Constructing a Collocation Database for the CEFR-J Wordlist

英語学習用活用語彙リストの提案 — CEFR-J Wordlist のコロケーション・データセットの試み —

福田 航平（東京外国語大学）†
投野 由紀夫（東京外国語大学）††

Constructing a Collocation Database for the CEFR-J Wordlist

Kohei Fukuda （Tokyo University of Foreign Studies）
Yukio Tono （Tokyo University of Foreign Studies）

要旨

外国語学習において、語彙を運用する力を学習者が身につけるためには語彙リストで単語を個別に記憶するだけでなく、使用頻度の高いフレーズで提示するのが重要である。近年、コーパス準拠英語語彙・フレーズリストが発表されているが、対象学習者レベル、受容・産出語彙の区別は明確ではない。本研究では、CEFR 準拠英語学習語彙表 CEFR-J Wordlist の活用度を上げ、学習に資するコロケーション選択をするための資料となるコロケーション・データセットを整備する。1 億語のイギリス英語均衡コーパスである British National Corpus から Universal Dependencies に基づいた構文解析を行い、共起フレーム別（例 : 動詞＋名詞、形容詞＋名詞）に共起語セットを抽出した。教育的に有用なコロケーションを教師や学習者が選定するための情報源として、単純共起頻度以外に検索語-共起語ペアについて、各語の CEFR-J Wordlist に基づいた CEFR レベル情報、共起統計（MI, MI2, MI3, t-score, z-score , logDice, log-likelihood, chi-squared）と散布度指標（DP）の情報を付与した。

１．はじめに

コロケーションは外国語学習において重要な側面であることは広く認識されているが、学習者にとって習得が困難な領域である（Martinez and Murphy, 2011）。コロケーション学習における問題の一つは、学習者が習得すべきコロケーションを判断する基準の合意がないこと （Rogersa, Brizzarda, et. al., 2015）、また個々のコロケーションの学習レベルに関する研究が少ないことである。本研究では、コロケーションの学習レベルを CEFR 準拠英語学習語彙表を活用して付与すると同時に多様な共起統計を付属したコロケーション・データセットを整備することで学習すべきコロケーション特定のための言語教育資料を提供することを目的とする。

２．コロケーション

２.１ コロケーションとは

コロケーションは、最も簡単には「語の自然な組み合わせ」[1]と定義される（McCarthy & O’Dell, 2017, p.4）。コロケーションは研究によって様々な定義がなされているが、大きく２つの方向性に分類できる。構成性に焦点を当てたフレイジオロジー的アプローチと、コーパス情報に基づく頻度情報に焦点を当てた統計的アプローチである（Evert, 2008; Granger &

† fukuda.kohei.r0[at]tufs.ac.jp ※[at]は@に置き換え
†† y.tono[at]tufs.ac.jp

[1] 原文は“a natural combination of words”

Paquot, 2008; Gablasova, Brezina & McEnery, 2017）[2]。前者のアプローチでは語結合の構成性（compositionality）に着目し、自由結合句（free combination）とイディオムの中間に位置するものとしてコロケーションを捉える。この見方では自由結合句とコロケーションとイディオムは連続体として考えられ、客観的にコロケーションとそうでないものを線引きして区別することはできない。後者のアプローチでは、コーパス内での共起頻度や、後述する共起統計指標を用い、数値化してコロケーション性を判断する。これにより一定の数値基準を設け、その基準を満たしているかどうかでコロケーションとそうでないものに区別することができる。設定する基準はその目的に応じて変わり、ある程度恣意的なものとならざるを得ないが、妥当な基準を設定できれば、比較的客観的かつ効率的なコロケーション選択が可能となる。本研究では教育的に有用なコロケーションをその用途に応じて客観的に判断するための統計情報を提供するのが主目的であるため、後者の統計的アプローチでコロケーションを捉える。

２.２　フレーズリスト研究

英語教育分野では語彙表の研究は長い歴史を持ち、多くの教育用語彙表が提案されてきた（中條, 2015）。その後は 1 つ 1 つの単語の振る舞いを学習辞典の形式でまとめる工夫が多く提案され、その中でコロケーションもコーパス研究の成果を活かして積極的に辞書記述に取り入れられていった。日本では戦前にすでに勝俣銓吉郎による『英和活用大辞典』（1939）があり、欧米では BBI Dictionary（revised edition）of English Word Combinations（1997）、the LTP Dictionary of Selected Collocations（1999）、Oxford Collocations Dictionary for Students of English（2009）、 Longman Dictionary of Collocations & Thesaurus（2013）などがコロケーションの辞書としては知られているが、一方で Academic Formulas List（Simpson-Vlach & Ellis, 2010）、 Phrasal Expression List（Martinez & Schmitt, 2012）のような語彙表と同じようなコロケーションやフレーズのリストの提案は比較的最近のことである。特にこれらの資料はコーパスからの機械的な n-gram 情報とそれらをフレーズとしての機能やまとまりなどの観点から人手で評価する情報を合成している。こういった合成の判断基準も研究によってまちまちであり、ある意味で異なる情報の重み付けは恣意的に行われているのが現状である。

３．データ整備

３.１　共起語抽出

コーパスからコロケーションを抽出する方法には大きく 4 つの方法がある（Bhalla & Klimcikova, 2019）。（1）ウインドウスパンを用いた方法、（2）n-gram, （3）正規表現と POS タグを用いた方法（regex-over-pos）、(4) 構文解析による統語依存関係（syntactic dependency）を利用した方法である。(1) のウインドウスパンを用いた手法は中心語（node）の前後数語に共起する語を抽出する手法で、英語コーパス研究で最もよく用いられる手法である。Sinclair（1991）によるコロケーションの定義もこの手法を前提としている[3]。(2) の n-gram は任意の語数の隣接する句を機械的に抽出する手法である。(3) は品詞タグと正規表現によるパターン抽出によって特定の文法関係にある語を抽出する手法で、コーパス検索 Web イ

[2] Evert（2008）では、後者によるコロケーションを”empirical collocations”と呼んでいる。
[3] Sinclair（1991）ではコロケーションを“the occurrence of two or more words within a short span of each other in a text”（p.170）と定義している。

ンターフェースである Sketch Engine の Word Sketch 機能で利用されている（Kilgarriff, Bisa et al., 2014）。

コロケーション抽出方法を比較した研究では、(4) の手法が最も正確にコロケーション抽出をすることができるという結果が報告されている（Bartsch, Evert et al., 2014; Bhalla & Klimcikova, 2019; Evert, Uhrig et al., 2017）。従って、本研究では大規模コーパスに構文解析を行い、統語依存関係に基づき共起フレーム別に共起ペアを抽出した。今回は Python の自然言語処理ライブラリである Stanza（Qi, Zhang et al., 2020）を利用し構文解析を行った。Stanza の構文解析器は Universal Dependencies treebanks を用いて訓練されており、統語依存関係の解析結果は Universal Dependencies に基づいた表示がなされる（de Marneffe, Manning et al., 2021）。

今回のデータ整備に当たっては、British National Corpus XML edition（BNC Consortium, 2007）のテキストデータを用いた。テキスト全体に対して構文解析を行い、解析結果を用いて（a）動詞＋目的語、（b）形容詞＋名詞、（c）名詞＋名詞、（d）副詞＋動詞、（e）副詞＋形容詞の共起ペアを抽出し、全ての共起ペアの頻度を算出した[4]。なお全てレマ単位で抽出・集計している。

３.１.１　動詞＋目的語の抽出

Stanza での構文解析結果を用い obj の係り受け関係（dependency relation）で結ばれている主辞（head）と依存部を「動詞＋目的語」のペアとして抽出した。また、受動態の動詞とその主語、名詞を修飾する過去分詞とその被修飾語も動詞＋目的語のペアとして抽出するために、nsubj:pass の係り受け関係で結ばれている主辞と依存部、acl の係り受け関係で結ばれている主辞と依存部（依存部の語の品詞が xpos で VBN のものに限定）も動詞＋目的語の共起ペアとして抽出した。また、conj の依存関係タグを用いて、1 つの動詞に対して複数目的語が並列されている場合も考慮した。目的語となる名詞は upos のタグが NOUN のものに限定し、代名詞や固有名詞は含んでいない。データセットでは obj のシートにこの共起フレームのデータを収録している。

以下に BNC 中の実際の文から例を示す。1 つ目の括弧は BNC のジャンルとファイル名である。2 つ目以降の括弧には該当文から抽出される共起ペアとその際の係り受け関係タグである。

(1) The pictures look like war photos and the **costs** will not easily be **assessed** but the first estimates to **repair** the **damages** run between $2.5 and $4 million.（W misc; J1B）（nsub:pass; assess + cost）（obj; repair damage）

(2) The firm that **makes** the **ties** **sold** in the country's department stores, is optimistic about prospects at home.（S brdcast news; K6H）（obj; make + tie）（acl; sell + tie）

(3) If anyone could **post** the half-time **score** and final **result** when it comes through I'd appreciate it.（W email; J1G）（obj; post + score）（conj; post + result）

[4] 詳しい処理は Python コード参照のこと。

(4) They therefore **put** forward a wide **spectrum** of policies to **cover** all politically significant **aspects** of nationallife, as well as foreign **affairs**.（W ac polit law edu; GV5）（obj; put + spectrum）（obj; cover + aspect）（conj; cover + affair）

上記の (1) ～ (4) は今回の抽出方法で抽出可能な動詞＋目的語のペアの一例を示している。（3）のように 1 つの動詞に対して複数目的語が付いていて、目的語となる名詞が比較的離れている場合でも抽出可能である。(4) のように as well as で結ばれた aspects と affairs を両方とも cover の目的語として抽出することができている。もちろんこの抽出方法でも完璧ではなく、(4) の例では put と spectrum が動詞＋目的語ペアと判定されている。これは実際には put forward という句動詞の目的語であると判定すべきあり、さらには意味内容を考慮すると、spectrum よりも policy を目的語とする方が正確である。しかしこれはウインドウスパンを用いた抽出方法でも同様であり、例えば名詞を中心語（node）として左 5 右 5 のスパンで共起する動詞を抽出した場合、spectrum の共起語として put と cover が抽出されることになる。以上のことを考えると統語依存関係に基づいた共起語抽出の方が比較的過不足なく目的となる共起ペアを抽出することが可能であると考えられる。

３.１.２　形容詞＋名詞の抽出

Stanza での構文解析結果を用い amod の係り受け関係で結ばれている主辞と依存部を「形容詞＋名詞」のペアとして抽出した。名詞は upos が NOUN のものに限定し、代名詞や固有名詞は含んでいない。conj のタグを用い 1 つの名詞に複数の形容詞が並列されている場合も考慮している。データセットでは amod のシートにこの共起フレームのデータを収録している。

(5) A **fate** **worse** than a cookbook. （W commerce; CBU）（amod; bad + fate）

(6) The **thin**, **pale** **sunshine** gave **little** **warmth**, but it was a **welcome** **extension** of the autumn, and unusually fine for the **last** **day** of October.（W fict prose; GVT）（amod; thin + sunshine）（amod; pale + sunshine）（amod; little + warmth）（amod; welcome + extention）（amod; last + day）

(7) It wasn't simply because he was ugly — he knew **men** equally **ugly** and ten times more **lovable**. (W fict prose; GVT)（amod; ugly + man）（conj; ugly + lovable）

(8) **Individual**, **social** and **historical** **contexts** are also key.（W misc; G34）（amod; individual + context）（conj; social + context）（conj; historical + context）

（5）の例のように形容詞が後ろから前の名詞を修飾する場合も amod の係り受け関係で抽出可能である。（6）の thin, pale のように 2 つの名詞が連続して 1 つの名詞を修飾している場合も amod の係り受け関係で両方抽出可能である。（7）の例では amod と conj のタグを用いることで ugly と lovable を men に対する形容詞修飾として抽出に成功している。(8) のように 1 つの名詞に対して 3 つの形容詞が修飾している場合も 3 つも抽出することができる。

３.１.３　名詞＋名詞の抽出

Stanza での構文解析結果を用い compound の係り受け関係で結ばれている主辞と依存部を抽出した。主辞・依存部ともに upos で NOUN のタグが付いているものに限定している。これによって下記の（9）のように名詞＋名詞で 1 つの表現を成しているペアを抽出可能である。データセットでは nounmod のシートにこのデータを収録している。

(9) One arises out of the need for **state provision** of education to be efficient and economical, which means that **policy considerations** might outweigh individual parental preference.（W ac polit law edu; AN5）（compound; state + provision）（compound; policy consideration）

３.１.４　副詞＋動詞の抽出

Stanza での構文解析結果を用い advmod の係り受け関係で結ばれている主辞とその依存部を抽出した。advmod のタグは副詞が動詞を修飾する場合だけでなく形容詞や他の副詞を修飾する場合も含むため、主辞の upos タグが VERB のものに限っている。依存部の語の品詞は where などのような疑問副詞が入らないように、xpos のタグが RB, RBR, RBS のものに限っている。conj の依存関係を用いて 1 つの動詞に対して複数の副詞が修飾する場合も考慮している。データセットでは advmod_verb のシートにこの共起フレームのデータを収録している。

(10) Saturday, 13th: **Arrived** at Delhi airport over two hours **late**, at 06.30.（W misc; KAN）（advmod; late + arrive）

(11) If she opts for cardiotocography it must be **done properly** and not **haphazardly**.（W ac medicine; EA2）（advmod; do + properly）（conj; do + haphazardly）

３.１.５　副詞＋形容詞の抽出

Stanza での構文解析結果を用い advmod の係り受け関係で結ばれている主辞のうち upos が ADJ のものと、その依存部を副詞＋形容詞の共起ペアとして抽出した。副詞＋動詞の場合と同様に副詞の品詞は RB, RBR, RBS に絞っている。conj の係り受け関係を利用して 1 つの形容詞に対して複数の副詞が修飾する場合も考慮している。データセットでは advmod_adj のシートにこの共起フレームのデータを収録している。

(12) The action of the sea, however, is concentrated within **comparatively narrow** limits.（W ac nat science; GV0）（advmod; comparatively narrow）

(13) **Initially** the field-worker's relations with respondents in the field were **especially** and **unusually warm**, which runs counter to the norm in ethnographic research.（W ac polit law edu; A5Y）（advmod; initially + warm）（advmod; especially + warm）（conj; unusually + warm）

（13）の例では initially と warm を副詞＋形容詞のペアとして抽出している。それと同時に especially と unusually も warm を修飾する副詞として抽出することに成功している。

３.２　共起統計指標

コーパスを用いた統計的アプローチでのコロケーション抽出では広く共起統計指標（association measures; AM）が用いられる[5]。これまで様々な AM が提案されており、どの AM がコロケーション抽出に適しているか研究がなされている（e.g. Evert, 2004）。教育目的のコロケーションリスト作成においても、コロケーション候補抽出の際に AM を用いて基準設定をすることが多い。例えば、Ackermann & Chen（2013）はアカデミック英語用（EAP）のコロケーションリスト作成において、MI と t-score を用いて基準を設定し、一定の値を超えたものをリストに入れる候補にしている。

AM にはそれぞれ特徴があるため目的に応じて使い分けるべきである（Brezina, 2018）。例えば MI や z-score は低頻度語を構成素に含む共起ペアを評価しやすい一方で、t-score は高頻度語を構成素に含む共起ペアに対して高い値を返す（Evert, 2008; 石川, 2008; Brezina, 2018）。本研究ではデータセットの利用者がその目的に応じて必要な AM を選択できるように全ての共起ペアに対して MI, MI2, MI3, t-score, z-score, logDice, log-likelihood, chi-squared の 8 つの AM を算出し情報を付した[6]。

AM の算出には共起頻度、各構成素の頻度、コーパス総語数が必要である。本研究での共起語抽出は統語依存関係による共起語抽出[7]であるため、Evert（2008）に習い、各構成素の頻度はコーパス全体における当該語の総頻度ではなく、共起フレームごとに見て、その共起フレーム内に出現する頻度にしている。データセットでは、w1_in_rel と w2_in_rel の列が各共起フレーム内での構成素頻度を表している。コーパス総語数は、統語依存関係による共起語抽出では共起フレームの出現回数である。表 1 は共起フレーム別の出現総頻度である。

表 1　共起フレーム別総頻度

共起フレーム	総頻度
動詞＋目的語　（obj）	4,633,147
形容詞＋名詞　（amod）	5,188,122
名詞＋名詞　（nounmod）	2,295,774
副詞＋動詞　（advmod_verb）	2,948,140
副詞＋形容詞　（advmod_adj）	890,696

３.３　散布度指標

教育的に有用な語句を抽出するにあたっては頻度のみではなく散布度（dispersion）も考慮することが重要である。これは単語だけでなくコロケーションにも同様に言えることである（Rogersa, , Brizzarda, et. al., 2015）。共起統計指標は全て頻度に基づいて算出したものであるため、頻度とは別の軸で教育有用度を測る 1 つの指標として散布度の指標を用いることは重要である。散布度の指標を用いることで、限定的な場面でしか使用されない表現ではなく、様々な場面で使用されるコロケーションを選択することができる。しかしながら、著者の知る限りでは散布度指標を明示したコロケーション・データセットはほとんどない。従っ

[5] association measures はより一般的に「相関度指標」「関連度指標」とも訳される。
[6] 各指標の詳しい計算方法については Evert（2008）を参照。logDice については Rychlý（2008）または、恒川（2020）を参照。
[7] Evert（2008）ではこれを’syntactic cooccurrence’と呼んでいる。

て、本研究のデータセット整備では、教育的に有用なコロケーションを選択するために重要な情報として、散布度指標を算出し明示した。

本研究では散布度指標として DP（Deviation of Proportions）を用いた[8]。DP はコーパス内のサブコーパスやファイルごとの語数を考慮に入れた上で、各サブコーパスやファイル間で、特定の語句がどれだけ均等に用いられているかを数値化した指標である。0 から 1 の値を取り、0 に近いほど均等に用いられていることを示す。BNC は各ファイルや各ジャンルの語数が均等ではないことと、膨大なデータを処理しなければならないことを考慮し、算出が比較的簡素である DP を本研究では取り入れた。

DP を算出するにあたっては、コーパス分割の基準を定める必要があるが、今回は BNC のジャンル分けに従った。British National Corpus XML edition では classCode のタグ内に各ファイルのジャンルが示されている。タグを参照しジャンルごとに処理することで、各ジャンルの総語数と、各ジャンルにおける共起ペアの共起頻度データを取り出し DP を算出した。

３.４　CEFR-J Wordlist の利用

今回のコロケーション・リストを作成する際の学習レベル情報は、ヨーロッパ言語共通参照枠（CEFR）を日本の英語教育に適用した CEFR-J プロジェクトで構築された CEFR レベル別語彙表 CEFR-J Wordlist Version 1.6[9]を用いた（投野, 2013）。これはアジア圏の英語教科書分析および海外の CEFR 語彙表データを比較して学習語彙表として整備されたもので、A1 から B2 まで 7801 語（品詞別の語数、表層形では 6868 語）を選定している。CEFR-J Wordlist を用いることで、共起ペア内の単語の CEFR レベルを明示することができ、有益なコロケーションが 7801 語の語彙表内の単語でどの程度カバーできるか、語彙表の単語を身につけることの生産性と、コロケーションによる活用度・有用度の向上を目指した。

４．データセットの概要

本節では今回整備したコロケーション・データセットの概要を説明する。本データセットの中に含まれる情報は、図 1 の列名で示されるように各共起ペアの構成素（w1, w2）、構成素の CEFR-J レベル（w1_CEFR, w2_CEFR）、共起フレーム（relation）、共起頻度（cooccurrence）、共起フレーム内での各構成素の頻度（w1_in_rel, w2_in_rel）、散布度指標（DP）、期待値頻度（expected_freq）、共起統計指標（MI 以降）である。全部で 5 通りの共起フレーム（amod, obj, nounmod, advmod_verb, advmod_adj）を各シートに分けている。扱いやすいデータサイズにするために、収録しているコロケーションデータは共起頻度 5 以上かつ、共起頻度が期待値頻度以上（cooccurrence ≧ expected_freq）のものに限定している[10]。本データセットに収録されている共起ペア総数は表 2 の通りである。

[8] DP についての詳しい説明や計算方法は Gries（2008）を参照。

[9] http://www.cefr-j.org/download.html#cefrj_wordlist

[10] これに限定しない全データを研究用に公開する。付録参照。

表２　共起フレーム別 共起ペア数

共起フレーム	共起ペア数
動詞＋目的語（obj）	114,582
形容詞＋名詞（amod）	135,940
名詞＋名詞（nounmod）	72,340
副詞＋動詞（advmod_verb）	43,992
副詞＋形容詞（advmod_adj）	16,180
合計	**383,034**

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	w1	w2	w1_CEFR	w2_CEFR	relation	cooccurrence	freq_w1	freq_w2	w1_in_rel	w2_in_rel	DP	expected_freq	MI
2	last	year	A2	A1	amod	14280	78446	176485	59008	52433	0.388692426	596.355765	4.58167892
3	first	time	A1	A1	amod	8622	124338	189496	78511	56321	0.205860426	852.2964632	3.338595312
4	last	night	A2	A1	amod	8575	78446	38745	59008	13826	0.608845761	157.2523946	5.76898278
5	same	time	A1	A1	amod	7641	61855	189496	48995	56321	0.162705428	531.8778924	3.844594483
6	local	authority	A2	B1	amod	7152	47279	28451	43662	14385	0.525139906	121.0607364	5.884543799
7	long	term	A1	B1	amod	6467	82846	46578	38307	19778	0.329249197	146.0327737	5.46873252
8	next	year	A2	A1	amod	6455	45646	176485	35459	52433	0.352138152	358.3612234	4.170930652
9	other	hand	A1	A1	amod	5578	182918	59458	128190	14167	0.279434234	350.0433741	3.994142331
10	last	week	A2	A1	amod	5503	78446	48163	59008	14998	0.513578273	170.5823387	5.011678139
11	many	people	A1	A1	amod	4966	89532	117876	64619	40160	0.247660507	500.2000801	3.311507068
12	long	time	A1	A1	amod	4652	82846	189496	38307	56321	0.244187168	415.8515445	3.483710601
13	other	people	A1	A1	amod	4535	182918	117876	128190	40160	0.200474189	992.2878452	2.192271964
14	young	man	A1	A1	amod	4156	35531	93946	26878	31477	0.373086437	163.0722651	4.671612316
15	great	deal	A1	A2	amod	4061	66218	28841	52014	7528	0.18531049	75.47266468	5.749737005
16	nineteenth	century	–	A2	amod	3959	4207	27308	4058	24393	0.472953046	19.07950391	7.696968597
17	young	people	A1	A1	amod	3762	35531	117876	26878	40160	0.369457095	208.0561097	4.176455289
18	many	year	A1	A1	amod	3614	89532	176485	64619	52433	0.219059021	653.0625199	2.468303489
19	high	level	A1	A2	amod	3592	64571	41241	44473	18363	0.345744296	157.4091162	4.512196349
20	past	year	B1	A1	amod	3521	26267	176485	9267	52433	0.364867557	93.65558693	5.232476357
21	large	number	A1	A1	amod	3500	49196	61044	42006	14692	0.345426329	118.9548264	4.878869209
22	few	year	A2	A1	amod	3453	46810	176485	38943	52433	0.198886735	393.5717624	3.133151714
23	other	word	A1	A1	amod	3411	182918	42876	128190	12459	0.35514389	307.8414906	3.46993516
24	old	man	A1	A1	amod	3333	69388	93946	48501	31477	0.411698942	294.2617728	3.501649277
25	next	week	A2	A1	amod	3328	45646	48163	35459	14998	0.488149559	102.5060864	5.020873955

図１ コロケーション・データセットのサンプル画像

５．データセットを用いた調査例

本研究で整備したデータセットは、学習者や教師が学ぶべきコロケーションを選択するための情報として役立てることを想定しているが、研究目的も含め、その目的に応じて様々な利用が可能であろう。本節では学習に有用なコロケーション選択のために使用することを念頭に置いたうえで、教育的に有用なコロケーションを抽出するために本データセットをどのように用いるのがよいか考えるための、簡易的な調査を行う。

コロケーションを選択するためには共起頻度や AM や DP の値を用いてソートした上で、その中から使用者が必要であると思われるものを選択するのが基本であるが、その際にどの AM を使用するのか、適切な選択が必要である。また、DP を用いた適切な基準値で境界線を設けることで、共起頻度や AM の値は高いが、特定のジャンルに偏っているものを除外することができる。つまり適切な AM の選択と DP の値の基準値を考える必要がある。従って以下の 2 点を調査する。

1. 本データセットを用いて教育的に有用なコロケーションを抽出するにはどの AM を用いるのが最適か。

2. 本データセットを用いた教育的に有用なコロケーションの抽出に当たって、基準値となる適切な DP の値はいくつか。

５.１　ゴールドスタンダード

コロケーション抽出の評価にはゴールドスタンダードを使った手法がよく用いられる（e.g. Evert, 2008; Bhalla & Klimcikova, 2019; Evert, Uhrig et al., 2017）。本研究は CEFR-J Wordlist 内の語の組み合わせでできるコロケーションを教育的に有用なコロケーションと定義する。今回は Oxford Collocations Dictionary を用いて（McIntosh, Francis & Poole 2009）、この辞書に記載があり、かつ CEFR-J Wordlist に掲載されている語同士の組合わせのコロケーションをゴールドスタンダードと仮定する。

今回はデータセットの形容詞修飾（amod）と名詞修飾（nounmod）を対象とする。データセット中の amod と nounmod の共起ペア総数は 205,310 である。Oxford Collocations Dictionary では名詞＋名詞と形容詞＋名詞を区別していないため、この二つの共起フレームを同時に扱う。具体的には Oxford Collocations Dictionary の名詞を見出し語とするエントリーの中の ADJECTIVE に分類されている共起ペアを対象とする。これに該当するコロケーションを Oxford Collocations Dictionary から全て取り出したうえで、構成単語の片方もしくは両方が CEFR-J Wordlist にないものを削除し、残ったものをゴールドスタンダードとした。ゴールドスタンダードとなったコロケーション数は 38,362 であった。

５.２　AM と DP の評価

上記の手順で作成したゴールドスタンダードと作成したデータセットを用いて、n-best リスト（共起頻度や特定の AM の値順に並べ一定の数で区切ったもの）中の適合率（presicion; n-best リスト中のゴールドスタンダードの割合）、再現率（recall; n-best リスト中のゴールドスタンダードの数の、全ゴールドスタンダードに占める割合）を算出した。AM の中から共起頻度（cooccurrence）、MI、MI3、t-score、z-score、logDice、log-likelihood をそれぞれ用い、precision と recall を算出した。precision と recall のグラフを DP による基準値別に表したものが図 2-1 から図 2-6 である。

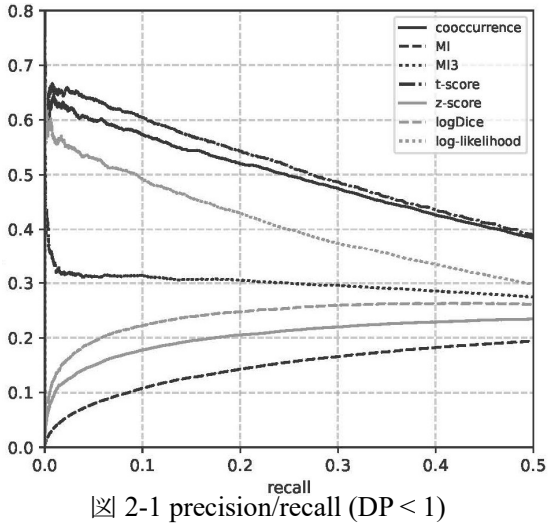

図 2-1 precision/recall (DP < 1)

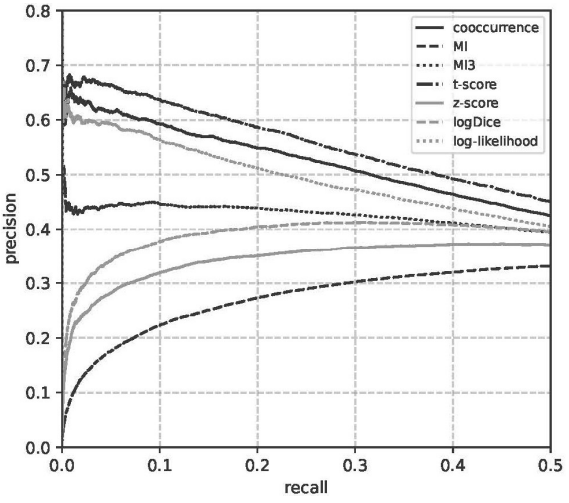

図 2-2 precision/recall (DP < 0.8)

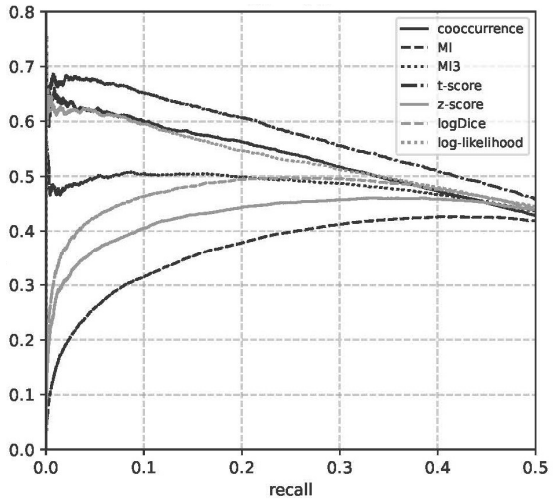

図 2-3 precision/recall (DP < 0.7)

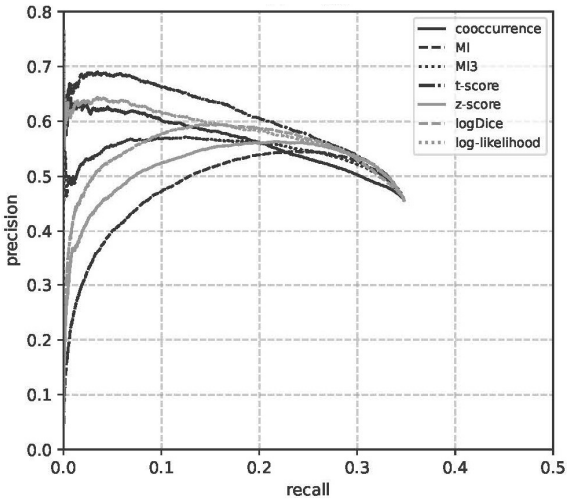

図 2-4 precision/recall (DP < 0.6)

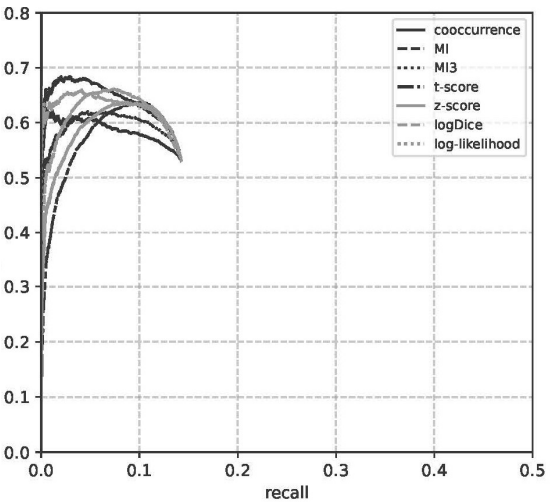

図 2-5 precision/recall (DP < 0.5)

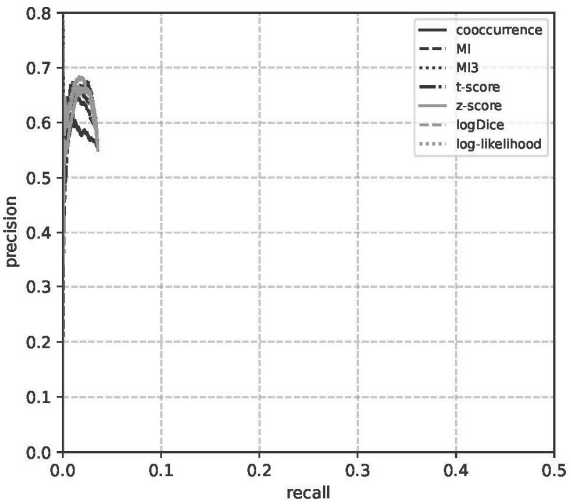

図 2-6 precision/recall (DP < 0.4)

図 2-1 から図 2-6 を見ると、常に t-score がよい成績であることが分かる。従って教育的に有用なコロケーションを選択するという点では t-score を用いるのが良いであろうということがグラフから推察される。DP < 0.6、DP < 0.5、DP < 0.4 のグラフは途中で切れているが、これは DP の値を低く設定すると、多くのコロケーションが除外され、一定以上 recall が上がらないことを意味する。例えば、DP を 0.4 以下に設定すると、2,487 の共起ペアしか残らないため、これ以上 n-best リストを長くすることはできない。この中に含まれるゴールドスタンダードの数は 1,366 であるため、DP < 0.4 の場合の recall は約 3.6%を超えることがない。従って、DP < 0.4 や DP < 0.5 は学ぶべきコロケーションのうちの多くを除外することになる。

グラフだけでは具体的な数値で AM・DP の成績を評価することができないため、precision と recall の観点から AM・DP の成績を数値化して評価したものが表 3 である。表 3 の AP30 は recall = 0.3 までの precision の平均値である。Evert, Uhrig et al.（2017）は、recall = 0.5 ま

での precision の平均値である AP50 によって各種のコロケーション抽出法の評価を行っているが、教育目的でのコロケーション抽出では数を絞る必要があることを勘案し、recall = 0.3 に設定した。DP の値を 0.5 未満に設定すると recall が 0.3 には達さないため AP30 は DP < 1.0、DP < 0.8、DP < 0.7、DP < 0.6 の 4 つの基準値別に算出した。

表 3　DP の基準値別 AP30

DP	AM	AP30
DP < 1.0	共起頻度(cooccurrence)	54.2%
	MI	9.7%
	MI3	30.9%
	t-score	56.5%
	z-score	17.3%
	logDice	21.6%
	log-likelihood	45.2%
DP < 0.8	共起頻度(cooccurrence)	56.7%
	MI	21.0%
	MI3	43.9%
	t-score	60.5%
	z-score	31.3%
	logDice	36.8%
	log-likelihood	53.3%
DP < 0.7	共起頻度(cooccurrence)	57.6%
	MI	30.4%
	MI3	49.7%
	t-score	62.1%
	z-score	39.9%
	logDice	45.3%
	log-likelihood	56.7%
DP < 0.6	共起頻度(cooccurrence)	57.2%
	MI	45.8%
	MI3	55.1%
	t-score	62.0%
	z-score	51.1%
	logDice	54.9%
	log-likelihood	59.3%

DP の基準値別に見てみると、0.7 に基準を設定するのが適切であると思われる。AP30 で最も高い値が出るのは DP < 0.7 の t-score の場合である。DP < 0.6 にすると、わずかであるが t-score の AP30 の値が下がる。他の AM は、DP の値を低く設定するほど成績が良くなっていることが分かるが、t-score の値を超えることはない。

以上のことを踏まえると基本的には DP < 0.7 辺りに設定するのが妥当であると考えられ

るが、実際には対象となる学習者のレベルに応じて、選択するコロケーションの数は絞ることになる。その場合 DP < 0.6 や DP < 0.5 に設定することも考えられるが、t-score を使用する場合は、DP を低く設定し必要なものを除外してしまうことを考えると、やはり DP < 0.7 あたりが妥当な目安になるであろう。

もちろん対象となる学習者やデータセットの使用の目的によって、AM や DP やその他の情報の使用法は様々である。対象となる学習者レベルが高い場合は、MI や logDice の値を用いることも十分考えられる。これらの評価例を参考に、目的に応じて有用なコロケーションを CEFR レベルを勘案して選定できるのがこのデータベースの新規性であるといえよう。

６． おわりに

本研究では、CEFR 準拠英語学習語彙表と多様な統計指標を活用して、学習すべきコロケーション特定のためのコロケーション・データセット作成の報告を行い、データセット内での DP と AM の評価を行った。データセットの使用及び DP・AM の教育的な観点での評価は暫定的なものであり、対象となる学習者レベルを考慮した調査を行い、構成素のレベルも踏まえたうえで、どのようなコロケーションが抽出されるのかより子細に調査する必要があるが、一応の目安として t-score を用い、DP＜0.7 に設定すると、CEFR の A1～B2 レベルの語彙で構成された学習レベル的に有益なコロケーションが選定可能になると考えられる。

付　録

本研究で作成したデータセットとプログラムコードは以下のリンクから利用可能である。全データについては、共起フレーム別に CSV 形式で保存してあるものと、Python プログラム上で呼び出し可能な pandas DataFrame の形で保存してある。公開しているデータは本稿を適切に引用することで研究教育目的にのみ無償で利用可能である。
(https://drive.google.com/drive/folders/1CxLqPpAL9UGTYd234siJgCOJNGv85Duw?usp=sharing)

参考文献

Kirsten Ackermann and Yu-Hua Chen (2013). “Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach”, *Journal of English for Academic Purposes*, 12:4, pp.235-247.

Sabine Bartsch and Stefan Evert (2014). “Towards a Firthian notion of collocation”, In A. Abel and L. Lemnitzer (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, number 2/2014 in OPAL - Online publizierte Arbeiten zur Linguistik, pp.48-61. Institut für Deutsche Sprache, Mannheim. (https://www.stephanie-evert.de/PUB/BartschEvert2014.pdf よりダウンロード可能)

Vishal Bhalla and Klara Klimcikova (2019). “Evaluation of automatic collocation extraction methods for language learning”, In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp.264-274, Florence, Italy. Association for Computational Linguistics. (https://aclanthology.org/W19-4428/よりダウンロード可能)

BNC Consortium (2007). *British National Corpus, XML edition*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554.

Vaclav Brezina (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Stefan Evert (2004). “The Statistics of Word Cooccurrences: Word Pairs and Collocations”, Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. (https://www.stephanie-evert.de/PUB/Evert2004phd.pdf よりダウンロード可能)

Stefan Evert (2008). “Corpora and collocations”, In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58, pp.1212-1248. Mouton de Gruyter, Berlin. (http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf よりダウンロード可能)

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl (2017). “E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification”, In *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference*, pp.531‑549. (https://www.stephanie-evert.de/PUB/EvertUhrigEtc2017.pdf よりダウンロード可能)

Dana Gablasova, Vaclav Brezina, and Tony McEnery (2017). “Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence”, *Language Learning*, 67:S1, pp.155-79. (https://onlinelibrary.wiley.com/doi/full/10.1111/lang.12225 よりダウンロード可能)

Sylviane Granger and Magali Paquot (2008). “Disentangling the phraseological web”, In Sylvia Granger and Fanny Meunier (eds.) *Phraseology: An Interdisciplinary Perspectives*, pp.27-49, Benjamins.

Stefan Th Gries (2008). “Dispersions and adjusted frequencies in corpora”, *International Journal of Corpus Linguistics*, 13:4, pp.403-437. (https://stgries.info/research/2008_STG_Dispersion_IJCL.pdf よりダウンロード可能)

石川慎一郎 (2008). 『英語コーパスと言語教育 ―データとしてのテクスト』大修館書店.

Adam Kilgarriff, Vít Baisa, Jan Bušta et al. (2014). “The Sketch Engine: ten years on”, *Lexicography ASIALE*, 1, pp.7-36. (https://link.springer.com/article/10.1007/s40607-014-0009-9 よりダウンロード可能)

Ron Martinez and Norbert Schmitt (2012). A Phrasal Expressions List. *Applied Linguistics*, 33:3, pp.299-320.

Ron Martinez and Victoria A. Murphy (2011). Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. *TESOL Quarterly*, 45:2, pp.267-290.

Michael McCarthy and Felicity O'Dell (2017) *English Collocations in Use Intermediate* [second edition] Cambridge University Press.

Colin McIntosh, Ben Francis and Richard Poole (2009). *Oxford collocations dictionary: for students of English (Second)*, Oxford University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman (2021). "Universal Dependencies", *Computational Linguistics*, 47:2, pp.255-308. (https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies よりダウンロード可能

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp.101-108, Association for Computational Linguistics. (https://aclanthology.org/2020.acl-demos.14/よりダウンロード可能)

James Rogersa, Chris Brizzarda, Frank Daultonb et al. (2015). "On using corpus frequency, dispersion, and chronological data to help identify useful collocations", *Vocabulary Learning and Instruction*, 4:2, pp.21-37.

Pavel Rychlý (2008) A Lexicographer-Friendly Association Score, In Proceedings of Recent Advances in *Slavonic Natural Language Processing*, RASLAN, pp.6-9. (https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf.よりダウンロード可能)

Rita Simpson-Vlach and Nick C. Ellis (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31:4, pp.487-512.

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

投野由紀夫 (編) (2013). 『CAN‐DO リスト作成・活用 英語到達度指標 CEFR‐J ガイドブック』大修館書店.

恒川元 (2020). 「logDice 係数はどのような共起指標か」『言語文化論究』45, pp.35-44. (https://catalog.lib.kyushu-u.ac.jp/opac_detail_md/?lang=0&amode=MD100000&bibid=4104141 よりダウンロード可能)