日本語話し言葉における定式表現の分析
―CSJ・CEJC・J-TOCC の比較を中心に―

蘇 振軍（江蘇大学・大阪大学）†

Formulaic Language in Spoken Japanese: A Comparative Study of CSJ, CEJC, and J-TOCC

Zhenjun SU (Jiangsu University・Osaka University)

要旨

本研究は，日本語話し言葉の定式表現の文体差を明らかにするため，CEJC（雑談）・J-TOCC（話題別）・CSJ（学会講演）各 100 万語から 2〜7 語の上位 1,000 を抽出し，表現文型・文法コロケーション・「実質語 2」・「実質語 1：その他」・「実質語 1：完全句」・非定式表現に分類して分析した。その結果，語列が長いほど定式表現が増え，とりわけ固定的定式表現は J-TOCC で 5 語 77.20%・6 語 78.64%，CSJ で 5 語 61.90%・6 語 73.30%・7 語 88.00%，CEJC でも 4 語 60.90%に達した。文体差は明らかであり，CEJC は相槌の繰り返しと「実質語 2」からなる語彙コロケーション，J-TOCC は 3〜5 語で文法コロケーション，CSJ は 4 語以降に「名詞 - 動詞」，「動詞 - 名詞」の内容志向連鎖が多かった。完全句は CEJC に偏って出ているのに対し，CSJ ではほぼ皆無である。

１．はじめに

近年，複数の語が組み合わさって 1 つのまとまった意味や機能を表す定式表現（formulaic language）[1]に関する研究が，コーパス言語学や言語心理学を中心に盛んに行われている（e.g., Wei & Zhong, 2023; 蘇, 2019）。とりわけ，英語を対象とした研究では，定式表現の役割に関して多様な検討が進んでいる（e.g., Nattinger & DeCarrico, 1992; Wray, 2002, 2008, 2009）。コーパス言語学においては，大規模コーパスから高頻度の定式表現を抽出し，その使用頻度や文法的・機能的特徴を詳細に分析する研究が多くなされている。その結果，英語では表現のうち 59%から 80%が定式表現に該当するとされ（畑佐, 2022: 92），定式表現が文の構成において重要な役割を果たすことが明らかになった（e.g., Biber et al., 1998; Vela-Rodrigo, 2023）。これらは，コーパスから言語的事実や法則性を導き出そうとするコーパス駆動型研究に位置づけられ，定式表現は一般に語彙束（lexical bundles）[2]と呼ばれる。語彙束については，使用頻度や割合，文法構造上の特徴，談話構築常の機能に関して多くの知見が得られている

† sosingun<@>yahoo.co.jp / 1000005091<@>ujs.edu.cn

[1] 定式表現に関する名称は prefabs, fixed・semi-fixed expression, phraseology, lexical bundles, formulaic sequences など，60 ほど存在する（Wray, 2002, 2008），本稿では全て定式表現と呼ぶ。

[2] Biber et al.（1998）は，コーパスにおける高頻度の語連鎖を指す用語として，初めて「語彙束（lexical bundles）」を導入した。先行研究を踏まえると，語彙束には表現文型に相当する機能語連鎖や実質語どうしのコロケーションなどが含まれるため，本研究ではこれを包括的に「定式表現」と呼ぶ。

（e.g., Altenberg, 1998; 蘇, 2019, 2024; Vela-Rodrigo, 2023）。

以上のように，定式表現研究は英語を対象として発展しており，言語研究・言語習得における重要性が広く認識されている。他方，日本語を対象とする定式表現研究では，慣用句やコロケーションなどの複数の語の結びつきに対する記述的アプローチや，特定の語彙や文法項目を対象とするコーパス基盤アプローチが中心であり（蘇, 2019），英語研究に見られるようなコーパス内での使用頻度，種類，割合，機能に関する研究は十分とは言いがたい。

そこで，本研究では，日本語話し言葉コーパスにおいて定式表現がどの程度使用されているか，また文体間にどのような差異が見られるかを明らかにすることを目的とする。これにより，日本語話し言葉における定式表現の文体差に関する理解が深まることを期待する。

２．先行研究と本研究の課題

２.１　先行研究の概観

定式表現の研究は，1990 年代以降コーパス言語学の発展とともに注目を集めてきた。特に英語研究においては，Sinclair（1991），Biber et al.（1998），Wray（2002, 2008）などが理論的基盤を構築し，その後，多くの実証研究が行われている。

そのうち，文体差に焦点を当てた研究としては，Biber らによる一連の成果が挙げられる（e.g., Biber et al., 1998; Biber et al., 2004; Biber & Barbieri, 2007; Biber, 2009）。Biber et al.（1998）は，日常会話とアカデミックライティングという 2 つの文体を比較し，会話では定式表現の使用数・種類が多く，句や従属節が中心となる一方，アカデミックライティングでは名詞句や前置詞句が多いことを示した。これを発展させた Biber et al.（2004）は，話し言葉と書き言葉の特徴を併せ持つ大学の授業，および教科書を加えて分析した。その結果，授業では名詞句と動詞句が共存するハイブリッド性と比較的単純な句構造の多用，教科書では複雑な名詞句型と参照表現が多かったことを報告した。さらに，Biber & Barbieri（2007）は大学場面をより広く対象に，定式表現の頻度と機能を検討し，話し言葉の方が書き言葉より定式表現が多く，教室管理や受付場面に集中しやすいこと，話し言葉では立場表明表現（例 : I don’t know）が多用されることを示した。Biber（2009）は，会話における「連続固定序列」（例 : I don’t know）と，論文における「スロットを持つ文型」（例 : the _er the _er）が多く，さらに機能としては前者が立場表明，後者が参照に集中することを明らかにしている。

また，ジャンルや分野の差異に着目した研究も重視されてきた。Hyland（2008）は，電気工学・生物学・経営学・応用言語学の研究論文・博士論文・修士論文から成る約 340 万語コーパスを用い，分野・ジャンル別に 4 語を比較した。その結果，共通の定式表現はほとんど見られず，各分野特有の表現が多数あることがわかった。Durrant（2017）は事前に分野区分を仮定せず分野間の差異を検討することを目的として，24 分野で 1,558 ほどのレポートを対象に 4 語を焦点に当てて分析した。その結果，レポートが人文社会・理工・生命科学・商学の 4 つに大別されることを報告した。他に，Ren（2021）は応用言語学と薬科学の研究記事を対象に 4 語を比較した結果，後者がより文型に依存することを報告した。これらは，レジスター差を越えて「学問分野」が定式表現の選択に影響することを示し，英語指導における分野別アプローチの必要性があると考えられる。

一方，日本語に関しては，杉浦（2001），杉浦・朴（2002），李・長谷部（2017），蘇（2024）などの先行研究があるものの，話し言葉を対象とする定式表現の包括的分析は不足している。特に，異なる文体間における定式表現の頻度・構造・機能の比較はほとんど行われておらず，日本語話し言葉における文体差を体系的に検討する必要があると言えよう。

２.２　本研究の課題

以上を踏まえ，本研究では，日本語話し言葉における定式表現の文体差を検討することを目的とし，異なる性格を持つ 3 つのコーパスを対象とする。具体的には，話題別の『日本語話題別会話コーパス』（Japanese Topic-Oriented Conversation Corpus，以下 J-TOCC）（中俣，2021），場面別の日常生活の『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation，以下 CEJC）（小磯他，2022），アカデミックスピーキングの『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese，以下 CSJ）（国立国語研究所，2006）である。これらのコーパスを対象に，以下の課題を設ける。

課題 1：日本語話し言葉において，3 つの文体に共通する定式表現はどの程度の割合で用いられ，どのような種類が見られるのか。

課題 2：各コーパスにおいて特徴的な定式表現は何か，その種類と割合はどのように異なるのか。

３．コーパスと方法

３.１　データ

上述したように，本研究では，J-TOCC は全話題，CEJC は「雑談」のみ，CSJ は「学会講演」のみを対象とし，それぞれから無作為に 100 万語を抽出した（表 1）。表 1 のトークンはテキスト中の表記形の総語数を指す。タイプは，同一表現が複数回出現しても 1 語として数え，全体に含まれる異なり語の数である。G.I.（Guiraud index）は Root TTR とも呼ばれ，タイプ数をトークン数の平方根で割って算出する指標である。トークンの影響を比較的受けにくく，値が大きいほど表現の種類が豊富であることを示す（中俣，2015）。

表 1：各コーパスの語数情報

	トークン	タイプ	G.I.
CEJC	1,000,000	22,967	22.97
J-TOCC	1,000,000	18,780	18.78
CSJ	1,000,000	17,302	17.30

３.２　*n*-gram の抽出手順

先行研究を参照に，*n*-gram モデルを用い，以下の手順で 2-gram から 7-gram を抽出した。*n*-gram モデルとは，語列長 n を決め，2-gram なら「are you」，3-gram なら「are you crazy」のように，連続する n 語を単位として一定の頻度基準で抽出する手法である。抽出・処理は PyCharm 2024.2.4（Professional Edition）を用い，Python3.11 で実行した。

ステップ 1. サブコーパスの作成。J-TOCC から全話題の txt データを無作為抽出して合併する。CEJC では「雑談」のデータのみを合併する。CSJ では分野（人文・工学）を問わず「学会講演」のデータを無作為抽出して合併する。

ステップ 2. *ノイズ処理。*3 サブコーパスに含まれる記号類，メタ情報など言語内容に関わらない情報を削除する。

ステップ 3. *n*-gram の抽出。3 サブコーパスから 2-gram から 7-gram を抽出する。頻度の閾値は Biber et al.（1998）に従い，2 から 4-gram は 100 万語あたり 10 回以上（PMW≧10），5 から 7-gram は 5 回以上（PMW≧5）とした。

３.３　定式表現の認定方法と認定手順

定式表現の分類については，蘇（2019）によると，コロケーション，表現文型，慣用句，ことわざといったものがある。それぞれの定義や認定基準は研究によって異なる。本研究では，次のように定める。

まず，コロケーションは，国広（2007）の「連語」項（「二つ以上の単語が文法的関連のもとで固定的に結びつくもの」）を概念的に踏襲する[3]。但し，本研究の *n*-gram 抽出は形態素に基づき，助詞等を含むため，操作的には「2 つ以上の実質語から成る連続語列」をコロケーションと定義する。さらに，本研究ではコロケーションを包括的範疇とし，慣用句・ことわざもその一部に位置付ける。

また，表現文型についても呼称・定義が諸説ある。森田・松木（1989）の「複合辞」（「いくつかの語が複合して，ひとまとまりで辞的機能を果たす表現」）を採用し，『現代語複合辞用例集』（国立国語研究所，2001）および「機能語用例データベース『はごろも』ver.4」（堀他，2016）を参照してコーディング用 Excel（参照用 Excel）を作成した。その結果，表現文型の参照項目は 421 項目となった。さらに，表現文型＋実質語 1 語から成る語列は，中俣（2014）に倣い文法コロケーションとする。

以上を踏まえ，2 から 7-gram に含まれる定式表現を次の手順で認定した。処理は PyCharm 2024.2.4 で実行し，各段階で得られた候補を目視確認した。

ステップ 1．表現文型・文法コロケーションの認定。参照用 Excel を用いて 2 から 7-gram とマッピングし，表現文型をコーディングする。次に，該当文型に実質語の有無を確認し，実質語を 1 語含み表現文型と結合する語列を「実質語 1：文法コロケーション」に分類する（例：想像に難くない）。参照リストの項目や，表現文型要素を含む語列（例：〜てもいいし／〜んじゃないか）は表現文型に分類する。

ステップ 2．コロケーションの認定。ステップ 1 に該当しない語列について実質語の数を判定し，実質語を 1 語含む語列を「実質語 1：その他」（例：がわかった/を目に），2 語以上を含む語列を「実質語 2」（例：話を聞い/心の中）に分類する。

ステップ 3．非定式表現の認定。ステップ 1・2 のいずれにも該当しない語列（実質語・表現文型を含まない語列），および「実質語 2」に含まれる無意味の語列の断片を非定式表現とする。

３.４　特有定式表現の抽出と統計検定

本研究の「各コーパスに特有の定式表現」は，当該コーパスとほかの 2 コーパス（結合）との相対頻度差が統計的に有意な *n*-gram を指す。検定には Dunning（1993）に基づく対数尤度比（LLR）検定を用い，有意水準は.001（LLR≧10.83，df=1）とした。検定対象は，３.２節におけるステップ 3 で設定した頻度閾値によって得られた *n*-gram とした。この枠組みは，日本語の産出差と定式表現研究に LLR を適用する先行研究（小西，2017；蘇，2024）も参照した。以後，「特有の定式表現」は，上記基準で LLR により有意と判定された語列をいう（抽出結果は４.２節で報告する）。

---

[3] 国広（2007）「連語」項目では，「二つ以上の単語が何らかの文法的な関連（主語＋述語、目的語＋動詞、修飾語＋被修飾語）の元に結びついて用いられるとき、意味的な要請もなく、はっきりとした理由もないのに結びつきが固定しているもの」と定義され，「風邪を引く」（風邪/を/引く）等が例示されている。「黄色い石」のような 2 語構造は 2-gram に多く含まれるため対象外とした。

４．結果と分析

４.１　3 つのコーパスに共通の定式表現

３.２節で述べた頻度閾値に基づき抽出した共通の *n*-gram の詳細を表 2 に示す。表 2 から分かるように，語列長が増すにつれて件数は急減する。PMW 閾値でフィルタした後は，高頻度語列への絞り込み効果が明らかである。しかし，2-gram のタイプ数は極めて多く，すべてを分析対象に含めるのはデータ量が膨大すぎるため，これからの分析では各語長につき頻度上位 1,000 項目（4 語以上は該当件数すべて）に範囲を限定し，これらに含まれる定式表現を検討する。定式表現の認定は３.３節の基準に従った。認定した結果は表 3 である。ただし，表 3 においうて「実質 1：その他」の括弧内の数字は，「そうですね」や「すいません」,「どうなの」の類に代表される完全句である。これらの表現は原則として定式表現に入るため，後述の分析では定式表現として扱う。

表 2：3 コーパスに共通する n-gram の詳細

gram	総数	PMW 閾値	PMW フィルタ後
2	14,750	10	1,631
3	10,014	10	542
4	3,273	10	70
5	691	5	28
6	126	5	1
7	12	5	0

表 3 が示すように。2-gram では,「実質語 1 : その他」が 678 と最も多く，非定式表現 118，表現文型 101，「実質語 2」68，文法コロケーション 26 が続く。このように，実質語 1 語＋機能語が広く分布し，固定度の高い結合は相対的に少ない。3-gram では，「実質語 1：その他」236，表現文型 143，文法コロケーション 76，「実質語 2」36，非定式表現 44 の順である。2 語と比べて表現文型と文法コロケーションが増加し，半固定・固定の枠組みが現れ始める。4-gram では，表現文型 25，文法コロケーション 24 がほぼ同じ，両者で 70.00%に達する。「実質語 1：その他」9，「実質語 2」6 がこれに続き，非定式表現は 5 まで減少する。5-gram では，文法コロケーション 15 が最多で，表現文型 9，「実質語 2」4 が続く。非定式表現と「実質語 1：その他」は 0 である。6-gram は文法コロケーションが 1 しかない。

表 3：3 コーパスに共通する構造型（トークン）

gram	実質語 0		実質語 1		実質語 2	合計
	非定式表現	表現文型	文法コロケーション	その他		
2	118	101	26	678(8)	68	1,000
3	44	143	76	236(7)	36	542
4	5	25	24	9(1)	6	70
5	0	9	15	0	4	28
6	0	0	1	0	0	1

以上のように，語列長が増加するにつれて表現文型と文法コロケーションの占める割合

が段階的に拡大し,4 語で両者が全体の 7 割,5 語では文法コロケーションが過半を占める。加えて,「実質語 1：その他」に含まれる完全句（2-gram＝8，3-gram＝7，4-gram＝1）は，短い語列に限って少数出現する。n の増大に伴い，半固定（表現文型）と固定（文法コロケーション）定式表現が多くなり，非定式的な連接は急速に減少する。

４.１.１　３コーパスに共通する「表現文型」の分析

本節では，表 4 に基づき,「表現文型」および「文法コロケーション」に焦点を絞って詳細な分析を行う。表 4 には，各 gram のトークン数，タイプ数，G.I.が示されている。

まず，トークン数を見ると，3-gram は 219，最も多い。次いで 2-gram（127）となっている。4-gram（49），5-gram（24），6-gram（1）と語列長の増加に伴って急減し，長い語列では限られた表現文型のみが上位に現れることがわかる。一方，タイプ数は 3-gram が最も多く（64），2-gram が 56，語列が短いほどタイプの多様性が高い。また，G.I.は 2-gram が最も高く（4.97），次いで 3-gram（4.32），4-gram（2.65），5-gram（1.63），6-gram（1.00）の順であり，語列が長くなるにつれてタイプの多様性が減少する傾向が見られた。つまり，短い語列ほど表現の種類が豊富であり，長くなるにつれて同一の文型が繰り返し使用されることが分かる。また，文法コロケーションを見ると，トークンは 3-gram（76）が最も多く，次いで 2-gram（26），4-gram（24）であり，5-gram 以上では急速に減少する。このことから，共通文型においては，3〜4 語で文法コロケーションの寄与が顕著になり，5 語以降は表現文型の大半を占めることが示唆される。このように，話し言葉においては短い語列では多様な表現が存在し，比較的高い自由度を持つが，長くなるにつれて特定の定式表現が繰り返し現れる傾向が顕著になることがわかった。文法コロケーションについても同様の傾向があった。

表 4：「表現文型」および「文法コロケーション」の詳細

gram	トークン	タイプ	G.I.	文法コロケーション（トークン）
2	127	56	4.97	26
3	219	64	4.32	76
4	49	18	2.65	24
5	24	8	1.63	15
6	1	1	1.00	1

４.１.２　３コーパスに共通する「実質語 2」の分析

表 5 に示すように，2-gram で最も高いのは「感動詞/感動詞」23.53%である（例：あーあー，えーえー，ああ，ええ）。相槌・反応・ためらいなど発話の立ち上げに用いられる表現が中心である。それから,「連体詞/名詞」11.76%（例：あの人，この・その辺，この人，こんな感じ）で，指示語＋名詞による話題の指示・同定が多い。「副詞/動詞」11.76%（例：こうやっ，そうする，ちょっと違う，どうする）が続き，評価・程度・方法を示す副詞が述部と結び付き，簡潔な行為記述を表す。3-gram では,「副詞/動詞」は 19.44%を占め（例：こうやって，どうして，どうなってる，どうやって），説明・問いかけの基本型が中心となる。次いで「名詞/動詞」16.67%（例：こともある，ものがある，感じになっ）が高く，名詞化＋述部による一般化・叙述が目立つ。さらに「形状詞/名詞」8.33%（例：みたいなもの，ようなこと・感じ）が一定割合を占め，比況・類似を用いて話題を柔らかく述べる。これらの結

果から，共通の「実質語 2」は，2 語では感動詞系列や指示表現が中心で応答・指示の機能を果たし，3 語では副詞＋述部＋助詞や名詞化＋述部が増えて行為・状況の記述へと移行することが分かる。つまり，語列は 1 語が増えることで，相互行為上の反応や話題提示から，命題内容の説明・一般化へと機能が拡張することが示唆される。

表 5：「実質語 2」における主な構造型および割合（%）[4]

gram	構造	割合	例
2	感動詞/感動詞	23.53	あーあー，えーえー，ああ，ええ
	連体詞/名詞	11.76	あの人，この/その・辺，この人，こんな感じ
	副詞/動詞	11.76	こうやっ，そうする，ちょっと違う，どうする
3	副詞/動詞/助詞	19.44	こうやって，どうして，どうなってる，どうやって
	名詞/動詞	16.67	こともある，ものがある，感じになっ
	形状詞/名詞	8.33	みたいなもの，ようなこと・感じ

４.２　3 コーパスにおけるそれぞれ特有の定式表現

４.２.１　「CEJC」に関する分析

本節では，CEJC における上位 1000 の特有定式表現を対象に，構造型を分析する（表 6）。分析にあたっては，5-gram には，「うんうんうんうんうん」「はいはいはいはいはい」など相槌の繰り返しが「実質語 2」に 306 項目も含まれ，残りはコロケーション 34 項目，完全句 127 項目に偏っていた。そのため，CEJC では 2-gram から 4-gram に限定した。

表 6：CEJC における構造型（トークン）

gram	実質語 0		実質語 1		実質語 2	合計
	非定式表現	表現文型	文法コロケーション	その他		
2	123	62	20	523(57)	215	1000
3	36	210	100	238(62)	354	1000
4	10	267	168	114(38)	403	1000

まず，2-gram においては，非定式表現（123）と「実質語 1：その他」（523）が合わせて 646 と多く占める。短い語列では，実質語を含まない語列や，実質語 1 語に機能語が付く語列が多いことを示す。一方，「実質語 2」が 215 見られ，複数実質語の結合も無視できない。表現文型（62），文法コロケーション（20）は相対的に少ない。これに対し，3-gram では，表現文型（210）と「実質語 2」（354）が大きく増加し，文法コロケーション（100）も増えた。非定式表現は 36 に減少し，「実質語 1：その他」 238（完全句 62）は 300 で，短い挨拶・応答などの完全句も一定量現れる。4-gram に至ると，表現文型が 267 と最も多く，「実質語 2」が 403 と高い割合を示している。文法コロケーションは 168 でこれに続く。非定式表現は 10，「実質語 1：その他」は 114 と大幅に減少する。文法的に固定化された表現文型

[4] 「構造」欄には実質語のみを示し，助詞・助動詞などの機能語は記載しない。「例」欄には，「/」は択一を表し，「この/その・辺」は「この辺」と「その辺」を意味する。「・」は結合位置の区切りであり，「A/B・C」は AC と BC（同様に「C・A/B」は CA と CB）を指す。

や複数実質語のコロケーションが優勢となる。さらに,「実質語 1：その他」における完全句を定式表現に算入すると，定式表現の割合は，2-gram35.40%，3-gram72.60%，4-gram87.60%となる。これらのことから，CEJC では，短い語列では自由度の高い語連接が多い一方，3 語と 4 語の範囲で表現文型・文法コロケーション・「実質語 2」が増え，定式表現が急速に集中することがわかる。

A. CEJC における「表現文型」の分析

表 7 は，CEJC の表現文型について，2〜4-gram のトークン数・タイプ数・G.I.，および内訳としての「文法コロケーション」トークン数を示す。トークン数を見ると，2-gram（82）から 4-gram（435）へと，語列長の増加に伴って明確に増加する。タイプ数は，2-gram は 32，3-gram は 42，4-gram は 39 となっている。G.I.は，2-gram が最も高い（3.53），4-gram では 1.87 まで減少する。これは，長い語列ほど少数の定式化された表現が繰り返され，タイプの多様性が失われることがうかがえる。最後に，文法コロケーションを見ると，2-gram は 20，3-gram は 95，4-gram は 168 まで増加する。CEJC では，3〜4 語の範囲で表現文型が量的に拡大するだけでなく，その中核をなす文法コロケーションの寄与が高まることが示唆される。これらにより，CEJC では短い語列では比較的多様な定式表現が現れる一方で，長い語列になるにつれて特定の文法コロケーションが支配的となり，定式性が一層強化されることが言えよう。

表 7：CEJC における「表現文型」および「文法コロケーション」の詳細

gram	トークン	タイプ	G.I.	文法コロケーション（トークン）
2	82	32	3.53	20
3	310	42	2.39	95
4	435	39	1.87	168

B. CEJC における「実質語 2」の分析

表 8 によれば，2-gram では「感動詞/感動詞」41.40%が最も高い。「ああ，あはい，いやいや」など，感動詞の反復や応答詞の連接が中心で，相手発話への即時反応として同意・否認・驚きを簡潔に示す用法が目立つ。次いで「感動詞/副詞」9.30%（例：あそ，あっそう）が続き，感情提示に程度・評価を添える型が一定数見られる。「感動詞/代名詞」8.84%（例：あこれ，あーあたし）は，指示語との結合によって対象の指示と気づきの提示を同時に行う点に特徴がある。3-gram では，「感動詞/感動詞」49.44%が引き続き最も高く，「いやいやいや，うんうんうん」のように反復の例が多かった。これは強い否認・驚き・積極的な相槌の表明に用いられる。「副詞/動詞」14.12%（例：こうやって，どうして）は説明・手順提示・問いかけの枠組みをつくる基本型である。さらに，「感動詞/副詞」7.91%（例：あそゆう，うんそうだ）が続き，了解・評価の提示を丁寧化する働きが観察される。4-gram では，「感動詞/感動詞」53.10%に達し，感動詞の反復が長い連接として定着している。一方，「感動詞/副詞」10.67%（例：あそうですか，うんうんそうそう）は，確認・応答を表す表現として一定割合を占める。このように，CEJC の「実質語 2」は各語長で感動詞を核とする結合が高い比率を占め，特に反復型は相槌・応答・驚き・否認を明示する機能を担う。これに対し，感動詞に副詞や代名詞が連接する型は，程度評価や対象指示を付与し，気づきの提示・話題共有の確認・発話導入を円滑にする働きを持つ。提示例の多くは発話冒頭や応答部に配されやすく，日常雑談における相互行為を滑らかに進める機能を果たす。

表 8：CEJC における「実質語 2」の主な構造型および割合（%）

gram	構造	割合	例
2	感動詞/感動詞	41.40	ああ，あはい，いやいや，はいありがとう
	感動詞/副詞	9.30	あそ，あっそう，うんまあ，はいどうぞ
	感動詞/代名詞	8.84	あこれ，あーあたし，うんそれ，はいこれ
3	感動詞/感動詞	49.44	あああ，いやいやいや，うんあそう，はいあはい
	副詞/動詞	14.12	ああゆうこと，よくわかん，こうやって
	感動詞/副詞	7.91	あそゆう，うんそうだ，はいそうだ，はいどうも
4	感動詞/感動詞	53.10	あああああ，いやいやいやいや，うんうんうんうん
	感動詞/副詞	10.67	あそうですか，うんうんそうそう，はいそうですね

４.２.２　「J-TOCC」に関する分析

J-TOCC は身近な話題を二者会話で扱うコーパスであり，場面別の CEJC と構成が異なる。本節では，表 9 の構造型分布に基づいて，語列長別の特徴を分析する。表 9 を見ると，2-gram においては，非定式表現（98）と「実質語 1：その他」（648）が合計で 746 となる。CEJC と同じく，短い語列では，実質語 1 語を核とする簡潔な連結が中心であり，複数実質語と表現文型は相対的に少ない。句レベルの定式表現（12）も少ない。

3-gram になると分布が変化する。表現文型 136，文法コロケーション 217，「実質語 2」233 がいずれも増加し，「実質語 1：その他」367（句レベル 42）と非定式表現 47 は減少する。3 語連接では，半固定の文型および文型＋実質語の結合が目立ち，2 語連接より機能が明確な語列が増える。4-gram では，文法コロケーション 374 が最も多く，「実質語 2」257，表現文型 178 が続く。「実質語 1：その他」169（句レベル 21），非定式表現 22 はさらに減少する。4 語連接では，文法要素と実質語の結合が中心となり，機能の定まった連接が優勢である。5-gram では，この傾向がより明確となり，文法コロケーション 487 が過半に迫り，「実質語 2」278，表現文型 170 が続く。「実質語 1：その他」57（句レベル 7），非定式表現 8 はごくわずかである。語列が長くなるほど，文法項目＋語彙による定まった結合が増加する。6-gram は総数 487 のうち，文法コロケーション 264，表現文型 94，「実質語 2」117，「実質語 1：その他」9（句レベル 2），非定式表現 3 であり，文法主導の連接が中心である。

最後に，「実質語 1：その他」に含まれる句レベル表現を定式表現に算入すると，定式表現の割合は，2-gram が 26.60%，3-gram が 62.80%，4-gram が 83.00%，5-gram が 94.20%，6-gram が 97.95%となる。これらのことから，J-TOCC では語列長の伸長に伴い，文法項目と実質語の結合を核とする語列が高比率となり，3 語以降は機能の明確な文型が大半を占めることが分かった。

表 9：J-TOCC における構造型（トークン）

gram	実質語 0		実質語 1		実質語 2	合計
	非定式表現	表現文型	文法コロケーション	その他		
2	98	60	30	648(12)	164	1000
3	47	136	217	367(42)	233	1000
4	22	178	374	169(21)	257	1000
5	8	170	487	57(7)	278	1000
6	3	94	264	9(2)	117	487

A. J-TOCC における「表現文型」の分析

本節では，J-TOCC の各 gram 上位となった「表現文型」の詳細を検討する。まず，表 10 のトークンを見ると，表現文型のトークンは 2-gram で 90 件，3-gram で 353 件，4-gram で 552，5-gram で 657 と語長の増加に伴って大きく増える。6-gram は 358 であり，上位集合が 487 であることを踏まえると相対比は 73.51%に達する。すなわち，J-TOCC では 3 語以上の語列で表現文型が広く用いられており，特に 4 語，5 語で顕著である。次に，多様性指標を確認すると，タイプは 20，32，35，43，31 と推移し，G.I.は 2.11，1.70，1.49，1.68，1.64 である。3 語と 4 語においてトークンが大きく増える一方で，G.I.は 1.49 まで減少しており，同一の表現文型が繰り返し用いられていることが示唆される。5 語ではタイプが 43 へと伸び，G.I.も 1.68 に回復するが，6 語ではタイプが 31 にとどまり，G.I.も 1.64 に減少する。6 語には，上位集合が少ないということの影響を受けている可能性が高い。

このように，J-TOCC における表現文型は，3 語から 5 語にかけて量的に多くなり，反復的使用が目立つ一方，5 語では種類が増えるという特徴を示す。6 語については総語数の制約があるものの，相対比では高い値を示しており，長い語列でも表現文型が重要な役割を持つことがわかる。

表 10：J-TOCC における「表現文型」および「文法コロケーション」の詳細

gram	トークン	タイプ	G.I.	文法コロケーション（トークン）
2	90	20	2.11	30
3	353	32	1.70	217
4	552	35	1.49	374
5	657	43	1.68	487
6	358	31	1.64	264

B. J-TOCC における「実質語 2」の分析

本節では，J-TOCC の「実質語 2」を 4-gram まで検討する（5-gram 以上は「そうそうそう」「はいはいはい」など相槌の反復が大半を占めるから除外した）。

2-gram では「感動詞/代名詞」13.41%が最も高い（例：あ俺，あ何，あそれ）。発話開始部で相手の注意を促しつつ，直後の指示語で対象を示す用法が多い。続く「感動詞/副詞」12.80%（あまあ，あそうなど）は，感動詞に程度・評価の成分を添えて態度を簡潔に表す。「感動詞/感動詞」9.15%（うんああ，ああ，いやいやなど）は，即時応答として肯定・否定や驚きを短く示す型である。3-gram では「感動詞/代名詞」12.45%（あそれは，うんそれは，あの何だ）が引き続き多く，注意喚起と対象指示を連続して行う。「感動詞/副詞」10.73%（あそういう，うんなるほどね，あそうなん）は，了解・評価を先に置いて発話を続ける前置きとして機能する。「副詞/副詞」9.44%（そうそう，まあまあまあ，なるほどねそう）は，評価副詞の反復により同意の度合いを調整する。4-gram では「副詞/副詞」15.95%が最も高い（そうそうそう確か，ああそっかそっか，まあまあそう）。副詞の連接が増え，同意・納得の強さを段階的に示す例がまとまって現れる。「感動詞/副詞」12.06%（うんそういう，あそうそうそう，ああそっかそっか）は，再認・了解を示しつつ次の内容へつなぐ働きがある。「感動詞/感動詞」9.34%（あああああ，ああはいはい）は，やや強い相槌や反応として用いられる。

J-TOCC の「実質語 2」は，2-gram で反応＋指示・評価の短い型が中心となり，3-gram で

はそれらが前置き・導入として用いられる。4-gram では副詞連接が増え，同意や納得の強さを細かく示す相槌が目立つ。いずれの型も発話冒頭や応答部に置かれやすく，相手の発話を受ける，対象を示して話題をそろえる，自身の態度を簡潔に述べるといった対人的機能を担っている。

表 11：J-TOCC における「実質語 2」の主な構造型および割合（%）

gram	構造	割合	例
2	感動詞/代名詞	13.41	あ俺，あ何，ああそれ，ああ何，ああ俺
	感動詞/副詞	12.80	あまあ，ああそう，ああそっ，ああちょっと
	感動詞/感動詞	9.15	あうん，あああ，あああの，ああいや
3	感動詞/代名詞	12.45	あそれは，あの何だ，うんそれは，うん何だろう
	感動詞/副詞	10.73	あそういう，あそうなん，うんなるほどね
	副詞/副詞	9.44	そうそう，なるほどねそう，まあまあまあ
4	副詞/副詞	15.95	そうそうそう確か，まあまあそう，なるほどねそう
	感動詞/副詞	12.06	あそうそうそう，ああそっかそっか，うんそういう
	感動詞/感動詞	9.34	あああああああ何，ああはいはいはい

４.２.３　「CSJ」に関する分析

本節では，CSJ における構造型を語列長別に概観する。表 12 を見ると，2-gram では，「実質語 1：その他」が 715 で最も多く，次いで「実質語 2」が 122，表現文型は 65，非定式表現は 95，文法コロケーションは 3 の順となる。最短語長では実質語 1 が中心とする単純な連接が多数を占める。3-gram では，「実質語 1：その他」487 が依然として最多であるものの，表現文型 156，「実質語 2」200，文法コロケーション 86 がいずれも増加する。3 語連接の段階から，説明・展開に用いる表現や，実質語どうしの結び付きが目立ち始める。4-gram に至ると分布がほぼ均衡し，「実質語 2」268，表現文型 239，文法コロケーション 223，「実質語 1：その他」241 が近い値を示す。一方，非定式表現は 29 まで減少する。4 語は，語彙連接・文型・文法結合が並行して用いられていることが明らかである。5-gram では，文法コロケーションが 393 で最も多く，表現文型 274，「実質語 2」226 が続く。「実質語 1：その他」88，非定式表現は 19 まで減少する。6-gram では，文法コロケーションが 567 で過半に達し，表現文型 213，「実質語 2」166 が続く。「実質語 1：その他」は 23，非定式表現は 31 である。7-gram ではこの傾向がさらに強まり，文法コロケーションが 700 に達し，「実質語 2」180，表現文型 87，「実質語 1：その他」29，非定式表現 4 となる。7 語の上位は，プレゼンテーションの構成要素や結論部で繰り返して用いられる定まった決まり文句が大半を占める。

以上をまとめると，CSJ では語長の増加に伴い文法コロケーションの割合が一貫して多くなり，4 語と 5 語で表現文型とあわせて主要な骨格を成すことが明らかとなった。また，「実質語 2」は 4 語をピックに達し，後は緩やかに減少する。「実質語 1：その他」の括弧内値が全て 0 である点は，謝辞や依頼のような句レベルの定式表現がほとんど現れないことを示しており，アカデミックな発話様式の特性が数値に反映されていると言えよう。

表 12：CSJ における構造型（トークン）

gram	実質語 0		実質語 1		実質語 2	合計
	非定式表現	表現文型	文法コロケーション	その他		
2	95	65	3	715(0)	122	1000
3	71	156	86	487(0)	200	1000
4	29	239	223	241(0)	268	1000
5	19	274	393	88(0)	226	1000
6	31	213	567	23(0)	166	1000
7	4	87	700	29(0)	180	1000

A. CSJ における「表現文型」の分析

表 13 の値に基づき，語列長別の出現量と多様性を確認する。まず，トークン数は 2-gram の 68 から 7-gram の 787 まで語列が長くなるほど表現文型が多用される。タイプ数は 2-gram の 36 から 4-gram の 60 まで増え，のち 5-gram で 51，6-gram で 43 と減少し，7-gram で 52 となり，やや回復する。G.I.は 2-gram の 4.37 が最も高く，3-gram の 3.28，4-gram の 2.79，5-gram の 1.97，6-gram の 1.54 と徐々に減少し，7-gram で 1.85 とやや持ち直す。これらの推移は，4-gram 以降は多様性より反復使用が相対的に強まることが示唆される。また，文法コロケーションを見ると，割合は連続的に上昇し，6〜7 語では表現文型の大部分が文法コロケーションで構成される。以上より，CSJ では語列長の増大に応じて表現文型の出現規模が拡大し，4-gram 以降は表現文型の反復が優勢となる。同時に，文法コロケーションの寄与が段階的に増大し，6〜7 語では表現文型の大部分がこのタイプで構成される。アカデミックスピーキングの構築において，定まった文型に内容語をはめ込む運用が強く働いていることを示す結果であろう。

表 13：CSJ における「表現文型」および「文法コロケーション」の詳細

gram	トークン	タイプ	G.I.	文法コロケーション（トークン）
2	68	36	4.37	3
3	242	51	3.28	86
4	462	60	2.79	223
5	667	51	1.97	393
6	780	43	1.54	567
7	787	52	1.85	700

B. CSJ における「実質語 2」の分析

表 14 は，CSJ における「実質語 2」のうち出現割合の高い構造型を語長別に示すものである。2-gram では「感動詞/名詞」（25.25%）が最も多く，「えーそれぞれ」「えー次」など，発話冒頭で聞き手の注意を引いたり間を稼いだりして直後の名詞で話題を指示する機能を果たすと思われる。次いで「動詞/名詞」17.17%（「いうこと」「いるもの」）が高く，命題の名詞化を用いて論点を提示する典型的な説明型の連結である。「副詞/動詞」は 10.10%となり，叙述の程度・様態を先行させて述べていく機能を果たすであろう。3-gram では「動詞/名詞」は 21.14%を占め，「いうこと＋助詞」や「いるもの＋助詞」など命題の提示や定義づ

け，「名詞/動詞」は 16.00%を占め，「ことがわかる」「ことが言える」のような結論・評価の明確化に用いられる。「感動詞/代名詞」は 21.27%となり，「えここで」「えこちらの」などのような注意喚起や指示対象の明示を同時に行う役割を担う。4-gram では，「名詞/動詞」17.54%と「動詞/名詞」11.19%が上位を占め，「名詞+動詞」または「動詞+名詞」の枠組みが説明・判断の骨格として定着する段階に入る。5-gram では「名詞/動詞」24.34%が再び最上位となり，「動詞/名詞」19.03%が続く。いずれも命題展開や根拠提示を進めるための定型的な述語構成である。「動詞/感動詞」17.26%（「ありますのでえー」「おりますのでえー」等）は，原因を表す接続助詞で間を稼ぐか，考えながら次の内容へ移行するかといった機能を果たすと考えられる。このように，CSJ の「実質語 2」は，短い語列では対人的管理を担う感動詞型や名詞結合が中心となり，語列が長くなるにつれて命題提示・説明・判断を行う「名詞+動詞」「動詞+名詞」の結合が高比率を占めるようになる。アカデミックスピーキング特有の説明志向が，これらの構造分布に明らかに反映されているとうかがえる。

表 14：CSJ における「実質語 2」の主な構造型および割合（%）

gram	構造	割合	例
2	感動詞/名詞	25.25	え/えー・それぞれ/次/今回
	動詞/名詞	17.17	いう・こと/もの/方法，いる・こと/もの
	副詞/動詞	10.10	ある・こと/場合/訳，この・こと/部分/場合/結果
3	動詞/名詞	21.14	いう・こと+助詞/もの+助詞/ふう
	名詞/動詞	16.00	ことが・わかる/考える/言える，ことを考える
	感動詞/代名詞	11.43	え・ここで/これは/これら/こちらの/これが/これを
4	動詞/感動詞	21.27	ありまして/いうのが/いうのはえー
	名詞/動詞	17.54	ことが・わかる/挙げる/考える/言える
	動詞/名詞	11.19	あることがわかる，いうことが・わかる/考える/言える
5	名詞/動詞	24.34	が必要になります，ことが・わかる/わかると思う/挙げる
	動詞/名詞	19.03	あることがわかる，いうことがわかる/挙げる/言える
	動詞/感動詞	17.26	ありますので/おりますので・えー

４.２.４　３コーパスの比較分析

本節では，3 コーパスにおける特有の定式表現を，（A）定式表現の分布と割合，（B）表現文型，（C）「実質語 2」コロケーションの 3 観点から比較する（図 1～図 4）。

A. ３コーパスにおける定式表現の分布と割合

まず，図 1 は定式表現（FL）と非定式表現（NonFL）の割合を示す。いずれのコーパスでも語列が長くなるにつれて定式表現の割合が上昇するという共通傾向が見られる。2 から 4-gram の推移を具体的にみると，CEJC は 35.40%, 72.60%, 87.60%, J-TOCC は 26.60%, 62.80%, 83.00%，CSJ は 19.00%，44.20%，73.00%である。これらのことから，短い語列では自由度の高い非定式表現（NonFL）が一定量残るが，3 語・4 語へ進むにつれていずれも定式表現が多くなることがわかろう。ただし，上昇の仕方には差があり，CEJC は 3-gram 時点で一気に 7 割超に達するのに対し，CSJ は 4-gram でもまだ 7 割程度にとどまる。J-TOCC は両者の中間で，段階的に定式表現が増えていく。

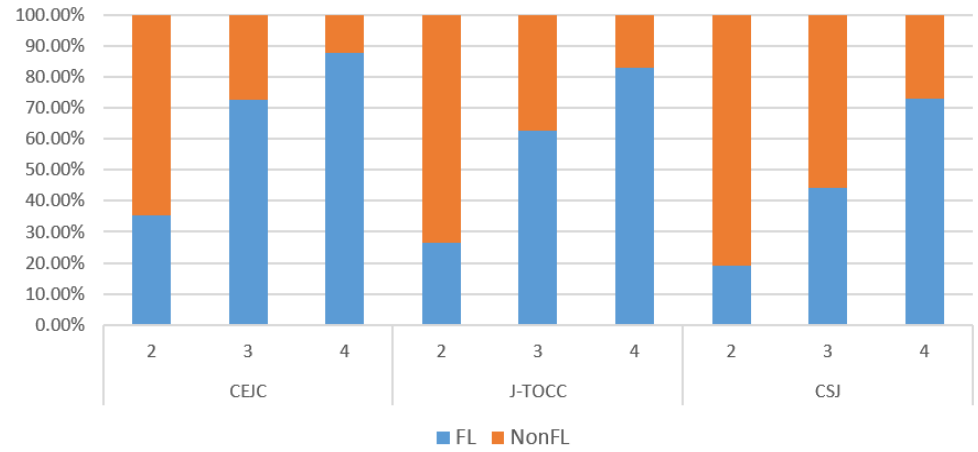

図 1：3 コーパスにおける定式表現の割合

図 2 は定式表現の内訳である。まず CEJC では，「実質語 2」（collocation）が 3-gram・4-gram で大きく伸び，次いで表現文型（pattern）が支える構図である。さらに注目すべきは完全句（phrase）が明確に確認される点で，3-gram で 62，4-gram で 38 と，日常の雑談に固有の決まり文句（例：ありがとうございます，お願いします など）が一定量を占める。J-TOCC でも完全句は出現するがそれほど多くない，代わって文法コロケーション（gramcollo）が 3-gram で 217，4-gram で 374 と多くなった。このことから，話題別の 2 者会話では，表現文型＋実質語で発話を運ぶ型が中核を成すことが分かる。一方，CSJ では完全句が 2〜4-gram のいずれにも観察されない。定式表現は表現文型とコロケーションでほぼ占められ，とりわけ 4-gram ではコロケーション（268）が表現文型（239）と並び立つ。これは，アカデミックスピーキングが挨拶・応答のような決まり文句をほぼ排し，説明・定義・評価など命題内容を展開するための文型と語彙を多用されることが言えよう。また，定式表現は語列の増大に伴って共通して高まるが，内訳はコーパスによって対照的である。CEJC は完全句と語彙コロケーションが多く，J-TOCC は文法コロケーションが多く，CSJ は完全句を欠いたまま，表現文型と実質語主導のコロケーションで比率を押し上げる。言い換えれば，親密な相互行為の維持（CEJC），トピックの段階的展開（J-TOCC），学術的説明の構築（CSJ）という談話目的の違いが，定式表現の構成比にそのまま反映されている。

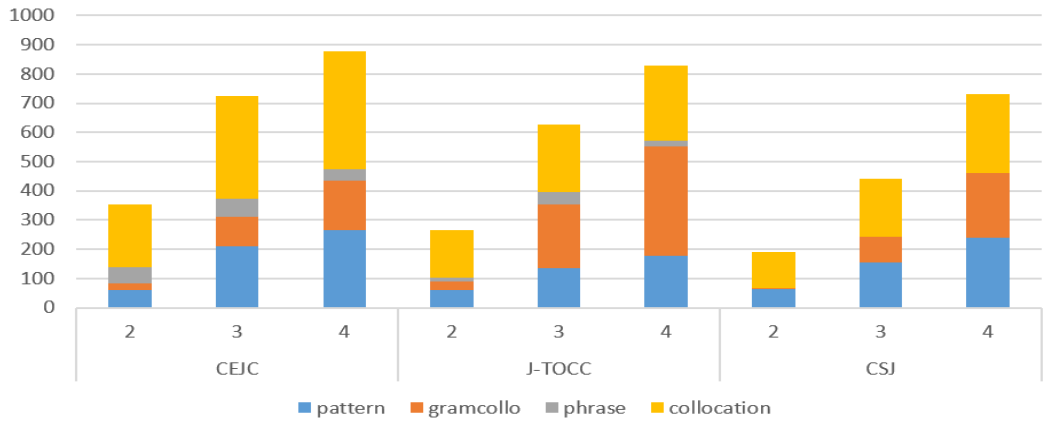

図 2：3 コーパスにおける定式表現の種類別の内訳

B. 3 コーパスにおける表現文型の比較

図 3 は，表現文型の数（token），異なり語数（type），および文法コロケーション（gramcollo）の推移を示す。まずトークンは，J-TOCC と CSJ で 3 語以降に多くなり，特に J-TOCC は 4 語・5 語で非常に多い。計画性の低い雑談である CEJC でも増加は見られるが，上記 2 コーパスに比べれば緩やかである。タイプは全体を通じて CSJ が最も高く，4 語でピークを示す。同じ長さの語列でも，アカデミックスピーキングでは相対的に多様な文型が使い分けられていることが明らかであろう。CEJC のタイプは中位で，5 語で J-TOCC よりやや少ない一方，J-TOCC は相対的に低く，限られた文型の反復が多い。

また，文法コロケーションに着目すると，コーパス間の対照がいっそう明確である。J-TOCC は 3 語から 5 語と段階的に増加し，5 語で最大となる。すなわち，文型に実質語を 1 語だけ差し込む文法コロケーションが，トピック導入や展開の足場として広く機能している。CSJ も語列長の増大に伴い着実に増え，長い語列では説明・評価の骨格を担う用法が中心となる。これに対し，CEJC は文法コロケーションの伸びが小さく，雑談では文型そのものの繰り返しと，実質語どうしの結合が中心になりやすい。よって，J-TOCC は「文型＋実質語 1」のような文法コロケーションが多く，CSJ は多様な文型の併用と実質語主導の構築，CEJC は限られた文型の繰り返しと語彙コロケーションの強さという文体差が，トークン・タイプ・文法コロケーションの 3 指標の組み合わせから定量的に裏づけられる。

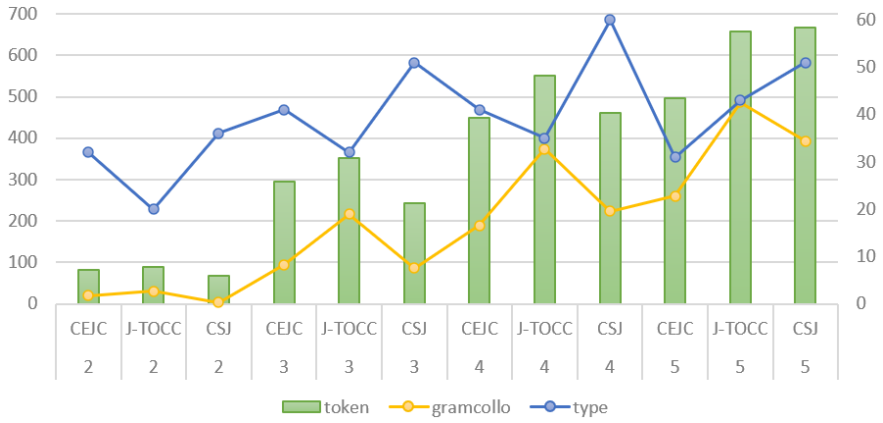

図 3：3 コーパスにおける表現文型の比較

C. 3 コーパスにおける「実質語 2」の比較

図 4 は，「実質語 2」の上位 3 パターンのみを取り上げ，語長別の分布を可視化したものである。2 語では，CEJC は「感動詞/感動詞」が圧倒的であり，応答詞の反復による相槌が中心となっている。これはターン交替の円滑化や対人配慮を担う相互行為資源として機能する。J-TOCC は「感動詞/名詞」「感動詞/代名詞」が上位で，感動詞によるターンの立ち上げ＋指示・照応の結合が目立つ。CSJ は「動詞/名詞」「感動詞/名詞」など命題的結合が相対的に強く，開始段から説明志向が現れる。上位三型の共有数は CEJC - J-TOCC で 3 型程度，J-TOCC - CSJ，CEJC - CSJ ではゼロである。3 語では，CEJC は引き続き感動詞連鎖が最上位で，感情表出と関与の維持が談話を進める役割があろう。J-TOCC は「副詞/副詞」「感動詞/代名詞」が多く，評価副詞や立場標識の反復による同意・了解の段階化，および参照の確定が談話運営の核をなすであろう。CSJ では「動詞/名詞」「名詞/動詞」が優勢で，定義・

言い換え・因果提示といったテキスト構築に直結する結合が支配的である。2-gram では 3 コーパスで上位型の重なりはほとんど見られなかったが，3-gram では文体の固有性が鮮明になった。上位三型の共有は，CEJC - J-TOCC が 1 型であり（感動詞起点の結合），J-TOCC - CSJ が 1 型（感動詞/代名詞），CEJC - CSJ はゼロである。J-TOCC が両者の中間に位置し，CEJC と CSJ は最も遠い。さらに，4-gram では，CEJC は「感動詞/感動詞」の連鎖が上位を占め，感情表出の強化が連続語列として顕在化する。J-TOCC は「感動詞/副詞」「副詞/副詞」が強く，評価副詞・談話標識を介した接続・展開操作が中心となる。CSJ は「名詞/動詞」に加えて「動詞/感動詞」が上位に入り，命題提示＋モダリティ標識（終助詞的要素）により説明のリズムを整える。上位型の共有は CEJC - J-TOCC が三型となっており，割合が異なる。両方とも CSJ との重なりが見られなかった。

このように，コーパス間は CEJC が CSJ から最も遠く，J-TOCC は中間位置であり，語長が伸びるほど文体差による分布の乖離が拡大する。CEJC は感動詞連鎖という相互行為管理が中心，CSJ は動詞/名詞・名詞/動詞といった命題構築が中心，J-TOCC は短い語長で感動詞起点，長い語長で評価副詞・談話標識を核とする談話運営型連鎖へと重心が移る。ホットマップは，相互行為の維持（CEJC）/談話運営の調整（J-TOCC）/学術的説明の構築（CSJ）という文体固有の機能要求が，「実質語 2」の構文選択に体系的な差異を生み出していることを，パターン共有の少なさと順位の乖離として裏づけている。

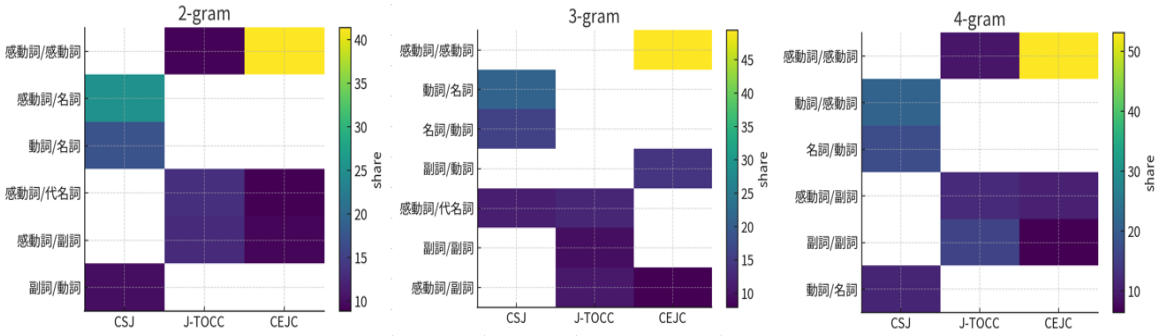

図 4：3 コーパスにおける「実質語 2」の対照

５．まとめと今後の課題

本研究は，日本語話し言葉における定式表現の文体差を明らかにすることを目的とし，CEJC（雑談），J-TOCC（話題別），CSJ（学会講演）から各 100 万語を無作為抽出して分析した。2〜7 語の *n*-gram を対象に上位 1,000 項目を対象に，表現文型・文法コロケーション・「実質語 2」・「実質語 1：その他」・非定式表現に分類し，「実質語 1：その他」に含まれる完全句は定式表現として再集計した。その結果，3 コーパスに共通して語列が長くなるほど定式表現の割合が上昇することがわかった。とりわけ固定的定式表現（文法コロケーション＋実質語 2＋完全句）は，J-TOCC で 5 語 77.20%・6 語 78.64%，CSJ で 5 語 61.90%・6 語 73.30%・7 語 88.00%に達し，CEJC でも 4 語 60.90%まで増加した。また，「実質語 2」の構成は文体により異なり，CEJC では感動詞の情動的反復と語彙コロケーションが中核，J-TOCC では 3〜5 語で文法コロケーションが多くなり談話運営を支え，CSJ では表現文型のタイプが最も多様で，4 語以降は「名詞 - 動詞」，「動詞 - 名詞」を核とする内容志向の連鎖が優勢であった。さらに，句レベルの完全句は CEJC に多く，CSJ ではほぼ出現しなかった。

これらのことから，雑談は相互行為を，話題別の会話では文型と実質語の結合を，学術講演は説明構築のための文型を相対的に多用することが示唆される。教育面では，文体別リストと典型表現の提示により状況適合的な表現選択の習得が期待される。

今後の課題として，本研究で用いた *n*-gram モデルは表層の形態素連接によるため，活用の正規化に基づく多語表現の抽出を併用して精緻化する必要があることが挙げられる。また，完全句の判定は規則と目視確認に依存しており，参照リストの拡充とアノテーション信頼性の検証が求められる。さらに，本研究は分布差を主として頻度から叙述した。機能や談話位置（発話冒頭・応答部等）との対応づけを進めたい。最後に，学習者コーパスへの適用や縦断的追跡を通じて，定式表現の習得を解明し，教育実践への還元を図ることが今後の課題である。

謝　辞

本研究は「江苏高校哲学社会科学研究基金项目『日语程式语心理表征研究（2021SJA2071）』」および「江苏大学人才引进科研启动基金项目『日语高频程式语语言学特征与心理表征关系（5501150006）』」の助成を受けたものである。

文　献

Altenberg, B. (1998) On the Phraseology of Spoken English: The Evidence of Recurrent Word-combinations. In A. P. Cowie (Ed.), *Phraseology Theory, Analysis and Application*, 173-194. Oxford: Oxford University Press.

Biber, D. (2009) A Corpus-driven Approach to Formulaic Language in English: Multi-word Patterns in Speech and Writing. *International Journal of Corpus Linguistics*, 14 (3): 275-311.

Biber, D., & Barbieri, F. (2007) Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, 26(3): 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004) Lexical Bundles in Speech and Writing an Initial Taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, 71-92. Harlow: Longman.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1998) *Longman Grammar of Spoken and Written English*. Essex: Pearson Education.

Dunning, T.（1993）Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), 61–74.

Durrant, P. (2017) Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories. *Applied Linguistics, 38*(2), 165–193.

Hyland, K. (2008) As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes, 27*(1), 4–21.

Nattinger, J. R., & DeCarrico, J. S. (1992) *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Ren, J. (2021) Variability and Functions of Lexical Bundles in Research Articles of Applied Linguistics and Pharmaceutical Sciences. *Journal of English for Academic Purposes, 50*, Article 100968.

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

Vela-Rodrigo, A. Á. (2023) A Lexical Bundle Analysis of Art-related Crowdfunding Projects. Ibérica

Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE), 46: 321-349.
Wei, Y., & Zhong, Y. (2023) The Processing Advantage of Multiword Sequences: A Meta-analysis. *Studies in Second Language Acquisition*, 46(2): 427-452.
Wray, A. (2002) *Formulaic Language and Lexicon*. Cambridge, UK: Cambridge University Press.
Wray, A. (2008) *Formulaic Language: Pushing the Boundaries.* Oxford: Oxford University Press.
Wray, A. (2009) Future directions in formulaic language research. *Journal of Foreign Languages*, *32*, 2-17.
国広哲弥（2007）「連語」「慣用句」, 飛田良文・遠藤好英・加藤正信・佐藤武義・蜂谷清人・前田富祺 (編著)『日本語学研究事典』. 171-172, 明治書院.
国立国語研究所（2006）「日本語話し言葉コーパスの構築法」国立国語研究所. https://doi.org/10.15084/00001357.
国立国語研究所（2001）「現代語複合辞用例集」国立国語研究所.
小西円（2017）「日本語学習者と母語話者の産出語彙の相違－I-JAS の異なるタスクを用いた比較－」『国立国語研究所論集』13, 79–106.
小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 「『日本語日常会話コーパス』設計と特徴」『国立国語研究所論集』24, pp.153-168, 2023.1.
杉浦正利（2001）コーパスを利用した日本語学習者と母語話者のコロケーション知識に関する調査」『日本語電子化資料収集・作成－コーパスに基づく日本語研究と日本語教育への応用を目指して－』平成 12 年度名古屋大学教育研究改革・改善プロジェクト報告書, 64-81.
杉浦正利・朴秀智（2002）「日本語学習者作文コーパスにおける形態素レベルの共起表現について」『日本語学習者辞書編纂に向けた電子化コーパス利用によるコロケーション研究』平成 13 年－平成 15 年科学研究費補助金基盤研究 (B) (二) 中間報告論文集, 1-10.
蘇振軍（2019）「定式表現研究の動向と今後の課題－日本語の定式表現研究の発展に向けて－」『教育学ジャーナル』24: 33-42.
蘇振軍（2024）「日语学习者与本族语者书面语词束研究」『高等日语教育』14, 13–24.
森田良行・松木正恵（1989）『日本語表現文型－用例中心・複合辞の意味と用法』アルク.
中俣尚己（2014）『日本語教育のための文法コロケーションハンドブック』くろしお出版.
中俣尚己（2015）「初級文法項目の生産性の可視化－動詞に接続する文法項目の場合」『計量国語学会』29（8）: 275-295.
中俣尚己（2021）「日本語話題別会話コーパス：J-TOCC 解説資料」http://nakamata.info/database/j_tocc_document.pdf（2025 年 6 月閲覧）.
畑佐由紀子（2022）『学習者を支援する日本語指導法－音声・語彙・読解・聴解－』くろしお出版.
堀恵子・李在鎬・長谷部陽一郎（2016）「機能語用例文データベース『はごろも』について」『計量国語』30(5): 275-285.
李在鎬，長谷部陽一郎（2017）「N-gram を使った文法項目の抽出と学習者コーパスに基づく妥当性検証」『計量国語学』31(2) : 116-127.