日伊コロケーション辞書の作成を目指す『現代日本語書き言葉均衡コーパス』からのコロケーションの検出と分析

STRAFELLA Elga Laura（奈良先端科学技術大学院大学）
松本裕治（奈良先端科学技術大学院大学）

Towards a Japanese-Italian Collocation Dictionary: Detection and Analysis of Collocations from the *Balanced Corpus of Contemporary Written Japanese*

Elga Laura Strafella (Nara Institute of Science and Technology)
Yuji Matsumoto (Nara Institute of Science and Technology)

１．はじめに

本研究では、日伊コロケーション辞典を編集するために、コロケーション検出に広く用いられる単純頻度、相互情報量 (PMI)、対数尤度比 (Log-Likelihood Ratio)、ダイス係数 (Weighted Dice) という4つの指標を使用し、大規模な『現代日本語書き言葉均衡コーパス』(BCCWJ) からコロケーションを検出しながら、文法と意味解析を行う。データセットとしては、日本語能力試験の過去の出題基準となっていた１級の語彙リストを用い、コロケーションを検討する。統計的なアプローチで BCCWJ から連語 （「名詞＋動詞」「名詞＋イ形容詞」）を検出した上、統語論・意味論の特徴によって各語句の分類を行う。

次に、本研究の背景（２章）について述べてから、分析の流れと検出方法に関する特徴について簡潔に説明する。また、検出したコロケーションのいくつかの例を挙げて、それらの意味と例文も挙げる（３章）。最後に、まとめと今後の課題について述べる（４−５章）。

２．研究の背景

本研究は日本語における「コロケーション」の使い方を身に付けるために行なっている。日本語学習者は、何かを言いたいと思って辞書を引き、言いたいことに当たる名詞を見つけても、その名詞と共にどの動詞を使えばよいか分からなければ困るのであろう。現在に存在している日本語辞典には名詞に連続する動詞の用例が載っていないし、載っていても例が短いため、実際の用法が分かりにくいことが多い。例えば、'Have you **made** any **plans** for the summer vacation? ' と聞きたいと思っても、「計画」の後にどんな動詞を使えばよいのか分からないと、「夏休みの**計画を**何か**しました**か」「夏休みの**計画を**何か**作りました**か」と言ったりしてしまいがちである。しかし、ここでは「計画を立てる」というコロケーションを使わないといけないのである。また、「頭」には「頭が**いい**」という形容詞の他に「頭が**悪い**」「頭が**固い**」「頭が**古い**」などのコロケーションがあり、これらを使うことによって日常生活で的確な表現をすることができる。こうした知識を得るために、まずコロケーションを整理する必要がある。

第３章「分析の流れ」は、検索コーパスやデータセットの設計などについて詳しく述べる。

３．分析の流れ

データセットとしては、日本語学習者の初級者にも役に立つ資料を編集するため、日本語能力試験の過去の出題基準となっていた１級の語彙リスト（約8,000語）を用い、『現代日本語書き言葉均衡コーパス』を検索した。一般の会話だけではなく、文学書や新聞などでも使われている現代日本語を検討するため、均衡コーパスを使うことにしたのである。

３.１　抽出

はじめに、 BCCWJ からの係り受けを抽出した。具体的にいうと、 BCCWJ Dependency Extraction Toolkit [1] というプログラムで CaboCha [2] で解析した結果から、「名詞が動詞またはイ形容詞に係っている事例」を抽出した。「名詞＋格助詞＋動詞」という係り受けは 385,470 あり、 「名詞＋格助詞＋イ形容詞」という係り受けは 13,353 あった。それから、それぞれの係り受けの頻度と PMI、Log-Likelihood Ratio、Weighted Dice を計算した。以前の研究でも明らかになった通り、コロケーションの検出ではよく使われているいくつかの統計的な指標は語と語の結びつきの強さについてそれぞれの特徴を現すので、一つだけを使うと分析結果を信頼しにくくなる。[3] 頻度だけでも判断してしまうと役に立てないデータが多くなる。例えば、「方が良い」(f = 7,012)、「方が多い」(f = 845)、「方が高い」(f = 254) などはとても高い頻度があっても、日本語としてはそのままには使えないと考えられる。つまり、その前に来る表現によって伝えられる意味が変わってしまう。例えば、「方が良い」の場合、「寝た**方が良い**」と言えるし、「出た方が良い」とも言え、伝えている意味は違う。

しかし、ここで連続している名詞「方が」とそれぞれの形容詞「良い」「多い」「高い」の単純頻度は比較的に高いため、一緒に現れる頻度も多い。こういう偏ったデータを避けるために各コロケーション候補に対して、今回利用した３つの共起尺度 (PMI, Log-Likelihood Ratio, Weighted Dice) による共起度の順番をもとめ、その逆数の平均値を示す Mean Reciprocal Rank (MRR) [4] を計算し、分析を進めた。 MRR によって各連語をソートすると高い数値を示すものは３つの共起尺度によって総合的にコロケーションである可能性が高いと支持された表現であると言える。

３.２　分析

抽出したデータの分析は一つずつ手動で行われているため、時間が非常に掛かる。それで、以前の研究 [5] では「名詞＋格助詞＋動詞」しか扱われてなかったため、今回は「名詞＋格助詞＋イ形容詞」だけを分析した結果について述べる。

まず、関連研究に基づいていわゆる frequency threshold [6] を適用することにした。広く用いられるのは $f \geq 3$、$f \geq 5$、$f \geq 10$ であり、データを削減するために一番高い数値を適用し

---

[1] 解析プログラムは奈良先端科学技術大学院大学　自然言語処理学研究室　博士後期課程の林部祐太に作られた.

[2] code.google.com/p/cabocha/

[3] Strafella, 2013.

[4] en.wikipedia.org/wiki/Mean_reciprocal_rank

[5] Strafella, Hayashibe, Matsumoto, 2012.

[6] Evert, 2007: 5.

た結果、997 の事例を得た。これを一つずつ分析し、統語論・意味論の特徴によって分類を行った。表１には7つの名詞に関する例を挙げている。MRR の一番高い数値は太字になっている。

表１データサンプル

連語	頻度	PMI	Log-Likelihood Ratio	Weighted Dice	MRR
数が多い	455	1.764420904	1439.231531	0.316377828	0.5
数が少ない	348	2.984089786	2216.775566	0.87863168	**1**
数が大きい	18	0.186751859	0.86599955	0.025617287	0.305555556
数が高い	15	-0.390397562	3.845040344	0.014466393	0.277777778
事が多い	339	1.499015408	808.4376951	0.224693869	**0.833333333**
事が良い	31	-1.059835123	78.46334711	0.010324712	0.388888889
事が嬉しい	14	2.089499352	49.58983067	0.061022288	0.611111111
運が良い	304	1.876938586	1106.195885	0.171873048	0.666666667
運が悪い	119	2.469259606	566.2383853	0.241212205	**0.833333333**
運が強い	15	0.852461806	12.2699912	0.025782384	0.333333333
感じが良い	117	0.882616398	112.1124388	0.055053262	**0.833333333**
感じが強い	20	1.100675457	25.47720714	0.037820416	0.666666667
感じが悪い	18	0.541039449	6.540102027	0.021985545	0.333333333
幅が広い	95	4.930777293	1134.423075	1.197958316	**1**
幅が大きい	46	2.196344774	180.0766338	0.100587431	0.305555556
幅が狭い	40	4.196325143	380.3666159	0.440739387	0.5
幅が小さい	16	2.311900078	66.07301857	0.068965517	0.277777778
頭が良い	253	0.915291199	259.4077687	0.135060678	0.25
頭が痛い	156	3.715031998	1291.841864	0.938112048	**1**
頭が悪い	82	1.318846185	143.5533177	0.138391098	0.261111111
頭が可笑しい	65	3.27056167	445.6198181	0.382279207	0.5
頭が重い	26	2.461868949	118.0045722	0.123633215	0.194444444
頭が固い	24	2.947460102	141.207311	0.128926889	0.244444444
頭が大きい	12	-0.314097396	1.929147147	0.014380595	0.130952381
頭が高い	10	-0.891246817	16.0490946	0.008077637	0.136904762
気が重い	94	3.407324739	688.5074666	0.488218187	**1**
気が強い	60	1.121004539	79.25590641	0.121707911	0.197619048
気が弱い	50	2.397068731	219.6111854	0.200706123	0.388888889
気が小さい	40	1.721740933	105.8618708	0.127700734	0.261111111
気が早い	30	1.330232622	52.29419161	0.084334986	0.163888889
気が短い	24	1.897810298	73.7165589	0.083967264	0.186507937
気が良い	20	-1.96210847	252.43963	0.005676478	0.25
気が荒い	12	3.161715443	78.3686147	0.042998051	0.26984127

表１に示してあるように、MRR の一番高い数値を示す連語はコロケーションであり、単純頻度と違っている場合もある（「運が悪い」「頭が痛い」）。

関連研究においてはデータをソートするためによく使われる方法がもう一つあり、１つの指標だけを選んでソートすることである。しかし、本研究は３つの指標を使っているため、それらを総合的に考慮するため MRR を用いることにした。

また、それぞれの指標を検討してみると数値がばらばらになっているため、統計的な解析ができないことが分かる。つまり、コロケーションの研究には統計的なアプローチでデータを得られるが、その後研究者の判断が必要になるため、手動で分析をする必要がある。次に、我々が分析した結果について述べる。

３.３　結果

上記に述べた 997 の「名詞＋格助詞＋イ形容詞」の連語を分析した結果、約 300 のコロケーションが得られた。これの中には「拡張語彙意味単位」と言えるものもあり、特別な意味を持っていないが、よく使われる表現もある。前者は、「頭が高い：傲慢な、誇りを持って」「腕が良い：有能な、上手な」のように全体的な意味は語と語の意味から少し離れていて、比喩的な意味を表す表現である。後者は、「水が欲しい」「車が欲しい」「山が多い」「火山が多い」のように日常生活でもよく使われる表現である。

はじめに述べた通り、本研究の最終的な目的は日本語の初級者にも役に立つ資料を作ることであるため、それぞれのコロケーションの実際の用法が分かるために用例を挙げないといけない。それは、『国立国語研究所』が構築した NINJAL-LWP for BCCWJ [7] という BCCWJ を検索するためにオンライン検索システムを用いて挙げたいと思っている。具体的な例を挙げると、上記に挙げた「**腕が良い**」の用法を表示する例としては、

a.「板前の**腕が**めっぽう**いい。**」

b.「その年で自分のお店を始められるなんて、よっぽど**腕がいい**んだ。」が挙げられる。

また、「**頭が高い**」の場合、

a.「**頭が高い**からお客さんがつかない。」

b.「**頭が**やや**高く**、少しイビツなフォーム。」という例が挙げられる。

この4つの用例はコロケーションの一つの特徴を示している。つまり、コロケーションはイディオムと違い、文型が固定してないため、中心語と関連する語のあいだに別の語が入ることが可能である。この例では、入っている語は副詞である。

こうした統語論・意味論の分析を行うと、それぞれのコロケーションの実際用法はもっとも分かりやすくなり、学習者に役に立つ情報も得ることができる。

４．まとめ

本稿で行った調査は博士課程の主要議題であり、2010年に始まった。はじめに、関連研究で Multiword Expressions の検出にあたって共起度を測るためによく使われるいくつかの指標を深く調べ、その中で四つ（単純頻度、PMI、Log-Likelihood Ratio、Weighted Dice）を用いることにした。しかし、連語といってもコロケーションだけでなく、自由結合・イディオムなども存在しているため、その中でコロケーションだけを検出するのはコンピュー

[7] nlb.ninjal.ac.jp/

タを使っても複雑なタスクであることが明らかになった。従って、統計的なアプローチでコーパスから強く結びついている連語を検出し、手動で一つずつ分析を行っている。こうして辞典の基になるコロケーションリストを作っている。

５．今後の課題

本研究は、日本語を学習している学習者達に役に立つ資料を作ろうとしているため、「名詞＋イ形容詞」だけでなく、「名詞＋ナ形容詞」「名詞＋動詞」「副詞＋動詞」なども検討する必要がある。今後はそれぞれの文型を検討し、コロケーションリストを完成することを考えている。しかし、データの量が多くなければ多くなるほど分析しにくくなり、それを自動的に分析ができる方法を考える必要がある。また、網羅性の高いコロケーションリストを作成したら、各コロケーションの用例も挙げて、イタリア語に翻訳をしたり、当のコロケーションとの同じ意味を表すイタリア語の表現を挙げたりすることを考えている。

謝辞

本研究は、日本語学術振興会「外国人特別研究員（欧米短期）」（平成 24～25 年度）による補助を得ている。

本研究の推進にあたっては、共同研究者各位、とくに自然言語処理学研究室博士後期課程林部祐太に多くのご指導・ご助言をいただいた。感謝を申し上げます。

文献

Evert, Stefan (2005) *The Statistics of Word Co-occurrences: Word Pairs and Collocations,* Ph.D. thesis submitted to the Institut für machinelle Sprachverarbeitung, Universität Stuttgart.

Evert, Stefan (2007) *Corpora and collocations*, Institute of Cognitive Science, University of Osnabrück, Germany. (Extended Manuscript, 13 October 2007).

姫野昌子 (2012) 『研究社日本語コロケーション辞典』研究社.

Pereira, Lis W.K., Manguilimotan, Erlyn, and Matsumoto Yuji (2013) *Collocation Suggestion for Japanese Second Language Learners*, 情報処理学会研究報告　第 210 回自然言語処理研究、Vol.2013-NL-210, No.3, pp. 1-5, January 2013.

Pereira, Lis W.K., Maguilimotan, Erlyn, and Matsumoto Yuji (2013) Data Coverage vs. Data Size: A comparison of two large-scale corpora in Collocation Suggestion for Japanese Second Language Learners. In *Proceedings of the Nineteenth Annual Meeting of the Association or Natural Language Processing (NLP-2013)*, Nagoya, Japan, March 2013.

Seretan, Violeta (2010) *Syntax-Based Collocation Extraction*. In N. Ide and Véronis J. (eds.) *Text, Speech and Language Technology*, vol. 44, Springer Dordrecht Heidelberg London New York.

Strafella, Elga Laura (2013) *Detection and Analysis of Collocations in Contemporary Japanese – A corpus-based language study*, Ph.D. thesis submitted to the Department of Asian, African, and Mediterranean Studies, University of Naples “L’Orientale”.

Strafella, Elga Laura, 林部祐太、松本裕治 (2012) 『現代日本語におけるコロケーション：検出と分析』、第 1 コーパス日本語学ワークショップ、 pp. 53-58、国立国語研究所、5-6 March 2012.

関連URL

『現代日本語書き言葉均衡コーパス』(BCCWJ)　http:// www.ninjal.ac.jp/corpus_center/bccwj/

BCCWJ Dependency Extraction Toolkit　https:// github.com/shirayu/misc/tree/master/bccwj

Mean Reciprocal Rank (MRR)　http://en.wikipedia.org/wiki/Mean_reciprocal_rank

MeCab　http:// mecab.googlecode.com/svn/trunk/mecab/doc/index.html

NINJAL-LWP for BCCWJ (NLB) 国立国語研究所　http:// nlb.ninjal.ac.jp/