言語研究（*Gengo Kenkyu*）**162**: 47–62（2022）　　doi: 10.11435/gengo.162.0_47

統語依存関係に基づく位相研究
――文章ジャンルの位相差を対象に――

李　文平　上海財経大学
劉　海濤　浙江大学
小森　早江子　中部大学

【要旨】 本研究では，異なる文章ジャンル間の相違点と共通点を把握するために，統語情報が付与された日本語コーパス，具体的には「現代日本語書き言葉均衡コーパス」（国立国語研究所）のコアデータのツリーバンクである「UD Japanese-BCCWJ」を利用し，白書，新聞，書籍，雑誌，Yahoo! ブログ，Yahoo! 知恵袋の 6 つのジャンル間で統語依存関係，統語複雑度，依存方向において位相差があるか否かを考察した。その結果，統語依存関係，統語複雑度，依存方向はジャンルごとに異なっており，これらの指標は内容語か機能語かといった指標と同様に，文章ジャンルの特徴をある程度反映していることがわかった。また，統語複雑度と関係する依存距離は最小化される傾向があるという位相間に共通の法則が見られた *。

キーワード： 位相，文章ジャンル，統語依存関係，統語複雑度，依存方向

1. はじめに

日本語位相論 [1] は菊沢（1933）によって提案された。菊沢は，日本語を支配する言語社会は常に変化しており，社会が様相を異にするごとに言語も変化すると考え，これを日本語の「位相」と定義した。そして，日本語位相論とは，位相の相違による変化の状況を究め，その間にはたらく法則を見出す研究分野であるとした（菊沢 1933: 6–7）。また，菊沢（1933）は位相論を様相論と様式論に分類し，前者は言語社会を背景としてその位相の相違を考察し，後者は音声言語か文字言語か等の表現様式の相違を考察するものとした。菊沢は位相論が日本語の研究において音韻論，意義論，構成論と同様に重要な位置を占めると指摘している（菊沢 1933: 8–9）。

前田（1988）は菊沢（1933）を踏まえ，位相の定義を「言語の使用者の所属する社会集団の違い，言語を使用する場面の違いなどによって，言語がいろいろな形をとることを言語の位相という」（前田 1988: 24）と整理している。

田中（1999: 1）は位相に基づく言語上の差異を「位相差」と呼び，位相差をもたらす要因を社会的位相（性別によるもの，世代によるものなど），様式的位相（書きことば・話しことばの差異によるもの，文章ジャンル・文体の差異によるものなど），心理的位相（忌避の心理によるもの，美化の心理によるものなど）の 3 つに

* 本研究の執筆にあたり 3 名の査読者の先生方より論文の構成，統計的な手法による検定などに関して的確で建設的な意見をいただき，心よりお礼を申し上げる。

[1] 本研究では「日本語」を「国語」と同義とする。

分類している（田中 1999: 9–10）。また，文章の位相差をもたらす要因の一つとして，様式的位相の下位分類である「文章ジャンルによるもの」をあげている（田中 1999: 9–10）。

本研究では，菊沢（1933）の言う「言語のすがたが変化する」という立場から位相を捉え，文章ジャンルの違いにより言語のすがたも異なるという観点を出発点として，言語のすがたの差異の中に一定の法則を見出すことを試みる。その際，言語の最も重要な表現形式の 1 つである統語に注目する。統語的側面における位相差とはどのようなものかという問題を明らかにするために，本研究では統語情報が付与されたコーパスを用いて，統語依存関係，統語複雑度，依存方向におけるジャンル間の位相差を分析する。

2. 先行研究

近年，国立国語研究所の「現代日本語書き言葉均衡コーパス」（Balanced Corpus of Contemporary Written Japanese，以下 BCCWJ とする）を用いて文章の位相差を分析する研究が増えてきた。BCCWJ は書籍，雑誌，新聞，白書など 13 のジャンルにまたがる約 1 億 490 万語を格納した大規模コーパスである（Maekawa et al. 2014，国立国語研究所コーパス開発センター 2015）[2]。BCCWJ の中の 1% のサンプルは，人手により修正され解析精度が高くなっている。このデータは「コアデータ」と呼ばれる（山崎編 2014: 83）。コアデータの文章ジャンルは，白書，新聞，書籍，雑誌，Yahoo! ブログ（以下ブログ），Yahoo! 知恵袋（以下知恵袋）の 6 つである。

BCCWJ を利用した文章の位相差の研究としては，小磯ほか（2009），宮内（2012），丸山（2015）があげられる。小磯ほか（2009）では，BCCWJ の白書，新聞，書籍，知恵袋，国会会議録のデータと，国立国語研究所「日本語話し言葉コーパス」（Corpus of Spontaneous Japanese: CSJ）の学会講演，模擬講演のデータを資料とし，名詞率・機能語率・接続詞率などの 7 つの指標を用いて各ジャンル間の文体の差や類似性を調べた結果，名詞率が「書籍，新聞，白書」の順に高くなることを明らかにしている。また，機能語率の逆数が Halliday（1985）の提案した語彙密度 [3] に相当するとし，「書籍，新聞，白書」の順に語彙密度が高くなることから，この順に文章として複雑でフォーマル（あるいはより綿密に計画されたもの）であると述べている。知恵袋に関しては，和語率・漢語率・名詞率において書籍よりも白書や新聞に近い値を示し

[2] 書籍全般，雑誌全般，新聞，白書などを「レジスター」と呼ぶ研究（例えば，山崎 2014，丸山 2015，Maekawa et al. 2014，国立国語研究所コーパス開発センター 2015 など）がある一方，「ジャンル」と呼ぶ研究（例えば，小磯ほか 2009 や宮内 2012 など）もある。本研究では，小磯ほか（2009），宮内（2012），BCCWJ の公式サイトの呼び方にしたがい，「ジャンル」と呼ぶ。BCCWJ の公式サイトの URL は以下のとおりである。https://pj.ninjal.ac.jp/corpus_center/bccwj/index.html［2021 年 1 月 8 日アクセス］

[3] Halliday（1985: 63–64）は内容語（lexical item）の数を内容語と機能語（grammatical item）の合計の数で割る比率のことを語彙密度と定義している。文章が複雑であるほど，内容語の割合が高い傾向にあると述べている。

ていることを報告し，質問という行為においてはある程度改まった言葉遣いをすることや，回答・解説における専門性の高さなどが影響したものと説明している。

宮内（2012）は，小磯ほか（2008）や佐野・丸山（2008）などの先行研究を踏まえ，「書籍 2（文学），書籍 1（文学以外），新聞，白書」の順によりフォーマルであること，また，この逆順に話し言葉的な特徴が強くなること（宮内 2012: 43）を前提として，接続助詞の特徴について分析を行い，個々の接続助詞には明確に文体的特徴に関わる違いがあることを明らかにしている。また，接続助詞の出現傾向と文体的特徴の関連性に基づいて知恵袋の文体的特徴を考察し，知恵袋の文体には「フォーマルでない」「話しことば的」（宮内 2012: 50）などの特徴があると指摘している。

丸山（2015）は BCCWJ のコアデータと CSJ の学会講演，自由会話データを用いて格助詞の使用を調べ，BCCWJ における格助詞の使用に関して，白書は「デ」の代わりに「ニオイテ」を用いるなど，かたい書きことばとしての性質を持っており，知恵袋と対極の関係にあることを報告している。格助詞の使用において，新聞は多少白書に似た性質を持つ一方で，知恵袋とブログはともに話しことば的な性質を帯びていることを明らかにしている。

小磯ほか（2009），宮内（2012），丸山（2015）の研究結果をまとめると次のようになる。白書は最も書きことば的，そして複雑でフォーマルなジャンルである。新聞は白書と似た性質を持っている。書籍と雑誌は書きことば的であるが，新聞ほどフォーマルではない。ブログはさらにフォーマルさが低くなり，話しことば的な性質を帯びている。BCCWJ を利用した位相に関するほかの研究（例えば，佐野・丸山 2008，鯨井 2012，秋本 2016 など）の結果も，この結果とほぼ一致している。

しかし，先行研究の研究対象は主に内容語や機能語に集中し，内容語や機能語以外を対象とした位相差に関する研究は十分に行われているとは言えない。言語の位相は，内容語や機能語にばかりでなく，音韻・文字・統語・文章・文体などの各分野において現れるものである（前田 1988，田中 1999）。本研究では，内容語や機能語以外の要因として，統語的側面における位相差について考察する。そのために，本研究では統語情報が付与（アノテーション）されたコーパス，すなわちツリーバンク（Abeillé ed. 2003）を使用し，統語的側面におけるジャンル間の位相差を分析する。

3. 研究方法とデータ

3.1. 統語依存関係に基づく分析

統語上の依存関係とは，文を作る過程で要素（単語）A と要素（単語）B がどのように関係するかということである（Tesnière 1959，児玉 1987，2007，Hudson 2007，Liu 2008）。統語上の依存関係は自然言語処理において最も広く応用された統語モデルであり，構文解析器の開発・評価及び解析精度・解析効率の向上に使われている（牧野・納富 1991，平川 2005，若林 2014 など）。統語依存関係は 2 つの要素の間に成立する関係であり，2 つの要素において 1 つが主要素（head）で，もう 1 つが従要素（dependent）であるとされている（Tesnière 1959, Hudson 2007, Liu

2010）。

例文「二事業の指定取り消しを決めた」[4] の統語依存関係を図 1 に示す。主要素から従要素への弧線は依存関係の方向を示す。主要素と従要素の依存関係の内容は弧線の上に示す。例えば，「obj」は目的語を，「nmod」は名詞修飾語を，「aux」は助動詞を表す。

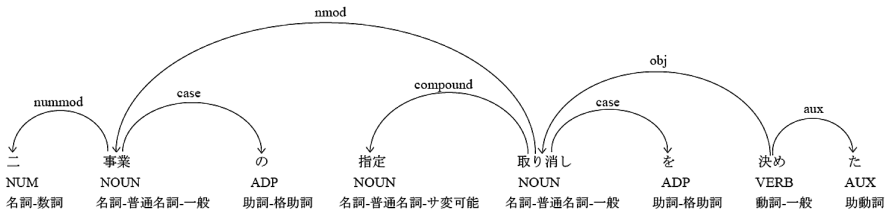

図 1　例文「二事業の指定取り消しを決めた」の統語依存関係図

図 1 は表 1 のようにツリーバンク形式でも示すことができる。

表 1　例文「二事業の指定取り消しを決めた」のツリーバンク形式の表記方法 [5]

従要素			主要素			依存関係	依存距離
語番号	語	品詞	語番号	語	品詞		
1	二	NUM	2	事業	NOUN	nummod	1
2	事業	NOUN	5	取り消し	NOUN	nmod	3
3	の	ADP	2	事業	NOUN	case	-1
4	指定	NOUN	5	取り消し	NOUN	compound	1
5	取り消し	NOUN	7	決め	VERB	obj	2
6	を	ADP	5	取り消し	NOUN	case	-1
7	決め	VERB	0	/	/	root	/
8	た	AUX	7	決め	VERB	aux	-1

表 1 は従要素，主要素，依存関係，依存距離という 4 つの項目をまとめたものである。表の左端に語番号を示した。主要素と従要素の語番号の差は両者の間の距離（依存距離 dependency distance）を表す（Hudson 1995: 15–16, Liu 2008: 164–166）。隣接する主要素と従要素の依存距離の絶対値は 1 になる。

統語依存関係に関する研究においては，統語上の特徴の考察には 2 つの指標が有効であると考えられている（Liu et al. 2009, Jiang & Liu 2015, Futrell et al. 2015, Wang & Liu 2017, Poiret & Liu 2020, Yadav et al. 2020）。1 つは文の統語複雑度或いは統語難

[4] 例文は BCCWJ の PN3e_0005-11 より引用したものである。

[5] 品詞と依存関係の略号の意味は次のとおりである。NUM は数詞，NOUN は名詞，ADP は接置詞，VERB は動詞，AUX は助動詞を意味する。nummod は数詞，nmod は名詞修飾語，case は格標識，compound は複合語，obj は目的語，root は文の主辞，aux は助動詞を意味する。

易度を測る指標である「依存距離」，もう 1 つは言語類型論的な特徴を測る指標である「依存方向」である。

まず，依存距離については，1 文の平均依存距離（mean dependency distance）を式（1）のように算出する（Liu 2008, Liu et al. 2009, Jiang & Liu 2015）。

（1）$$MDD_{sentence} = \frac{1}{n}\sum_{i=1}^{n}\left|DD_i\right|$$

式（1）における n は 1 文における依存関係の数を表す。DD_i は 1 文における i 個目の依存関係の依存距離を表す。つまり，式（1）は 1 文におけるすべての依存距離の絶対値の和を依存関係の数で割った平均値を意味する。ここで注意すべき点は，文の主辞（root）は他の要素の従要素となることがなく，対応する主要素がないため，平均依存距離を計算する際には主辞を除く必要があるということである。例えば，先の「二事業の指定取り消しを決めた」には 8 つの語があるが，7 つ目の語「決め（る）」は文の主辞であり，対応する主要素がないため，平均依存距離を計算する際には除外する。「決め（る）」以外の 7 つの語の間の依存距離はそれぞれ表 1 の一番右の列に示してある。式（1）により，表 1 の文の平均依存距離を算出すると「(1+3+1+1+2+1+1) / 7 = 1.43」となる。

式（1）はジャンル全体にも応用できる。

（2）$$MDD_{genre} = \frac{1}{n}\sum_{i=1}^{n}\left|DD_i\right|$$

式（2）における n はジャンル全体の依存関係の数を表す。DD_i はジャンルにおける i 個目の依存関係の依存距離を表す。つまり，式（2）はジャンルにおけるすべての依存距離の絶対値の和を依存関係の数で割った平均値を意味する。

次に，依存方向については，主要素は従要素より前に位置する場合もあれば，従要素より後に位置する場合もある。主要素が従要素より前に位置する場合は「主要素前置」であり，依存距離はマイナスである。これに対し，主要素が従要素より後に位置する場合は「主要素後置」であり，依存距離はプラスである。主要素前置は主に助詞・助動詞を従要素にする場合である。表 1 において助詞「ノ」と「ヲ」，及び助動詞「タ」の依存方向は主要素前置であるが，ほかの語の依存方向は主要素後置である。依存方向を比較する際には，一般に主要素前置または主要素後置の依存関係の割合を用いる（Liu 2010, Jiang et al. 2019）。

3.2. 使用するデータ

本研究で使用するツリーバンクは，BCCWJ のコアデータのツリーバンクである（Asahara et al. 2018, 浅原ほか 2019）。このツリーバンクは Universal Dependencies（UD）ガイドライン 2.0 の基準にしたがい，構築されたものであるため，「UD Japanese-

BCCWJ」とも呼ばれている（浅原ほか 2019: 24）。UD Japanese-BCCWJ の文章ジャンルは，白書，新聞，書籍，雑誌，Yahoo! ブログ，Yahoo! 知恵袋の 6 つである。

UD Japanese-BCCWJ における語の単位は BCCWJ と同様に短単位である[6]。品詞体系は Universal POS Tags version 2（UPOS）を採用している。例えば，名詞は NOUN で，動詞は VERB で表示される。前述したように，依存関係は UD ガイドライン 2.0 基準に基づいたものであり，例えば，名詞句主語は nsubj，目的語は obj と表示される。UD Japanese-BCCWJ の依存関係情報は機械的に自動変換した後，人手による修正が 3 回行われた（Asahara & Matsumoto 2016）。UD Japanese-BCCWJ は現段階で日本語ツリーバンクの中で解析精度が最も高く，データ規模が最も大きいツリーバンクである。本研究では UD Japanese-BCCWJ の train データを使用する。データの概要を表 2 にまとめる。

表 2　UD Japanese-BCCWJ の概要[7]

文章ジャンル	語の数	文の数	平均文長	百語あたりの文の数
新聞	267,374	13,491	19.82	5.05
書籍	156,834	7,196	21.79	4.59
雑誌	157,245	9,546	16.47	6.07
白書	150,996	4,456	33.89	2.95
ブログ	46,025	3,362	13.69	7.30
知恵袋	43,636	2,839	15.37	6.51
合計	822,110	40,890		
平均			20.11	5.41

3.3. データ処理手順

UD Japanese-BCCWJ のデータを以下の手順によって処理した。

(i)　データを新聞, 書籍, 雑誌, 白書, ブログ, 知恵袋の 6 つのジャンルに分類する。
(ii)　ツリーバンクにおけるテキストデータと統語情報データを分ける。
(iii) 句読点を削除し，統語情報データを表 1 の形に整形する。
(iv) 文長や依存距離などを計算する。

4. 結果と考察

4.1. 統語依存関係における位相差

本節では，「主語・述語」，「述語・目的語」，「修飾語・被修飾語（形容詞修飾，

[6] 短単位は，言語の形態的側面に着目して規定した言語単位である（国立国語研究所コーパス開発センター 2015: 75）。意味を持つ最小の単位のことで，例えば「公害紛争処理法」を「公害／紛争／処理／法」と細かく分割する（国立国語研究所コーパス開発センター 2015: 5）。
[7] 本研究では，UD Japanese-BCCWJ の 20180327 バージョンを使用した。語数は句読点を削除した数値である。大村・浅原（2018）の示す語数や文の数との違いは UD Japanese-BCCWJ のバージョンや語数の計算方法などの違いによるものと考えられる。

副詞修飾, 名詞修飾)」の 3 つの統語関係を取り上げ, これらの統語関係におけるジャンル間の位相差を分析する[8]。これらの統語関係の数と, それが各ジャンルの統語関係の総数に占める割合を表 3 にまとめた。

表 3　統語関係の数と割合[9]

	主語・述語	述語・目的語	修飾語・被修飾語			統語関係の総数
			形容詞修飾	副詞修飾	名詞修飾	
白書	2,569	4,731	3,494	1,172	**18,971**	144,613
	(1.78%)	(3.27%)	(2.42%)	(0.81%)	**(13.12%)**	
新聞	5,888	**8,459**	4,452	3,009	30,755	253,055
	(2.33%)	**(3.34%)**	(1.76%)	(1.19%)	(12.15%)	
書籍	**3,771**	4,675	3,970	3,947	15,260	148,824
	(2.53%)	(3.14%)	(2.67%)	(2.65%)	(10.25%)	
雑誌	3,455	4,499	3,644	3,370	15,667	146,714
	(2.35%)	(3.07%)	(2.48%)	(2.30%)	(10.68%)	
ブログ	815	768	973	**1,388**	4,168	41,686
	(1.96%)	(1.84%)	(2.33%)	**(3.33%)**	(10.00%)	
知恵袋	984	816	**1,158**	1,335	3,742	40,657
	(2.42%)	(2.01%)	**(2.85%)**	(3.28%)	(9.20%)	

まず,「主語･述語」の関係について分析する。表 3 を見ると, 書籍における主語･述語の関係の割合は 2.53% で, 6 つのジャンルで最も高い。これに対し, 白書における主語・述語の関係の割合は 1.78% で, 6 つのジャンルで最も低い。先の表 2 を見ると, 白書における百語あたりの文の数は 2.95 で, 6 つのジャンルで最も少ないことから, 文の数が少ないと主語・述語の関係の割合も低くなると考えられる。文の数のほかに, 1 文における主語・述語の関係の数も主語・述語の関係の割合に影響を及ぼすと考えられる[10]。主語・述語の関係を含まない文, すなわち主語を含ま

[8] 本研究における統語依存関係の集計は表 1 における依存関係のラベルによって行った。具体的には, 主語･述語の関係のラベルは nsubj であり, 述語･目的語の関係のラベルは obj である。形容詞修飾の関係のラベルは amod であり, 形容詞と形容動詞の連体修飾と連用修飾が含まれる。副詞修飾の関係のラベルは advmod であり, 名詞修飾の関係のラベルは nmod である。例えば, (i)「初めて見た写真でも, ほぼ正確に分類できた（PN1c_00005 より）」, (ii)「例えば, 財団法人の郵政弘済会がある（PN2b_00005 より）」という 2 つの文において, 主語・述語の関係にあるのは (ii) の「会」と「ある」である。形容詞修飾の関係にあるのは (i) の「正確」と「(分類) できる」である。副詞修飾の関係にあるのは, (i) の「初めて」と「見る」,「ほぼ」と「正確」, 及び (ii) の「例えば」と「ある」である。名詞修飾の関係にあるのは (ii) の「法人」と「会」である。紙幅の関係上統語依存関係とラベルの詳細については浅原ほか (2019) を参照されたい。

[9] 各依存関係において統語関係の割合が最も高い箇所は太字で, 最も低い箇所は下線で表した。なお, 前述のように UD Japanese-BCCWJ における語の単位は短単位で, 助詞・助動詞の統語関係は全体の 37.59% を占めている。

[10] 査読者から主語・述語の関係の多寡を比較する際に, 単文と複文・重文を区別する必要があるとの指摘を受けた。筆者はこの指摘に同意する。しかし, 現段階では UD Japanese-

ない単文（便宜上「無主述関係文」と呼ぶ）の数，主語・述語の関係を 1 つだけ含む文，すなわち主語を含む単文（便宜上「単主述関係文」と呼ぶ）の数，そして，主語・述語の関係を 2 つ以上含む文，すなわち複文・重文（便宜上「複主述関係文」と呼ぶ）の数のそれぞれがジャンル全体の文の数に占める割合を計算したものが表 4 である。表 4 を見ると，書籍における無主述関係文の割合は 60.37% で最も低い。白書は 65.78% で書籍に次ぎ，2 番目に低い。これに対し，ブログの無主述関係文の割合は 80.64% で最も高い。このことから，ブログは主語の省略が多いが，書籍と白書は主語の省略が少ないと言えよう。また，複主述関係文の割合を見ると，白書における複文・重文の出現率が高い（15.19%）のに対し，ブログにおける複文・重文の出現率は低い（3.69%）ことがわかった。

表 4　無主述関係文・単主述関係文・複主述関係文の数と割合

	無主述関係文の数（割合）	単主述関係文の数（割合）	複主述関係文の数（割合）	ジャンル全体の文の数
白書	2,931（65.78%）	848（19.03%）	677（15.19%）	4,456
書籍	4,344（60.37%）	2,128（29.57%）	724（10.06%）	7,196
新聞	9,104（67.48%）	3,235（23.98%）	1,152（8.54%）	13,491
雑誌	6,781（71.03%）	2,191（22.95%）	574（6.01%）	9,546
ブログ	2,711（80.64%）	527（15.68%）	124（3.69%）	3,362
知恵袋	2,032（71.57%）	666（23.46%）	141（4.97%）	2,839

次に，「述語・目的語」の関係について分析する。表 3 を見ると，述語・目的語の関係の割合は，新聞が 3.34% と最も高い。白書は 3.27% で新聞に次ぎ，2 位である。ブログと知恵袋は述語・目的語の関係の割合が低く，約 2% である。述語・目的語の関係の数において 6 つのジャンルの間に有意な差が見られるか否かを検証するために，カイ二乗検定を行った。その結果，新聞と白書の間には有意な差が見られなかった（$\chi^2 = 1.46, p > .05$）が，白書と書籍の間には有意な差が見られた（$\chi^2 = 4.01, p < .05$）。また，書籍と雑誌の間には有意な差はないが（$\chi^2 = 1.37, p > .05$），雑誌と知恵袋の間には有意な差があった（$\chi^2 = 129.65, p < .001$）。ブログと知恵袋の間には有意な差が見られなかった（$\chi^2 = 2.96, p > .05$）。この結果から，述語・目的語の関係を指標とした場合，6 つのジャンルは大きく「新聞と白書」，「書籍と雑誌」，「ブログと知恵袋」という 3 つのグループに分けられることがわかる。

この結果は丸山（2015）の格助詞「ヲ」に関する調査結果とほぼ一致している。丸山（2015）は BCCWJ における格助詞「ヲ」の割合を調べた結果，新聞と白書は格助詞「ヲ」の割合が高いのに対し，ブログと知恵袋は格助詞「ヲ」の割合が低い

BCCWJ には単文，複文，重文のタグが付いておらず，短期間で UD Japanese-BCCWJ ほどの大規模ツリーバングに単文，複文，重文などのタグを手作業で付けることは現実的ではない。今回は，単文と複文・重文の分布傾向を把握するために，1 文における主語・述語の数を計算した。

と述べている。UD Japanese-BCCWJ では述語・目的語の関係は係助詞「ハ」，格助詞「ヲ」，無助詞の 3 つの形式で表されるが，格助詞「ヲ」の出現傾向はほぼ述語・目的語の関係の出現傾向と一致していると言えよう。

最後に，「修飾語・被修飾語」の関係については次のような結果になった。白書は名詞修飾の割合（13.12%）が高いのに対し，副詞修飾の割合（0.81%）が低い。新聞も白書と同様に名詞修飾の割合（12.15%）は高いが，副詞修飾の割合（1.19%）は低い。知恵袋は白書と新聞に比べて副詞修飾（3.28%）の割合が高いのに対し，名詞修飾（9.20%）の割合が低い。ブログも副詞修飾（3.33%）の割合が白書と新聞に比べて高いのに対し，名詞修飾（10.00%）の割合は白書と新聞より低い。

4.2. 統語複雑度における位相差

本節では統語複雑度におけるジャンル間の位相差を分析する。統語複雑度を測る指標として，主要素と従要素との間の依存距離を用いる（Liu 2008, Liu et al. 2017, Yadav et al. 2020）。

6 つのジャンルの平均依存距離（MDD_{genre}）を表 5 にまとめた。表 5 を見ると，白書の平均依存距離は 3.761 で，6 つのジャンルで最も長い。新聞の平均依存距離は 3.148 で，白書に次いで 2 位である。白書と新聞の平均依存距離は 3 を超えている。ブログ，書籍，雑誌の 3 つのジャンルの平均依存距離は近似しており，いずれも約 2.9 である。知恵袋の平均依存距離は 2.677 で，6 つのジャンルで最も短い。

表 5　6 つのジャンルの平均依存距離

文章ジャンル	平均依存距離（MDD_{genre}）
白書	3.761
新聞	3.148
ブログ	2.965
書籍	2.924
雑誌	2.909
知恵袋	2.677

6 つのジャンルにおける個々の文の平均依存距離（$MDD_{sentence}$）を図 2 の箱ひげ図に示した。また，個々の文の平均依存距離において 6 つのジャンルの間に有意な差が見られるか否かを検証するために Brunner-Munzel 検定[11] を行い，その結果を表 6 にまとめた。表 6 の右端の Cliff's delta は，ノンパラメトリック検定における効果量の指標である。Cliff's delta の効果量から，すべてのジャンルの間の差は small 以

[11] Brunner-Munzel 検定はデータ分布の正規性も等分散性も仮定しない 2 群の差を比較するノンパラメトリック検定である。2 群の比較において，前者の方が大きい場合 Brunner-Munzel の結果はマイナスであり，後者の方が大きい場合 Brunner-Munzel の結果はプラスである。

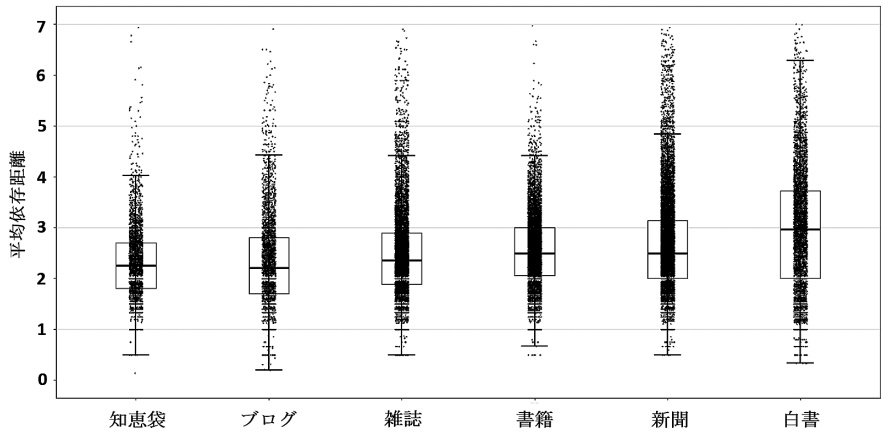

図 2　各ジャンルにおける個々の文の平均依存距離（$MDD_{sentence}$）

表 6　平均依存距離（$MDD_{sentence}$）に関する Brunner-Munzel 検定と Cliff's delta の結果

	Brunner-Munzel	*p*	Cliff's delta	
白書　対　新聞	-17.089	< .0001	0.1836	(small)
新聞　対　書籍	-0.450	0.6529	0.0037	(n.s.)
新聞　対　雑誌	-12.190	< .0001	0.0944	(negligible)
新聞　対　ブログ	-15.889	< .0001	0.1802	(small)
書籍　対　雑誌	-11.545	< .0001	0.1036	(negligible)
書籍　対　ブログ	-15.298	< .0001	0.1955	(small)
雑誌　対　ブログ	-7.788	< .0001	0.0951	(negligible)
雑誌　対　知恵袋	-7.623	< .0001	0.0909	(negligible)
ブログ　対　知恵袋	1.038	0.2993	-0.0155	(n.s.)

下であることがわかった[12]。

検定の結果，次のことがわかった。

第一に，白書の平均依存距離（$MDD_{sentence}$）は 6 つのジャンルで最も長い。白書における文章の統語複雑度は最も高いと言える。フォーマルな文章の平均依存距離が長くなる傾向があることが知られており（Hiranuma 1999, Wang & Liu 2017），白書は 6 つのジャンルにおいて最もフォーマルなジャンルであると言えよう。

第二に，新聞全体の平均依存距離（MDD_{genre}）は書籍よりやや長いが（表 5），

[12] n.s. は有意でないことを意味する。small や negligible の判断は Romano et al.（2006）によるものである。なお，依存距離を対象とした言語間の大規模研究の結果から，数多くの言語において隣接する単語の依存関係の割合は 50% 以上を占めていることが明らかになっている（Liu 2008）。大量の短距離依存関係が存在するため，言語間の平均依存距離の差が小さく，同一言語内のジャンル間の平均依存距離の差がさらに小さい結果になると考えられる。本研究では平均依存距離において効果量が「negligible」であっても，「n.s.」（有意でない）に比べれば効果がないとは言えない。

Brunner-Munzel の結果，新聞と書籍の間には有意差がないことが確認された。このことから，新聞の統語複雑度や文章のフォーマルさは書籍に似ていると言える。また，新聞と書籍の平均依存距離（$MDD_{sentence}$）は雑誌やブログに比べて長いことから，新聞と書籍の統語複雑度は雑誌やブログに比べて高いことがわかった。

第三に，ブログ全体の平均依存距離（MDD_{genre}）は書籍や雑誌に比べて長いが（表5），個々の文の平均依存距離を対象とした Brunner-Munzel 検定の結果，ブログの統語複雑度は書籍や雑誌よりも低いことがわかった。これはブログにおける個々の文の平均依存距離のばらつきが大きいためと考えられる。Maekawa et al.（2010）は白書，新聞，書籍，ブログ，知恵袋の一部のデータの品詞分布と文長を対象に線形判別分析を行い，ブログにおける文のばらつきが大きいため，ブログを入れた文章ジャンル分類の精度がブログ抜きの際に比べて低くなったと報告している。ブログにおける文のばらつきが大きいということは，ブログには少数の長距離依存関係を表す例と，大量の短距離依存関係を表す例が同時に存在することを意味する。少数の長距離依存関係を表す例が存在する結果，ブログ全体の平均依存距離（MDD_{genre}）が長くなる。一方，大量の短距離依存関係を表す例が存在するため，Brunner-Munzel 検定を行う際にブログから取り出した文の平均依存距離は雑誌から取り出した文の平均依存距離よりも短い確率が高くなると考えられる。

第四に，ブログと知恵袋の間に平均依存距離（$MDD_{sentence}$）について有意差がないことから，ブログは知恵袋に似た性質を持つと言える。

佐野・丸山（2008），宮内（2012），丸山（2015）などの内容語か機能語かを指標とした位相研究では，「ブログ，雑誌，書籍，新聞，白書」の順に文章として複雑でフォーマルになると述べている。本研究の結果はこれらの研究とほぼ一致している。知恵袋の文体的特徴に関しては先行研究では研究結果が分かれているが。本研究では，依存距離において知恵袋はブログに似た性質を持つという結果になった。

6 つのジャンルは平均依存距離においてある程度の位相差が見られる一方で，次のような共通する傾向も見られた。まず，図 2 からわかるように，ほとんどの文の平均依存距離は 7 以下である。また，表 5 からわかるように，各ジャンルにおけるすべての文を 1 つのまとまりとして得られたジャンル全体の平均依存距離は 4 以下である。人間の認知メカニズムや文法の制限により（Ferrer-i-Cancho 2004, 2016, Liu 2008, Gildea & Temperley 2010），依存距離は最小化される傾向があるということかもしれない（Liu 2008, Futrell et al. 2015）。依存距離という指標は，文章ジャンルの相違による統語上の相違を示すと同時に，異なる文章ジャンルの間の共通性も見出すこともできるということである。

4.3. 依存方向における位相差

本節では，依存方向を指標としてジャンル間の位相差を分析する。6 つのジャンルにおける主要素前置と主要素後置の割合を図 3 に示す。ジャンル間に有意な差が見られるか否かを検証するために主要素後置の割合についてカイ二乗検定を行い，

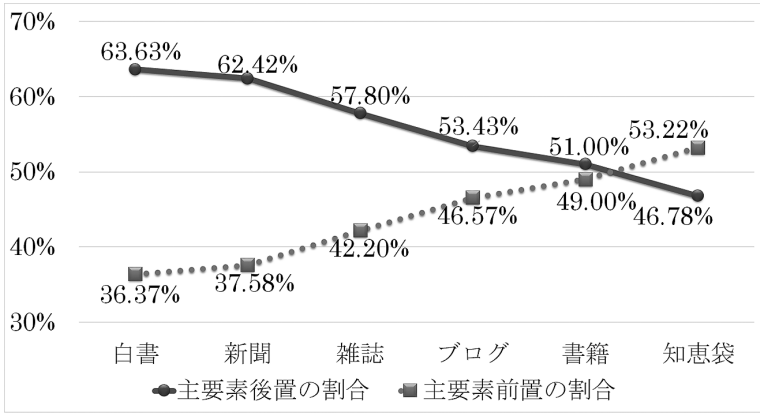

図 3　各ジャンルにおける依存方向の割合

表 7　主要素後置に関するカイ二乗検定の結果

			χ^2	p
白書	対	新聞	57.51	< .001
新聞	対	雑誌	829.77	< .001
雑誌	対	ブログ	252.63	< .001
ブログ	対	書籍	76.88	< .001
書籍	対	知恵袋	228.32	< .001

その結果を表 7 にまとめた。

図 3 から，知恵袋以外の 5 つのジャンルにおいて主要素後置の割合が主要素前置より高いことがわかる。そのうち，白書の主要素後置の割合は 63.63% で，6 つのジャンルで最も高い。新聞の主要素後置の割合は 62.42% で，白書に比べてわずかに低い。雑誌，ブログ，書籍の主要素後置の割合はいずれも 50% 強である。知恵袋の主要素後置の割合は 46.78% で，6 つのジャンルで最も低い。また，表 7 からすべてのジャンルの間に有意な差が確認された。

6 つのジャンルにおいて，主要素前置の割合が主要素後置より高いジャンルは知恵袋のみである。これは知恵袋のデータに助詞と助動詞が多いということである。また，統語複雑度の結果と異なり，書籍の主要素後置の割合はブログより知恵袋の方に近い特徴が見られた。これは本研究では書籍の中の文学作品の会話文とそれ以外を区別していなかったためと考えられる。宮内（2012），丸山（2015）は知恵袋が話しことば的な性質を帯びていることを指摘している。宮内（2012）は書籍の中の文学作品には話しことば的な要素が多く含まれることも指摘している。主要素前置の割合が高いことと話しことば的な要素が多く含まれることには関係があるのかもしれない。書籍の中の文学作品の会話文とそれ以外を区別することで，異なる文

章ジャンルの位相差をより詳しく分析することができると思われる。

5. おわりに

菊沢（1933: 7）は日本語位相論を提案して，位相間の相違を究め，位相間にはたらく法則を見出すことで，日本語の全貌及び本質の一端を把握することができると述べている。また，田中（1999）は位相差をもたらす要因を，社会的位相，様式的位相，心理的位相の 3 つに分類した。本研究では，菊沢（1933）の位相論の理論的枠組みのもと，田中（1999）があげる位相差をもたらす要因のうち，文章ジャンルの違いによる言語の位相差を分析し，異なる文章ジャンルの間に見られる特徴を見出そうと試みた。

本研究では，UD Japanese-BCCWJ をデータとして，白書，新聞，書籍，雑誌，Yahoo! ブログ，Yahoo! 知恵袋の 6 つの文章ジャンルを対象に，統語依存関係の内容，統語複雑度，依存方向を指標とした位相差の分析を行った。研究結果は次の 3 点にまとめられる。

第一に，統語依存関係において 6 つのジャンルの位相差が明らかになった。具体的には，主語・述語の関係では，書籍と白書は主語の省略が少ないのに対し，ブログは主語の省略が多い。述語・目的語の関係では，新聞と白書は述語・目的語の関係の割合が高く，ブログと知恵袋は述語・目的語の関係の割合が低い。修飾語・被修飾語の関係では，白書と新聞は名詞修飾の出現率が高く，副詞修飾の出現率が低い。これに対し，知恵袋とブログは副詞修飾の出現率が高く，名詞修飾の出現率が低い。

第二に，統語複雑度において 6 つのジャンルはある程度の位相差が見られると同時に，位相間に一定の傾向も見られることが明らかになった。白書の平均依存距離が 6 つのジャンルで最も長いことから，白書は統語複雑度が高く，ほかのジャンルと比べて文章がフォーマルであると判断される。一方，知恵袋の依存距離が最も短いことから，知恵袋で使われた文の統語複雑度は低いと言える。依存距離において知恵袋はブログに似た性質を持ち，Web 上の文章として共通する特徴が見られた。このことから，依存距離という指標は先行研究で使われた内容語か機能語かといった指標と同様に，文章ジャンルの特徴をある程度反映していると言えよう。ジャンル間にこのような位相差がある一方で，個々の文の平均依存距離は 7 以下で，ジャンル全体の平均依存距離は 4 以下であるという位相間に共通する傾向も見られた。このことから，6 つのジャンルの依存距離は最小化される傾向がある。これは人間の認知メカニズムや文法の制限によるもので，日本語のみならず言語に共通する本質の 1 つであろう。

第三に，依存方向においても 6 つのジャンルは異なる位相差を示すことが明らかになった。具体的には，白書の主要素後置の割合は 6 つのジャンルで最も高く，新聞の主要素後置の割合は白書に次いで 2 位である。知恵袋は主要素後置の割合が 6 つのジャンルで最も低い。統語複雑度の結果と異なり，書籍の主要素後置の割合は

ブログより知恵袋の方に近い特徴が見られたが，これは本研究では書籍の中の文学作品の会話文とそれ以外を区別していなかったためと考えられる。今後の研究では，異なる文章ジャンルの位相差を分析する際に，書籍の中の文学作品の会話文とそれ以外を区別する必要があろう。

本研究では統語情報が付与された大規模日本語コーパスを使用し，統語的側面から異なる文章ジャンル間の位相差を分析した。本研究で使用した方法は，年代による位相差や書きことばと話しことばの位相差の分析にも応用できる可能性がある。田中（1999）はことばの史的変遷の過程において位相差を究明することが重要な課題の１つであると述べている。今後は，「日本語歴史コーパス」（国立国語研究所）を対象に統語依存関係に基づき，奈良時代から明治・大正時代にかけて日本語の統語複雑度や依存方向がどのように変化してきたかなどの歴史的な観点からの研究が求められよう。

参照文献

Abeillé, Anne (ed.) (2003) *Treebank: Building and using parsed corpora*. Dordrecht: Kluwer Academic Publishers.

秋本瞳（2016）「コーパスにみる話しことばと書きことばの連続性：BCCWJ と CSJ におけるチョット / スコシ，ヨ / ネの出現頻度の比較を通じて」『言語と文明』14: 21–41.

Asahara, Masayuki and Yuji Matsumoto (2016) Bccwj-DepPara: a syntactic annotation treebank on the balanced corpus of contemporary written Japanese. *Proceedings of the 12th Workshop on Asian Language Resources*: 49–58.

Asahara, Masayuki, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura and Yugo Murawaki (2018) Universal Dependencies Version 2 for Japanese. *LREC-2018*.

浅原正幸・金山博・宮尾祐介・田中貴秋・大村舞・村脇有吾・松本裕治（2019）「Universal Dependencies 日本語コーパス」『自然言語処理』26(1): 3–36.

Ferrer-i-Cancho, Ramon (2004) Euclidean distance between syntactically linked words. *Physical Review E* 70: 056135.

Ferrer-i-Cancho, Ramon (2016) Non-crossing dependencies: least effort, not grammar. In: Mehler Alexander, Andy Lücking, Sven Banisch, Philippe Blanchard and Barvara Job (eds.) *Towards a theoretical framework for analyzing complex linguistic networks*, 203–234. Berlin Heidelberg: Springer.

Futrell, Richard, Kyle Mahowald and Edward Gibson (2015) Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33): 10336–10341.

Gildea, Daniel and David Temperley (2010) Do grammars minimize dependency length? *Cognitive Science* 34: 286–310.

Halliday, Michael Alexander Kirkwood (1985) *Spoken and written language*. Oxford: Oxford University Press.

平川秀樹（2005）「選好依存文法（PDG）における文解析能力の評価方式について」『情報処理学会論文誌』46(11): 2744–2752.

Hiranuma, So (1999) Syntactic difficulty in English and Japanese: a textual study. *UCL Working Papers in Linguistics* 11: 309–322.

Hudson, Richard (1995) *Measuring syntactic difficulty*. Unpublished paper. Downloadable at: https://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf [accessed January 2019].

Hudson, Richard (2007) *Language networks: the new word grammar*. Oxford: Oxford University Press.

Jiang, Jingyang and Haitao Liu (2015) The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency

treebank. *Language Sciences* 50: 93–104.
Jiang, Jingyang, Jinghui Ouyang and Haitao Liu (2019) Interlanguage: A perspective of quantitative linguistic typology. *Language Sciences* 74: 85–97.
菊沢季生（1933）『国語位相論』東京：明治書院.
児玉徳美（1987）『依存文法の研究』東京：研究社出版.
児玉徳美（2007）「依存関係の見直し」『立命館文学』601: 100–114.
小磯花絵・小木曽智信・小椋秀樹・冨士池優美・宮内佐夜香（2008）「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第22回研究大会発表論文集』192–195.
小磯花絵・小木曽智信・小椋秀樹・宮内佐夜香（2009）「コーパスに基づく多様なジャンルの文体比較：短単位情報に着目して」『言語処理学会第15回年次大会発表論文集』594–597.
国立国語研究所コーパス開発センター（2015）「『現代日本語書き言葉均衡コーパス』利用の手引第1.1版」国立国語研究所.
鯨井綾希（2012）「同一名詞の反復から見たジャンル間の文体差とその要因：コーパスを用いた定量的分析を通して」『言語科学論集』16: 13–25.
Liu, Haitao (2008) Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2): 159–191.
Liu, Haitao (2010) Dependency direction as a means of word-order typology: a method based on dependency Treebanks. *Lingua* 120(6): 1567–1578.
Liu, Haitao, Chunshan Xu and Junying Liang (2017) Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews* 21: 171–193.
Liu, Haitao, Yiyi Zhao and Wenwen Li (2009) Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics*, 45(4): 509–523.
前田富祺（1988）「武士言葉の世界：位相から見た軍記物語の語彙」『国語学』154: 24–33.
Maekawa, Kikuo, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso and Yasuharu Den (2010) Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. *Proceedings of LREC2010, Malta*: 1483–1486.
Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka and Yasuharu Den (2014) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48(2): 345–371.
牧野武則・納富一宏（1991）「語彙依存文法について」『情報処理学会研究報告自然言語処理』85(4): 25–32.
丸山直子（2015）「コーパスにおける格助詞の使用実態：BCCWJ・CSJにみる分布」『計量国語学』30(3): 127–145.
宮内佐夜香（2012）「接続助詞とジャンル別文体的特徴の関連について：『現代日本語書き言葉均衡コーパス』を資料として」『国立国語研究所論集』3: 39–52.
大村舞・浅原正幸（2018）「UD Japanese-BCCWJの構築と分析」『言語資源活用ワークショップ2018発表論文集』: 161–175.
Poiret, Rafaël and Haitao Liu (2020) Some quantitative aspects of written and spoken French based on syntactically annotated corpora. *Journal of French Language Studies* 30(3): 355–380.
Romano, Jeanine, Jeffrey Kromrey, Jesse Coraggio and Jeff Skowronek (2006) Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? *Annual meeting of the Florida Association of Institutional Research*.
佐野大樹・丸山岳彦（2008）「システミック文法に基づく書きことばの複雑さ測定：日本語大規模コーパスを用いた語彙密度計測」『言語処理学会第14回年次大会発表論文集』1097–1100.
田中章夫（1999）『日本語の位相と位相差』東京：明治書院.
Tesnière, Lucien (1959) *Éléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.（小泉保監訳（2007）『構造統語論要説』東京：研究社.）
若林啓（2014）「部分統語構造を考慮した階層的確率オートマトンに基づく教師なしチャンキング」『情報処理学会論文誌データベース』7(2): 61–69.

Wang, Yaqin and Haitao Liu (2017) The effects of genre on dependency distance and dependency direction. *Language Sciences* 59: 135–147.
Yadav, Himanshu, Ashwini Vaidya, Vishakha Shukla and Samar Husain (2020) Word order typology interacts with linguistic complexity: a cross-linguistic corpus study. *Cognitive Science* 44(4): e12822.
山崎誠（編）（2014）『書き言葉コーパス：設計と構築』, 講座日本語コーパス 2. 東京：朝倉書店.

執筆者連絡先：
李　文平
上海財経大学
e-mail: lwplovely1023@gmail.com

［受領日　2020 年 8 月 18 日
最終原稿受理日　2021 年 5 月 11 日］

Abstract

A Phase Research Based on Syntactic Dependency Relations: For the Phase Differences of Text Genre

Wenping Li　*Shanghai University of Finance and Economics*
Haitao Liu　*Zhejiang University*
Saeko Komori　*Chubu University*

In order to analyze the differences and similarities of text genres, this study used the treebank of the core data of *Balanced Corpus of Contemporary Written Japanese (UD Japanese BCCWJ)*. Comparing six genres, white papers, newspapers, books, magazines, Yahoo! Blogs, and Yahoo! Chiebukuro, we investigated the phase differences in syntactic dependency relations, syntactic complexities, and dependency directions. The result shows that there are obvious differences in syntactic dependency relations, syntactic complexities, and dependency directions among text genres. It suggests that these three indicators reflect the characteristics of text genres to some extent as well as indicators like content words or function words. This research used the Japanese corpus with syntactic annotations to analyze the phase differences of text genres in syntactic aspects. Through these analyses, we found that the dependency distance, which indicates the syntactic complexity, tends to be minimized among different text genres. This could be a common law among phases.