誤記修正版 (2013-03-19)

複合動詞・構成動詞間の意味的な共通性を計る指標の評価

山口 昌也
国立国語研究所

1. はじめに

本稿では，日本語の複合動詞とその構成動詞の意味的な共通性を計るための指標を提案し，評価する。なお，本稿で扱う複合動詞は,「生み出す」のような「動詞（連用形）＋動詞」タイプの複合動詞である。

複合動詞の統語的・意味的な構造を分析する上で，複合動詞と構成動詞との間の意味的な共通性は，重要な要素である。例えば，山本 [1] は，複合動詞辞書の試案において，複合動詞の格支配構造を記述するために，複合動詞と構成動詞の格支配構造間の関係を記述している。また，由本 [3] は，複合動詞の概念語彙構造 (Lexical Conceptual Structure，以後，LCS と表記) を，構成動詞の LCS の合成に基づいて記述している。これらの研究において，格支配構造の関係記述や LCS の合成は，複合動詞と構成動詞の意味的な共通性が前提となっている。

その一方で，意味的な共通性は，内省により判断されているため，構造分析の結果を客観的に評価することが困難になる。また，共通性の「度合い」が規定されていないので，構成動詞が複合動詞の要素となったとき，構成動詞にどのような意味的変化が起こるのか，といったことを明らかにすることができない。

そこで，本稿では，複合動詞と構成動詞との間の意味的な共通性を数値的に表現する指標を提案し，評価する。本指標は，複合動詞と構成動詞の周辺文脈の共通性を，大量の用例に基づいて計算する。指標の評価は，人手による意味的関係性の判断結果と一致する度合いを，既存の語間の類似度計算法 [5] と比較することによって行う。さらに，本指標を用いて，構成動詞が複合動詞の要素となったときの意味的変化を分析する。

2. 分析対象とする複合動詞

複合動詞と構成動詞間の意味的共通性を計る指標について述べる前に，本論文で扱う複合動詞の種類を定めておく。

本論文では,「動詞（連用形）＋動詞」タイプの複合動詞のうち，影山 [2] が提案する「語彙的複合動詞」を分析の対象とする。

[2] では，複合動詞が生成文法中のどの部門で記述されるかに基づいて，語彙的複合動詞，統語的複合動詞の 2 種類に分類している。語彙的複合動詞は，語彙部門で派生され，レキシコンに記述される。語彙的複合動詞の例として,「使い回す」「出し抜く」「飛び散る」「踏ん張る」を挙げる。これらは，前項・後項動詞の合成時に，意味的な制限が加わったり，まったく別の意味を持つようになる（例：「出し抜く」）。

一方，統語的複合動詞は統語部門で形成される。その前項・後項動詞は,「食べ始める」（＝食べるのを始める）「使い慣れる」（＝使うのに慣れる）といったように補文的な関係を持ち,「意味関係は完全に透明かつ合成的」[2] であるとされている。統語的複合動詞は，このように，前項動詞と複合動詞との間の構文的関係と意味的な関係が明らかなため，本論文では語彙的複合動詞だけを分析対象とすることにする。

3. 結びつきを計る指標

3.1 重複度

本稿では，複合動詞と構成動詞の周辺文脈を比較し，その重複する度合い（以後,「重複度」）を両者の意味的な共通性の度合いと考える。その際，分析の対象が複合動詞ということから，複合動詞の周辺文脈を基準として，構成動詞の周辺文脈が重複する度合いを計算する。以上の 2 点を勘案し，次のように 2 種類の重複度を定義した。$Overlap_v$ は周辺文脈の頻度情報を考慮した重複度，$Overlap_s$ は頻度情報を考慮せず，存在のみを考慮した重複度である。

$$Overlap_v(w_c, w_s) = \frac{\sum_{w_a \in W_c \cap W_s} F_c(w_a)}{\sum_{w_b \in W_c} F_c(w_b)} \quad (1)$$

$$Overlap_s(w_c, w_s) = \frac{|W_c \cap W_s|}{|W_c|} \quad (2)$$

なお，w_c，w_s はそれぞれ複合動詞，構成動詞，W_c, W_s

はそれぞれ w_c，w_s と共起する語の集合，$F_c(w)$ は語 w が w_c と共起する頻度である。

3.2 既存の指標

語と語の間の類似性を計算する試みは，従来より類義語の自動獲得やシソーラスの自動構築などの研究で広く行われている。ここでは，相澤の研究 [5] を参考に，次の三つの指標を比較の対象とする。ただし，Simpson 係数については，多くの複合動詞において [1]，$min(|W_c|, |W_s|) = |W_c|$ となり，$Overlap_s$ と同一であるため，比較は $Overlap_s$ で代用する。

$$
\begin{aligned}
Simpson(w_1, w_2) &= \frac{|W_1 \cap W_2|}{min(|W_1|, |W_2|)} \\
Jaccard(w_1, w_2) &= \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|} \\
COS(w_1, w_2) &= \frac{\mathbf{w_1} \cdot \mathbf{w_2}}{|\mathbf{w_1}||\mathbf{w_2}|}
\end{aligned}
$$

なお，$\mathbf{w_1}$，$\mathbf{w_2}$ は，それぞれ語 w_1, w_2 と共起する語の出現頻度のベクトルを表す。

4. 実験

4.1 方法

指標の評価は，指標の計算値と人手による関係評価と比較することによって行う。指標を計算する際の周辺文脈としては，ヲ格の格要素を用いる。これは，(a) 対象語周辺の共起語を一律に周辺文脈とする手法などとは異なり，計算値に与える言語要素の影響を分析しやすいこと，(b) 語間の類似度計算に対して，格情報の有効性が検証されていること [5] を考慮した。

人手による関係評価は，複合動詞の用例の格要素を構成動詞に適用し，次のチェックを行った。3 種類以上の格要素でチェックが成り立った場合，複合動詞と構成動詞は関係ありと判断する。人手関係評価の結果を表 1 に示す。

- 当該の格要素が構成動詞の格要素としても適格か
- 当該の構成動詞・格要素の意味が，複合動詞の意味の一部として適切か

表 1: 人手関係評価の結果

	関係あり	関係なし
前項動詞	426	86
後項動詞	269	243

[1] $min(|W_c|, |W_s|) = |W_s|$ を満たしやすい構成動詞の例として，後項動詞「込む」（例：走りこむ）や古語（「にじり寄る」の「にじる」）を挙げておく。

4.2 実験データ

実験データとして,「複合動詞用例データベース」[6] を用いる [2]。このデータベースには，3362 語の複合動詞，および，その構成動詞 1040 語を収録している。それぞれの語には，格解析済みの用例が用意されている。なお，形態素解析は Juman (ver.6.0)，構文解析・格解析には KNP(ver.3.01) を用いた。

用例は，Baroni らの手法 [4] を応用して，Web から収集した。具体的には，個々の動詞ごとに Web コーパスを構築し，そこから対象とする動詞の用例だけを抽出している [6]。コーパス構築の際は，検索エンジンに収集対象語とランダムなキーワードを与え，特定のサイトへの偏りを防いでいる。また，Web 特有の過度の引用に対応するため，完全に同一の用例は重複登録しないようにしている。

本実験では，このうち十分な用例を持つ複合動詞からランダムに 512 語抽出して，実験対象の複合動詞とした。取り出した複合動詞の条件は，用例数 1000 例以上，かつ，ヲ格つきの用例 50 例以上である。なお，この条件に合致する複合動詞は，1817 語である。

4.3 実験結果

3 節で示した 4 種類の指標値のヒストグラム（縦軸は全体に対する割合）を図 1〜8 に示す。なお,「関係あり」の図は，人手により関係ありと判断された複合動詞と構成動詞間の指標値,「関係なし」の図は関係なしと判断された場合の指標値である。

また，指標値の平均値と，人手評価結果との誤差の平均値を表 2 に示す。なお，誤差は，指標値を v としたとき，$1-v$(関係ありの場合)，v(関係なしの場合) とする。

表 2: 指標値と誤差（ともに平均値）

	関係あり	関係なし	平均誤差
$Overlap_v$	0.58	0.16	0.34
$Overlap_s$	0.42	0.11	0.43
$Jaccard$	0.09	0.02	0.63
COS	0.26	0.05	0.52

5. 考察

5.1 構成動詞が誤差に与える影響

人手による判断結果との比較で言えば，表 2 より $Overlap_v$ (=0.34) が最も平均誤差が少なく，$Overlap_s$ (=0.43)，COS (=0.52)，$Jaccard$ (=0.63) と続く。

[2] このデータベースは，`http://csd.ninjal.ac.jp/comp` で一般公開している。

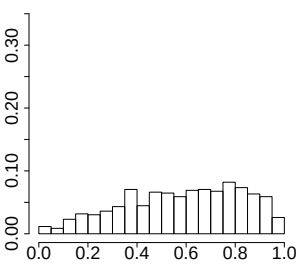

図 1: $Overlap_v$ (関係あり)

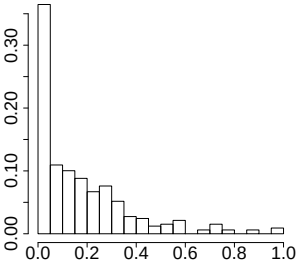

図 2: $Overlap_v$ (関係なし)

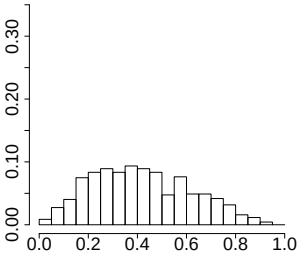

図 3: $Overlap_s$ (関係あり)

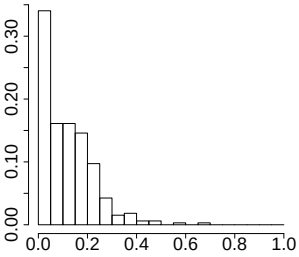

図 4: $Overlap_v$ (関係なし)

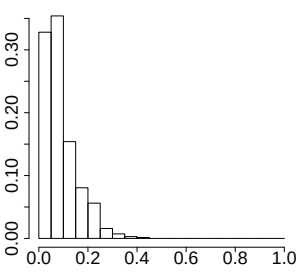

図 5: Jaccard (関係あり)

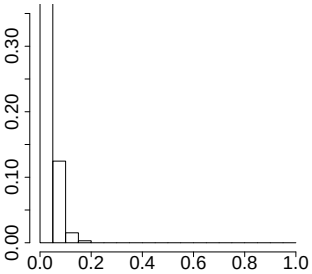

図 6: Jaccard (関係なし)

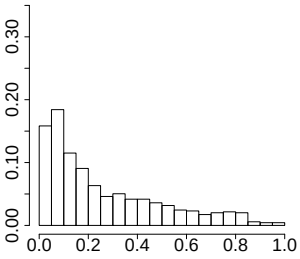

図 7: COS (関係あり)

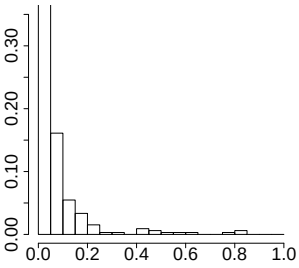

図 8: COS (関係なし)

この結果の主な要因は，構成動詞の扱いにある。具体的には，$Overlap_v$，$Overlap_s$ は，複合動詞の格要素集合に対する重複の度合いを計算している（式 1，2）が，COS，$Jaccard$ は構成動詞の格要素集合も含めて計算している。そのため，COS，$Jaccard$ では，構成動詞にしか現れない格要素集合の部分が誤差となる。具体的な例を次に挙げる。

構成動詞の多義性　「買い集める」と「買う」の場合，「買う」の語義のうち，「購入する」の意味が引き継がれるが，「怒りを買う」「板前の腕を買う」などの意味は引き継がれず，構成動詞側の用例にしか出現しない。

複合による格要素の制限　「飲む」の場合，飲み物一般が格要素となるが，「飲み明かす」の場合，一般的に酒類に制限される。このように，複合動詞の構成要素となると，語義的には同一でも格要素が制限され，構成動詞側にしか用例が出現しない場合がある。

以上のような構成動詞の扱いの影響は，指標値の分布にも現れており，COS，$Jaccard$ の分布（図 7，5）は，「関係あり」でも右肩下がりの分布になる。平均値を見ても，それぞれ 0.26，0.09 と，$Overlap_v$，$Overlap_s$ の 0.58，0.42 と比較して小さい。

5.2　頻度情報が誤差に与える影響

ここでは，頻度情報を考慮している $Overlap_v$ と，していない $Overlap_s$ を比較することにより，頻度情報が誤差に与える影響を考察する。

表 2 より，平均誤差の上では，$Overlap_v$ (=0.34) のほうが，$Overlap_s$ (=0.43) がよりも誤差は小さく，頻度情報を考慮したほうが性能が高い。この原因として考えられるのは，$Overlap_v$ では，複合動詞側の格要素の出現頻度が重みとして機能する (式 1 参照) ため，解析誤りなどの面から信頼性の低い，低頻度の格要素の評価が相対的に低くなることである。その一方で，$Overlap_s$ では，頻度に関係なく，一律に指標値が計算される。

ただし，構成動詞側の出現確率が複合動詞側よりも極端に低くなる格要素は，注意が必要である。なぜならば，構成動詞側の格要素の信頼性が低いにも関わらず，指標値への影響が高いからである。現在，$Overlap_v$ はこの問題に対する対策を取っておらず，今後の課題である。

5.3　構成動詞の複合時の意味的変化

5.1，5.2 の結果により，$Overlap_v$ の有効性が確認された。そこで，$Overlap_v$ を使って，構成動詞の複合時の意味的変化を分析してみる。

図 1〜8 を見ると，指標値と人手の判断結果とが大きく異なる場合がある。例えば，図 1 の 0 近辺，図 2 の 1 近辺である。ここは，構成動詞と複合動詞とで異なる格要素を取るにも関わらず，人手で意味的に共通していると判断された場所であり，複合時に意味的な変化が起きている可能性がある。

そこで，実際の用例を個別に人手で参照することにより，その原因を探る。実例を参照する対象となる複

合動詞 w_c，構成動詞 w_s の条件は，次の条件のいずれかを満たすものとする。

条件 1 人手判断結果「関係あり」の場合

$Overlap_v(w_c, w_s) \leq 0.3$

条件 2 人手判断結果「関係なし」の場合

$Overlap_v(w_c, w_s) > 0.3$

この条件を満たす複合動詞の用例から，人手の判断結果に反する格要素（出現頻度上位 5 位まで）を持つ用例を抽出する。その結果，人手判断結果と異なる原因として，次のものがあることがわかった。

> 複合動詞の多義性，構成動詞の多義性，格要素の出現確率の変化，システムの解析誤り，誤用

このうち，出現頻度が多い，三つの原因を詳しく説明する。

複合動詞の多義性　複合動詞が，人手判断した時に想定した語義とは異なる語義を持ち，その語義の用例の割合が高い場合である。この場合，複合時に構成動詞に意味的な変化が起きた可能性がある。

実際の例として,「打ち破る」と「打つ」の例がある。この例の場合,「塀を打ち破る」「塀を打つ」などの用例が考えられることから,「関係あり」と人手判定した。しかし,「打ち破る」は，抽象物を格要素とする語義（例：「現実を打ち破る」）もあり，この語義では「打つ」の格要素として不適格となる。この事例の場合，後者の語義の割合が高かったため，指標値の誤差が大きくなった。なお，後者の語義は，前者の語義のメタファーになっていると考えられる。

格要素の出現確率の変化　複合動詞の格要素としては用例が存在するが，構成動詞では出現確率が低く，構成動詞の用例としては，複合動詞用例データベースに登録されていなかった場合である。注意すべきことは，構成動詞の格要素としては適格であるということである。この場合，複合時に格要素の出現頻度に変化が起きている。

実際の例としては,「村を焼き払う」に対する「村を焼く」を挙げる。この場合，どちらの文も適格だが,「村を焼く」に相当する用例はデータベースに存在しなかった。

構成動詞の多義性　構成動詞の多義性により，誤って，重複利用される格要素と判定される場合である。そのため，構成動詞の語義が変化した根拠とはならない。

具体的な例として,「攻め立てる」と「立てる」の例を挙げる。この例の場合,「攻め立てる」の「立てる」は接辞的な用法であるため,「関係なし」と判断した。しかし,「立てる」に「相手を立てる」のような語義もあることから，指標値が上昇し，人手判断と異なる結果となった。

動詞の前項/後項，用例の参照条件（条件 1, 2) ごとに，原因の出現数を集計する（それぞれ出現頻度上位２位の原因）と次のようになる。なお，各項目ラベルの括弧内の数字は出現頻度である。

前項・条件 1 (47) 複合動詞の多義性 (57.4%)，格要素の出現確率変化 (53.2%)

前項・条件 2 (18)
構成動詞の多義性 (61.1%)，誤用 (22.2%)

後項・条件 1 (51) 格要素の出現確率変化 (74.5%)，複合動詞の多義性 (27.5%)

後項・条件 2 (46)
構成動詞の多義性 (80.4%)，解析誤り (13.0%)

以上の結果から，複合時の構成動詞の意味的変化に関して，次のことが示唆される。

- 複合時の構成動詞の意味変化として，複合動詞の多義語化への関与，および，格要素の出現確率の変化があること
- 前項・条件 1 と後項・条件 1 により，複合時の構成動詞の意味変化は，前項動詞，後項動詞ごとに分布が異なること

6. おわりに

本稿では，日本語の複合動詞とその構成動詞の意味的な共通性を計るための指標として，重複度を提案し，既存の語間の類似度計算法と比較して，その有効性を確認した。また，重複度を用いて，複合時の構成動詞の意味的変化として，複合動詞の多義語化への関与，格要素の出現確率の変化を確認した。

参考文献

[1] 山本清隆：複合動詞の格支配，都大論究，Vol.21，pp.32-49 (1984)

[2] 影山太郎：文法と語形成，ひつじ書房 (1993)

[3] 由本陽子：複合動詞・派生動詞の意味と統語 (2005)

[4] M. Baroni and S. Bernardini: BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004 (2004).

[5] 相澤彰子：大規模テキストコーパスを用いた語の類度似計算に関する考察，情報処理学会論文誌 Vol.49, No.3，pp.1426–1436 (2008)

[6] 山口昌也：複合動詞と構成要素動詞の格要素の対応関係分析，言語処理学会第 18 回年次大会予稿集 (2012)