．それを受けて，置換に対する処理の枠組みを設計し，全体のシステムおよび検索実験結果を示す．

2 イディオムの異形

2.1 異形のクラス

イディオムの異形については様々な言語学的考察があり [7, 8, 9]，連語を含むより広い複数単語からなる表現（MWE）についての研究も多い [2]．先行研究および翻訳者が直面するイディオムの異形を検討すると，概念的には以下のような異形の区分が成り立つ．

(1) 主題化や受動化等，外部的文法操作による異形（“pull strings” → “these are the strings he’d happily pull”など）．
(2) イディオムの構成要素に直接関わる異形（“go halves” → “go exact halves”など）．
(3) 言葉遊びなどの生産的な異形（“screwed on right” → “screwed on wrong”など）．

しかしながら，(1) を除いては，異形の規則をどう定式化すべきか明確にはなっていない（言語学的な異形の研究はあるが，「“shoot the breeze”や “kick the bucket” は受動化されない」，「“go halves”は挿入を許さない」など，現実に翻訳者が直面するテクストの異形の実態からは乖離している）．そこで我々は，主に (2) と (3) のタイプを想定し，英語母語話者 3 人に依頼して 2181 個のイディオム・サンプルについて異形データを作成し，分析した [4, 6]．その結果，イディオム構成要素の一つ以上を関連する要素で置換する異形が，挿入とともに異形の重要なクラスを構成することがわかった[1]．

2.2 置換による異形のタイプ

置換による異形のタイプは，置換前の構成要素と置換後の構成要素の関係から類型化できる．品詞別に類型を整理した結果を表 1 に示す．

表 1: 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

ただし，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．シソーラスが表示する基本的な関係である反意語，同義・類義語，同列語の置換で名詞置換の約 74 % (241/327)，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)

[1] 異形が「言葉遊び」によるものかどうかといった原因の分類は計算機処理に活用できないため，以下の議論では (2) と (3) の区別は解消して論ずる．

を占めることがわかる．ここから，高品質のシソーラスを用いれば，置換による異形のかなりを扱えることが示唆される[2]．以下では，これらの置換を対象とする．

3 置換イディオム処理の枠組み

置換による異形処理の基本は単純である．シソーラスを導入し，反意語，同義・類義語，同列語を展開して辞書とのマッチングを行う．本研究ではシソーラスとして WordNet を導入し，『グランドコンサイス英和辞典』のイディオムとマッチングを行う．なお，我々は，翻訳上の意志決定は翻訳者が行うことを前提としているため，イディオム検索システムの基本要求仕様は「一つの正解」を出すことではなく，漏れをなくし，多少の幅を持って翻訳者に有用な情報を提供することにある．したがって，過度に邪魔でない限り精度は多少低く過マッチングさせてよい．

WordNet の利用　まず WordNet が上記の置換に対してどれぐらいカバーできているかを評価した．置換による異形データから，反意・同義・同列語による置換 521 個について，WordNet の反意語，同義語，同列語の展開を行うことで，もとのイディオムが推定できるかどうかについて調べてたところ約 50%(263/521) の異形がカバーできた．逆に言うと，WordNet だけでは半分強の置換は扱えないが，これは複数のシソーラスを将来利用することでカバーする範囲を広げることで将来的に対処したい．また，WordNet には前置詞に関する反意語や類語関係は記述されていない．そこで，前置詞に関しては表 2 に示すような展開表を作成し，システムに組み込むこととした．

英和辞典の情報　『グランドコンサイス英和辞典』だけでなく通常の辞典には，イディオムの各構成要素について品詞情報は付与されていない．したがって，WordNet で展開してマッチングを行う際，品詞の異なるイディオムが検索される可能性がある．例えば “have one’s eye on” の “have” は動詞であるが辞書にはその記述が無いためシステムとしては入力文の単語に対して “have” の反意語，同義語，同列語として WordNet で検索されれば，品詞に関係なくマッチさせてしまう．イディオムに POS タギングを適用することも考えられるが，イディオム全体の文法単位があらかじめわかっていないとエラーが多い．そこで今回は，シソーラス展開に基づくマッチングを，品詞の合致は考慮せずに行う．

[2]表 1 中の「その他」の異形とは音の似ている単語への置換など連想力を働かせたものであり計算機での展開では容易ではない．

表 2: 前置詞の置換関係データ

前置詞	反意・類義語
at	to
above	up
atop	on
beyond	past
down	up
for	to
in	into, out
into	in
off	on
on	upon, atop, off, under
out	in
over	under
past	beyond
to	into, at, for
under	over, on
up	above, down
upon	on
with	without
without	with

展開に対する制約　以上のような条件のもと，WordNet による展開で不要なイディオム候補まで多数検索されすぎることを防ぐために，展開に以下の制約を設ける[3]．

- イディオムを構成する語が全て置換されることはない．
  例: “have on” が “take in” にはならない．
- 3 語以上からなるイディオムでは，動詞，名詞，前置詞，接続詞，副詞のいずれかが置換されずに残る．
  例: “have a seat” の異形として “take a seat” は可能であるが “take a stand” はあり得ない．
- イディオム中で be 動詞は置換されない．

4 置換と挿入を扱う検索システム

前節で概略を述べた，置換による異形を考慮したイディオム検索のメカニズムを，既に開発している挿入を考慮したイディオム検索システム [5] と統合する．置換処理では単語の展開を数多く行うため，処理に負荷がかかる．したがって，まず置換処理を行った後に挿入処理を行う．これにより，挿入も置換も同時に起こる異形イディオム [6] を扱うことも可能になる．挿入処理のアルゴリズムは文献 [5] に譲るとして，ここではシステムの全体像，共通の正規化処理，複数の候補をランキングする手法について説明する．

[3]すでに述べたように，イディオム異形の制約については言語学の研究で考えられているが [2, 7, 8, 9]，現実の異形はこれらの研究で述べられている制約を大幅に逸脱するため，「言語学的」制約は正しい候補を落としてしまう．

4.1 検索アルゴリズム

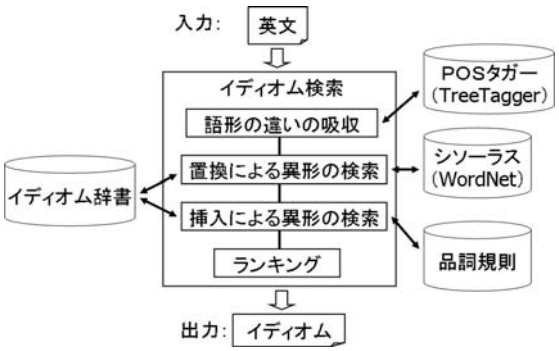

図 1: イディオム検索システム

検索システムの全体像を図 1 に示す．まず，TreeTagger を利用して活用形の正規化を行い，品詞を付与する．この情報をもとに WordNet で語を展開し，WordNet の展開されている語と正規化した語の中から辞書におけるイディオム・エントリーとのマッチングを行い置換による異形に対するイディオムを検索する．このとき，同時に挿入による異形の処理も行い，イディオムの構成要素が離れて現れる場合も，置換・挿入のあるイディオムとして候補に選んでおく．その後，品詞ベースで記述された挿入の制約規則 [5] を適用して，元のイディオムからの異形としてありえる挿入パターンに適合した異形イディオムのみを候補として残す（フィルタリング）．この段階で残された異形イディオム候補についてランキング処理を行い，その順序にあわせて提示する．図 2 に，展開からマッチングまでの様子を示す．

なお，現在速度向上の最適化を行っていないため，1 単文を処理するのに数秒かかる．現在，その高速化を検討中である．

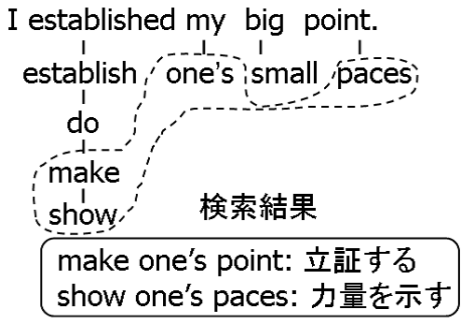

図 2: イディオム検索の様子

4.2 ランキング

複数の置換・挿入による異形イディオム候補が抽出されるため，何がしかの尺度でもっともらしい順序に並べ替える必要がある．まずどの範囲の候補をランキング対象にするかであるが，異形イディオム候補中で単語の位置が重なる部分を持つ候補集合をすべて対象とする．例えば図 3 のようなイディオム候補が検索された場合，“make one’s point”，“show one’s paces” がランキングの候補である．次にランキングのための評価式の決定であるが，基本的なアイデアとしてはテキスト中においてイディオムは異形よりも基本形で現れることの方が多いという傾向を重視する．つまり，置換語数が少ない，挿入語数が少ない，構成語数が多いイディオムが最もイディオムらしいという考えに基づいてランキングを行う．また挿入による異形の方が，置換による異形よりもテキスト中によく現れるという性質に従って，置換語数，挿入語数，構成語数に関しては，置換語数 ＞ 挿入語数 ＞ 構成語数 の順に重要であると仮定する．この仮定に基づくランキングスコア（H）を下記のように定義して各イディオム候補に対して評価を行う．

$$H = ((C - n1) * 100) + ((C - n2) * 10) + n3$$

ここで，n1：置換語数，n2：挿入語数，n3：構成語数，C：定数，である．

上記の式で計算したスコアを利用して，入力文中の同じ語を含むイディオムの中で最もスコアの高いイディオムを出力する．ランキング例を図 3 に示す（C=10）．イディオム候補にスコアを付与すると，“take the plunge = 1093”，“make one’s point = 1003”，“show one’s paces = 903” となる．これらをランキングすると，入力文中の各位置で最もスコアが高いイディオム，“take the plunge”，“make one’s point” が出力される．

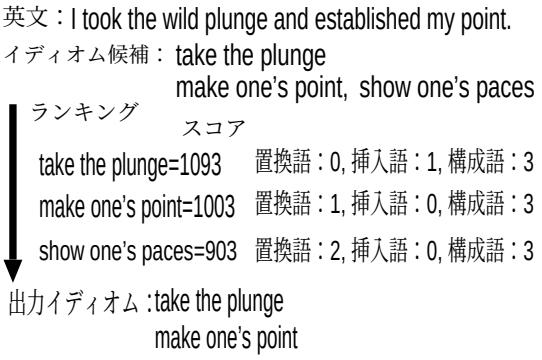

図 3: ランキングの例

5 検索実験と考察

構築した異形検索システムがどれだけ取りこぼしが少なくかつ候補を絞りこんで提示できるかを測るために，挿入による異形イディオムを含む文 (100 件)，置換による異形を含む文 (100 件) の 200 件について検索を行いその適合率と再現率を求めた (表 3).

表 3: 挿入・置換による検索実験

	適合率	再現率
異形を考慮せず	0.898	0.509
	(220/245)	(220/432)
挿入のみ考慮	0.803	0.792
	(342/426)	(342/432)
挿入と置換の	0.374	0.931
異形を考慮	(402/1075)	(402/432)

表 3 から挿入と置換を考慮した場合，再現率は 93.1% に達し，取りこぼしが非常に少ないことがわかる．一方で適合率の方は 37.4%と低くなるが，これは実用上問題ないと考えている．なぜならば，これは平均解候補は 3 つ程度であることを示しており提示する際に問題にならない数であるからである．

しかしながら表 3 のデータはイディオム検索システム構築において参考としたデータであり，未知データではないため精度が有利に働いている点がある．今回の異形データはその元のデータがグランドコンサイス辞書に記載されているもののみを選んである．上記の再現率が示すようにカバー率がかなり高い手法であることがわかるが，一方で，グランドコンサイス辞書のイディオムのカバー率というのは今回使用しているデータ全体に対して 34%程度であった．このカバー率を上げるには複数の辞書を利用する等の対処が必要であろう．今回はランキングについては精度を評価していないが複数辞書を利用して適合率がより下がった場合には重要度が増すと考えられる．

6 まとめ

翻訳支援ツールとしてテキスト文に現れるイディオムについて，語が挿入されたり，一部の単語が置換された場合についても，イディオム辞書から検索し提示出来るシステムを構築した．置換のための同義・同列語辞書として WordNet を利用し，過剰に生成される候補について品詞や他の候補との競合関係を考慮することで高い再現率を提示可能かつ候補数を絞り込めることを実験で示した．今後統語的な異形に対しても扱う予定である．

なお、最終的には，このシステムは，翻訳者のための統合的支援環境として開発が進められている QRedit[1] の一機能として組み込まれる．そこでは，翻訳元文書（英文）と翻訳文（日本語）を同時に表示し，マウスオーバーで語の意味やここで開発した異形イディオム検索の結果を示す．つまり翻訳支援ツールの一部として機能することになる．

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」(研究代表者:影浦峡)（課題番号 17200018）の支援を得て行われた．また,『グランドコンサイス英和辞典』のデータ利用を許していただいた（株）三省堂に感謝する．

参考文献

[1] 阿辺川武，影浦峡: QRedit: 英日ボランティア翻訳者向け統合エディタ，言語処理学会第 13 回年次大会，(2007: 発表予定).

[2] Baldwin, T.: Multiword Expressions, *Advanced course at the Australasian Language Technology Summer School* (2004).

[3] Carl, M. and Rascu, E. “A Dictionary Lookup Strategy for Translating Discontinuous Phrases” *EAMT-2006* (2006).

[4] Kageura, K. and Toyoshima, M.: Analysis of Idiom Variations in English for the Enhanced Automatic Look-up of Idiom Entries in Dictionaries, *Proceedings of the 12th Euralex International Congress*, pp. 989–995 (2006).

[5] 金平昂, 平尾一樹, 竹内孔一, 影浦峡: イディオムの異形規則を利用したイディオム検索システムの構築, 言語処理学会第 12 回年次大会発表論文集, pp. 711–714 (2006).

[6] 金平昂, 豊島実和, 竹内孔一, 影浦峡: 英語イディオムの異形を整理する, 言語処理学会第 12 回年次大会発表論文集, pp. 1019–1022 (2006).

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

[8] Nicolas, T. “Semantics of idiom modification,” In Everaert, et. al. eds. *Idioms: Structural and Psychological Perspectives*. Hillsdale: Lawrence Erlbaum Associates., 1995. p. 233-252 (1995).

[9] Numberg, G., Sag, I. and Wasow, Th. “Idioms,” *Language* 70(3), p. 491-538 (1994).

[10] 三省堂編集所 『グランドコンサイス英和辞典』(2004).