コリゲーションの抽出における形態統語情報の役割

千葉 庄寿 (麗澤大学)
schiba@reitaku-u.ac.jp

本発表では，フィンランド語と日本語を主な分析対象として，類型論的観点からコリゲーション抽出に必要な手順を検討するとともに，汎用的なコリゲーション抽出ツールの要件を提案する。

- コリゲーションの分析における形態統語的情報の位置づけ：コロケーションの単位は「語」であり，語形情報 (出現形，代表形) のみならず，語のもつ形態統語的情報もコリゲーション分析に用いる必要がある。
- 有意味なコリゲーション認定の条件：形態統語的情報をコリゲーションの統計的評価に導入する場合，語と語の単純な共起関係の評価に用いる統計的手法をそのまま用いることができない。

1 コロケーションとコリゲーション

コロケーションは一般に連語関係，つまり語と語の組み合わせ関係を表し，「一般的な文法規則に則って共起する語と語の慣用的な結びつき，またその語句のこと」を指す(『応用言語学辞典』p. 658)。「慣用的な結びつき」がしばしば「慣習的結合」(滝沢 2007:18)や「繰り返し現れる連鎖」(Kjellmer 1991:116)といった表現で置き換えられることから分かるように，コロケーションが指す現象の中心は，定型表現のもつ組み合わせの固定性や慣用句性を必ずしももたない「ある語がそれと共起しうる数多くの語のうち特定のものと共起する傾向」にある(村木 2007)。

コロケーションを構成する要素の種類として，Kjellmer (1991:114)は以下のような分類をおこなっている。

(1) 2 つ以上の語彙的な語によるもの。文法的な語が付随することもある

(2) 1 つの語彙的な語と 1 つ以上の文法的な語によるもの

このうち，(2)にあたる現象を特にコリゲーション colligation と呼ぶ。

> The collocation between a lexical word and a grammatical one is frequently termed ‘colligation’. (Hunston 2002:12, n.1)

コリゲーションの概念は Firth により提唱されたものを Hoey らが洗練させたもので(Hunston 2001; Hoey 2005)[1]，文法パターンの記述にコロケーションが関わることで，語彙論と文法論に垣根がないこと(Sinclair 1991)，また文法的振る舞いの分析にコーパスからの量的証拠が役立つことを示すものであるとされる (Hunston 2001:15, cf. Teubert 2007)。

Hoey (2005:43)はコリゲーションを以下の(3)～(5)のように定義する。

(3) the grammatical company a word or word sequence keeps (or avoids keeping) either within its own group or at a higher rank

(4) the grammatical functions preferred or avoided by the group in which the word or word sequences participates

(5) the place in a sequence that a word or word sequence prefers (or avoids)

この定義から，コリゲーションには語をとりまく文法環境だけでなく，語そのものがもつ文法機能も関わることがわかる。語を単位としてコロケーションを分析する場合，後者は純粋に検索語のもつ形態論情報として現れる (cf. 松村 2001)。

Hoey (2005)は英語の動詞 *ponder*「熟考する」が受動文で現れる傾向を挙げているほか，主語，目

的語，前置詞句といった特定の統語的位置を出現環境として好む名詞があることを論じている。

2 類型的言語特徴とコリゲーション

語と語の単純な共起関係を発展させた現在のコリゲーションの捉え方は，一語に多数の文法素性が共存する形態論的に複雑な言語には適用することができない。本節ではコロケーションの単位としての「語」が含みもつ情報が言語類型論的に斉一でないという問題を考察する。

(6) *Tommi Mäkinen käv-i tutki-ma-ssa*
T M.NOM visit-IMP.3SG research-3INF-INE
murheellise-n ojan=penka-n
sad-GEN ditch=edge-GEN
muutama tunti törmäykse-nsä jälkeen.
several hour.NOM crash.GEN-PX3 after
「T. マキネンは自身のクラッシュの数時間後，哀しみの道路脇を調べにやってきた」
(aamu1999) [2, 3]

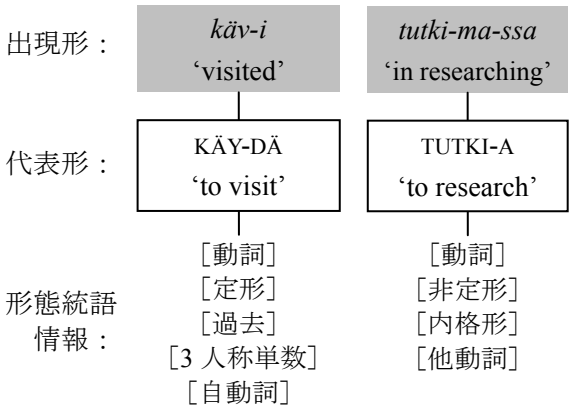

図 1：主動詞と不定詞のコリゲーション

日本語のコロケーション分析では，しばしば形態素区切り(短単位)による共起関係の評価が行われている(e.g. 深田 2007)。これは長単位の自動解析が一般的でないことにも起因するが，フィンランド語のように，接尾辞が語に密着し添加されて順序がはっきり決まっており，その自由度は低く，形態音韻的な操作が行われる言語には適切ではない。一方で，フィンランド語では英語の機能語にあたる要素が屈折語尾として添加され，内容語と一語で表示される。その結果，出現形や代表形といった単純な調査では形態統語情報がもつコリゲーションのパターンの全体像を効率よく得ることはできない。

なお，日本語の助詞は一般に付属語(後倚辞)として扱われ，形態論的に独立した語としての位置づけを与えうる(宮岡 2002)。一方，フィンランド語の小辞も後倚辞と分析されるが，文の情報構造と密接に関係し，文頭の最初の要素の後に添加される。

(7) ***On-ko*** *ohjelma-ssa-ni virhe?*
be.3SG-Q program-INE-PX1SG mistake.NOM
「私のプログラムの中に間違いは<u>ある</u>**か**？」

(8) ***Ohjelma-ssa-ni-ko*** *on virhe?*
program-INE-PX1SG-Q be.3SG mistake.NOM
「間違いがあるのは<u>私のプログラムの中に</u>**か**？」

(9) ***Virhe-kö*** *ohjelma-ssa-ni on?*
mistake.NOM-Q program-INE-PX1SG be.3SG
「私のプログラムの中に<u>間違いが</u>ある(という)の**か**？」

フィンランド語の小辞は文の位置情報と関係して添加され，日本語の後倚辞とは異なった出現傾向をもつ。このような要素のコロケーション分析での扱いは，言語によって大きく異なってくるといえる。

なお，形態統語的情報を厳密に指定するためには情報のもつ曖昧性を排除する必要がある。Farrar et al. (2003)などが提唱する言語学的オントロジーの階層構造を用いることで，形態統語情報を曖昧でなく分類処理することが可能となる。

3 コリゲーション認定の条件

「あるテキストにおいて共起している 2 ないしそれ以上の語」(Sinclair 1991: 170)の関係がどの程度有意味であるのか，有意味なコロケーションの発見の手順はできる限り自動化することが望ましい。これまでコロケーションの自動認定のための統計処理法について多くの提案がなされてお

り (Manning et al. 1999), 判定の枠組みは文法機能の共起情報の分析にかかわるコリゲーション情報の抽出にもある程度用いることができると考えられる。

コロケーション分析ツールの中には，共起語の頻度のみを抽出・表示するものも多い(Barlow 2003)。しかし，語の単純な出現頻度だけでは，その単語の出現頻度の大小により共起頻度が偏ってしまうため，共起情報のより厳密な評価に際しては，コーパスの総語数と実際の出現頻度から，期待頻度 (偶然出現した場合の頻度) を算出して利用することが一般的である。期待頻度と実際の出現頻度のずれを観察することで，特徴的な共起語をある程度見つけることができる。

$$(10)\ \text{期待頻度} = \frac{\text{共起語の出現頻度}}{\text{コーパスの総語数}} \times \text{基準語の出現頻度}$$ [4]

しかし，現在のコロケーション認定の統計的手法は，２節でみたような，一語に多数の文法素性が共存する形態論的に複雑な言語には適用できない。事前に各要素の出現頻度を計算することが難しいうえ，コーパスの総語数にあたる要素数を計算することが不可能になるためである。

期待頻度の算出には，多くの場合，基準語と共起語の頻度とコーパスの総語数を用いた期待頻度が統計処理に用いられる。しかし，３節でみたように，総語数にあたる数値の相対化をおこなうための要素の算出が難しい。そこで(特定の形態統語的情報をもつ)基準語と共起語の頻度情報と，これらの共起頻度のみで算出できる結びつきの強さの算出方法として，MI スコア(Barnbrook 1996: 98-100; cf. Hunston 2002: 70ff)から得られるランキング情報を利用することを提案する。[5]

$$(11)\ I = \log_2 \frac{\text{語Ａと語Ｂの共起頻度} \times \text{コーパスの総語数}}{\text{語Ａの頻度} \times \text{語Ｂの頻度}}$$

MI スコアにおいてはコーパスの総語数は単純に積算されるだけなので，スコアのランキング自体にはコーパスの総語数の情報は影響しない。

４ コリゲーション分析ツールの試作

基準語が決まっている場合，まず検索対象を含む用例をコーパスから取得した後，表示したい共起情報を選ぶことになる。この場合，コロケーションを検証するための共起情報として，以下のような内容を指定することが想定できる。

(12) a. 出現形：イディオム的共起パターン抽出
b. 代表形：いわゆるコロケーションパターンの抽出
c. 形態統語的素性：形態統語情報に基づくコリゲーションパターンの抽出

多くのコロケーション分析ツールは(12a)(12b)に対応している。(12b)はいわゆるコロケーションにあたる意味をもつ語同士の結びつきを検証することはもちろん，英語や日本語の助詞のような機能語によって形成される構造(valency, Teubert 2007, cf. ‘collocational framework’, Renouf et al. 1991:128-129)を含むコリゲーションパターンの抽出を行うことも可能である。しかし，本発表で提案する，類型論的に妥当性のある語を単位とするコロケーション分析においては(12c)の分析ができることが望ましい。

本研究では，基準語及び共起語の形態統語情報による用例の絞り込みを行ない，コリゲーションパターンの抽出と評価を行うためのツールのプロトタイプを作成した (図２参照)。Bank of English のピクチャ画面に準じ，検索結果は共起後としてランキングの高いものから順に表示をおこない，さらに共起語の形態統語情報を選択して用例の絞り込みをおこなうことでスパン内の統計情報を動的に更新し，MI スコアのランキング表示を行う。

なお，MI スコアで高い値をとる共起語はコロケーションに制限のある頻度の低い語である傾向があり(idem. p. 74)，このことは複雑な形態統語情報を指定した検索の際に障害となる可能性があり，共起語の頻度に一定の制限をつける(さらに共起語の頻度を色分け表示する)などの工夫をすることが肝要と考えられる。

参考文献

Barnbrook, Geoff (1996). Language and Computers: a Practical Introduction to the Computer Analysis of Language. Edinburgh: Edinburgh University Press.

Barlow, Michael (2003) Concordancing and Corpus Analysis Using M(onoConc)P(ro) 2.2. Houston: Athelstan.

Clear, Jeremy (1993). From Firth principles: computational tools for the study of collocation. In Baker, Mona, Gill Francis & Elena Tognini-Bonelli (eds.) Text and Technology: in Honour of John Sinclair. Amsterdam: John Benjamins, pp. 271-292.

Farrar, Scott & Terry Langendoen (2003). A linguistic ontology for the semantic web. Glot International 7: 97-100.

深田淳 (2007). 「日本語用例・コロケーション抽出システム『茶漉』」『日本語科学』22: 161-172.

Hoey, Michael (2005). Lexical Priming: A New Theory of Words and Language. London: Routledge.

Hunston, Susan (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Hunston, Susan (2001). Colligation, lexis, pattern, and text. In Scott, Mike & Geoff Thompson (eds.) Patterns of Text: In Honour of Michael Hoey. Amsterdam: John Benjamins, pp. 13-33.

Kjellmer, Göran (1991). A mint of phrases. in Aijmer, Karin & Bengt Altenberg (eds.) English Corpus Linguistics. London: Longman, pp. 128-143.

Manning, Christopher D & Hinrich Schütze (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

松村一登 (2001). 「フィンランド語の名詞の意味と場所格の使用頻度の関係について―コーパスのデータに基づく研究―」『梅田博之教授古希記念 韓日語文学論叢』ソウル: 太学社, pp. 1161-1206.

宮岡伯人 (2002). 『「語」とはなにか : エスキモー語から日本語をみる』三省堂.

村木新次郎 (2007). 「コロケーションとは何か」『日本語学』2007 年 10 月号. Pp. 4-17.

Renouf, Antoinette & John M. Sinclair (1991). Collocational frameworks in English. In Aijmer, Karin & Bengt Altenberg (eds.) English Corpus Linguistics. London: Longman, pp. 128-143.

Siepmann, Dirk (2005a). Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. International Journal of Lexicography. 18: 409-443.

Siepmann, Dirk (2005b). Collocation, colligation and encoding dictionaries. Part II: Lexicographical aspects. International Journal of Lexicography. 19: 1-39.

Sinclair, John M. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

滝沢直宏 (2007). 「コーパスを用いた英語研究の方法」『日本言語学会第 134 回大会予稿集』Pp. 18-23.

Teubert, Wolfgang (2007). Sinclair, pattern grammar and the question of *hatred*. International Journal of Corpus Linguistics 12: 223-248.

---

[1] 機能語を含むコロケーションを扱う研究でも，文献によってこの術語が用いられないことがある (Hoey 2005:43)。

[2] データは Kielipankki による。Kielipankki はフィンランドの教育省管轄の学術情報技術センター CSC (The Finnish IT Center for Science) が運営するフィンランド語を中心としたコーパスサービスである。URL: `http://www.csc.fi/tutkimus/alat/kielitiede`

[3] グロスに用いる略号は以下の通り：
1,2,3=人称; 1INF=第1 不定詞; 3INF=第3不定詞; ADE=接格; ALL=向格; APT=行為者分詞; GEN=属格; ILL=入格; IMP=過去; INE=内格; NOM=主格; PAR=分格; PASS=受動; PL=複数; PX=所有接尾辞(+人称,(数)); Q=疑問の小辞; SG=単数

[4] 単なる単語と単語の共起でなく，複数の語からなるスパンを想定する場合には，基準となる語の出現頻度にスパン分の語数をかける。

[5] 奈良先端科学技術大学院大学の松本裕治先生とのディスカッションによる。