Compound Adjectives in *Corpus of Contemporary American English*: A preliminary study

NISHIBU Mayumi

西 部 真 由 美

Faculty of International Communication, Aichi University

E-mail: mnishibu@vega.aichi-u.ac.jp

要旨

本稿ではアメリカ英語の大規模コーパス（Corpus of Contemporary American English）を用いて、ハイフンで結ばれている複合形容詞の特徴を探った。分析により、頻出する複合形容詞と多種類の項に結合する生産的な項の形態的・意味的特徴が明らかになった。数字を項に含んだ「時間・高さ・長さ・大きさ・深さ・年齢」などを表す複合形容詞の割合が大きい点や頻出の複合形容詞は英米語で共通していた。一方、アメリカ英語コーパスでは特に野球やバスケットボールに関連した例が頻出していることと、AmericanやU.Sが項として含まれる複合形容詞が多いことが特徴的であった。

1. Introduction

1.1. Compound Adjectives

Compound adjectives (CAs) are composed of two or more stems. Most CAs consist of two elements (e.g., *ready-made*, *ever-changing*), while a smaller number of them contain more than two (e.g., *up-to-date*, *step-by-step*). This study analyzes CAs of which elements are combined by hyphens. Hyphens are evidence that each combined element originates from separable stems. Approximately 80% to 90% of CAs modify nouns restrictively (Nishibu, 2015).

This study clarifies the morpho-semantic characteristics of frequently occurring CAs by providing qualitative and quantitative analyses of *Corpus of Contemporary American English* (*COCA*).

The following sections describe the corpus details (1.2) and the methods of this study (1.3). The subsequent sections explore overall frequencies (2.1), frequently appearing examples of CAs (2.2), and productive elements (2.3). Finally, the paper concludes with a summary of the research findings (3).

1.2. *COCA*

COCA consists of over one billion words in 485,202 texts, including 24–25 million words each year from 1990–2019, compiled by Mark Davies at Brigham Young University. It is evenly divided between the genres of TV and movie subtitles, spoken, fiction, popular magazines, newspapers, and academic journals. It provides a multi-functional online search engine. This corpus is the world's largest and most widely used balanced corpus of English.

The genres this study analyzed were the sub-corpora of authentic written registers, Fiction, Magazines, Newspapers, and Academic Journals. Their details can be summarized as follows (Davies, 2022).

Fiction (about 120 million words) includes short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books, and movie scripts. Magazines (about 127 million words) contains nearly 100 different magazines of specific domains, such as news, health, home and gardening, women, financial, religion, and sports (e.g., *Time, Cosmopolitan,* and *Fortune*). Newspapers (about 123 million words) consists of ten newspapers, including *USA Today*, *New York Times*, *Atlanta Journal Constitution*, and *San Francisco Chronicle*. Academic Journals (about 121 million words) contains nearly 100 different peer-reviewed journals. They were selected to cover all academic fields.

1.3. Method

The analysis was conducted using four sections of the corpus: Fiction, Magazines, Newspapers, and Academic Journals to make it comparable with my previous studies. The total number of words came up to 490,805,047 in the whole text.

The procedure was as follows. First, hyphenated restrictive adjectives were retrieved from the texts with the search engine using search forms: **-*_j NOUN* and **-*_j ADJ NOUN*. The search form with wildcards, **-*_j*, means adjectives consisting of any two or more elements combined with hyphens. The first form, **-*_j NOUN,* means hyphenated adjectives followed by a noun, and the second one, **-*_j ADJ NOUN,* indicates those followed by another adjective and a noun. The SEARCH function of the search engine was employed to obtain every sample.

Subsequently, all the nouns and adjectives following hyphenated adjectives were eliminated to extract hyphenated adjectives. The same hyphenated adjectives were integrated to examine their tokens and types. The search results were copied and pasted on Excel files for detailed examination.

Up to 3000 type samples per query can be retrieved online using the SEARCH function in the present version of *COCA*. The corpus is almost ten times larger than *British National Corpus* (*BNC*), and we cannot always retrieve or analyze all samples when the hit number is huge. Therefore, this study limited target samples to those that occurred ten times or more. The search forms were then broken down into forms starting with an alphabet (e.g., ***a*****-*_j NOUN*, ***b*****-*_j NOUN*, ...) and those starting with **?** (= any one letter) for numeral initial-elements (e.g., *?*-*_j NOUN* for one-digit numbers, *??*-*_j NOUN* for two-digit numbers, *???*-*_j NOUN* for three-digit numbers) to retrieve as many samples as possible.

In addition, even when the elements are combined with hyphens, a lexical item consisting of affixes with one stem is regarded as a derivative, not a compound. Therefore, derivatives must be removed from the retrieved samples to examine CAs. However, lexical items categorized into affixes differ considerably among linguists and lexicographers. Thus, this study set the smallest number of affixes, as shown in Table 1, and separated the derivatives from CAs. The affixes were selected based on *Oxford Advanced Learner's Dictionary of English*, wherein the number of affixes is relatively smaller, and *Combining forms*, the terminology specific to bound roots of neo-classical compounds (Bauer, 1983), are employed.

Table 1. Affixes

Prefix	Suffix
ant(i)-, ante-, be-, co-, de-, demi-, dis-, en-, ex-, extra-, hyper-, hypo-, infra-, inter-, intra-, mis-, non-, out-, over-, post-, pre-, pro-, re-, semi-, sub-, trans-, ultra-, un-, under-	-able, -ed, -esque, -est, -ful, -ic, -ing, -ish, -ist, -less, -ly, -ous, -ship, -some, -ward, -wise, -y,

Note: Samples with boxed affixes were found in the texts.

2. Analysis of CAs in *COCA*

2.1. Frequency

The raw frequencies of tokens and types in the texts are summarized in Table 2.

Table 2. Restrictive hyphenated adjectives in *COCA*

	Hyphenated ADJ		Compound ADJ			
	token	type	token		type	
all	524,560 (1068.77PMW)	6,614	499,580 (1017.87PMW) (95.2%)[1]		6,140 (92.8%)[2]	
numeral -stems	40,099 (81.70PMW)	562 97stems	40,099 (8.0%)[3]		562 (9.2%)[4]	
2 element	483,406	6,102	458,426	91.8%[5]	5,628	91.7%[8]
3 element	39,887	486	39,887	8.0%[6]	486	7.9%[9]
4+element	1,267	26	1,267	0.3%[7]	26	0.4%[10]

Notes: 1, 2: the number of CAs / the number of all hyphenated adjectives
3, 4: the number of numeral-stems / the number of all CAs
5-10: the number of CAs / the number of all CAs
PMW: per million words

It was revealed that 95.2% of tokens and 92.8% of types among all hyphenated adjectives were CAs. Furthermore, among all CAs, 91.8% of tokens and 91.7% of types consisted of two elements, while only small percentages of them were three or four or more elements (three-element: 8.0% of tokens, 7.9% of types; four (or more)-element: 0.3% and 0.4%).

Additionally, CAs beginning with Arabic numerals (e.g., *12-year-old*, *24-hour*) accounted for 8.0% of all tokens and 9.2% of all types of CAs, indicating that a variety of Arabic numerals became the initial stem of the CAs.

CAs appeared approximately 1018 times per million words, which is surprisingly frequent.

2.2. Frequent CAs

Let us examine CA samples that appeared at high frequency. Table 3 shows the 50 most frequent CAs, their frequencies per million words (PMW), and percentages in the total tokens. In Table 3, italicized CAs indicate that they were also ranked in BNC's 50 most frequent CAs. Boxed numeral words had different versions written with Arabic numerals, which were counted separately.

Table 3. The 50 most frequent CAs (*COCA*)

1–25	PMW	%	26–50	PMW	%
long-term	35.78	3.35%	one-way	3.07	0.29%
African-American	9.59	0.90%	low-fat	2.99	0.28%
full-time	8.54	0.80%	first-round	2.98	0.28%
all-star	7.80	0.73%	open-ended	2.94	0.27%
low-income	7.78	0.73%	long-distance	2.93	0.27%
follow-up	7.23	0.68%	*working-class*	2.89	0.27%
short-term	7.04	0.66%	3-D	2.89	0.27%
same-sex	6.39	0.60%	one-year	2.86	0.27%
two-year	5.62	0.53%	regular-season	2.73	0.26%
four-year	5.52	0.52%	black-and-white[3]	2.68	0.25%
part-time	5.47	0.51%	face-to-face[3]	2.62	0.25%
middle-class	5.35	0.50%	*large-scale*	2.60	0.24%
medium-high	5.12	0.48%	all-time	2.57	0.24%
hands-on	4.87	0.46%	*two-way*	2.53	0.24%
high-tech	4.63	0.43%	10-year	2.48	0.23%
five-year	4.58	0.43%	best-selling	2.43	0.23%
all-purpose	4.30	0.40%	high-risk	2.40	0.22%
day-to-day[3]	4.09	0.38%	after-school	2.26	0.21%
middle-aged	3.99	0.37%	for-profit	2.26	0.21%
decision-making	3.77	0.35%	cross-country	2.25	0.21%
in-depth	3.58	0.33%	first-time	2.25	0.21%
evidence-based	3.48	0.33%	3-point	2.19	0.21%
at-risk	3.40	0.32%	third-party	2.11	0.20%
three-year	3.35	0.31%	post-cold	2.10	0.20%
high-speed	3.10	0.29%	real-life	2.05	0.19%

Notes: % = the number of tokens / the number of all hyphenated adjectives
3: indicates a three-element CA

The most outstanding fact in Table 3 is that CAs with numerals expressing duration, *two/four/five/three/one-year* frequently appeared, as pointed out in my previous study on *BNC*. The most frequent CA was *long-term*, the same as in *BNC*. Similarly, *short-term*, *full-time*, *part-time*, *middle-class,* and *working-class* were highly frequent in *COCA* and

BNC. The most frequent three-element CA was *day-to-day,* which was also consistent with the results in *BNC*.

On the other hand, some CAs were peculiar to *COCA*. First, *African-American* was the second most frequent, while in *BNC*, *Anglo-Saxon* was ranked.

Second, sport-specific CAs, particularly basketball and baseball, were also conspicuous. Samples, such as *all-star* (*games*), *regular-season* (*games*), *first-round* (*draft/pick/playoff*), *3-point* (*range/line /goals/shooting*) appeared extremely frequently. Another sport-specific one, *cross-country* (*ski/skier/race*), was also ranked.

The third point was the difference in orthography between *BNC* and *COCA*. While *BNC* included *three-dimensional* in the top 50, *3-D* was much more common in *COCA*. The difference in time cohorts between the two corpora: *COCA* includes more recent texts than *BNC*, might have influenced the results in Table 3.

Lastly, *all-purpose* and *at-risk* were specific to *COCA*. It was because particular noun phrases, *all-purpose flour* in cooking recipes and *at-risk youth/children*, repeatedly appeared in *COCA*.

2.3. Productive elements

This section examines productive elements combined with many types of other elements. It means that the CAs are not closed sets, but each element can create a variety of combinations.

First, examine Table 4, which indicates the tokens and types of CAs beginning with Arabic numerals.

Table 4. Productive right elements with Arabic-numeral left elements

	PMW	Token%	Type	Type%
-year-old	25.54	2.39	85	1.29
-year	11.79	1.10	43	0.65
-yard	6.39	0.60	69	1.04
-point	5.16	0.48	22	0.33
-inch	4.48	0.42	22	0.33
-D	2.97	0.28	2	0.03
-hour	2.64	0.25	15	0.23
-degree	2.17	0.20	17	0.26
-minute	2.13	0.20	12	0.18
-day	1.93	0.18	16	0.24
-meter	1.74	0.16	11	0.17
-month	1.22	0.11	12	0.18
-quart	1.17	0.11	7	0.11
-month-old	1.07	0.10	17	0.26

Notes: Token%: the number of tokens / the total tokens of hyphenated adjectives
Type: the number of left element types
Type%: the number of left element types / the total types of hyphenated adjectives

It is evident that [number]-*year-old* was by far the most frequent with 1.29% of all types. It was also true to the results in *BNC*. Similarly, another CA expressing age, *-month-old*, also combined with many numbers. These CAs generally modify human nouns such as *boy*, *girl*, *son*, *daughter*, or people's names.

Right elements expressing duration were also remarkable. Five types of right elements, *-year*, *-hour*, *-minute*, *-day*, and *-month*, selected many different numbers for their left elements.

Expressions of measurement such as length, width, level, and quantity were also included in Table 4. They were *-yard*, *-inch*, *-degree*, *-meter*, and *-quart*.

Numeral left elements with these measurements are versatile and practical adjectives to describe the condition and situation of modified nouns.

Next, Table 5 displays productive left elements other than Arabic numerals.

Table 5. Productive left elements

POS		type	%	POS			type	%
Affix	non-	153	2.31%	Adj	Q	all-	43	0.65%
	anti-	86	1.29%			no-	37	0.56%
	post-	54	0.82%			full-	30	0.45%
	pre-	52	0.79%			half-	23	0.35%
	inter-	22	0.33%		M	high-	104	1.57%
	pro-	16	0.24%			low-	71	1.07%
Combining form	self-	85	1.29%			long-	37	0.56%
	multi-	23	0.35%			short-	22	0.33%
Prep	off-	29	0.44%			big-	18	0.27%
	in-	25	0.38%			lower-	17	0.26%
	on-	21	0.32%			deep-	16	0.24%
Adj/N (color)	red-	31	0.47%			medium-	14	0.21%
	white-	26	0.39%			top-	14	0.21%
	black-	15	0.23%	Adj		open-	22	0.33%
	blue-	14	0.21%			free-	19	0.29%
Adj/N (number)	two-	81	1.22%	Adj/N		public-	15	0.23%
	three-	75	1.13%	Adj/Adv		hard-	19	0.29%
	one-	63	0.95%	Adv		well-	58	0.88%
	five-	49	0.74%			cross-	22	0.33%
	four-	49	0.74%	Noun		U.S.-	20	0.30%
	six-	32	0.48%			drug-	18	0.27%

Adj/N (number)	eight-	17	0.26%	Noun	home-	17	0.26%
	seven-	16	0.24%		life-	15	0.23%
	first-	35	0.53%		water-	15	0.23%
	second-	31	0.47%		hand-	14	0.21%
	third-	21	0.32%		state-	14	0.21%
	single-	45	0.68%		teacher-	14	0.21%
	double-	31	0.47%		medium-	14	0.21%

Notes: %=the number of types / the total types of hyphenated adjectives
Q: quantity M: Adjectives referring to measurements

Again, many numeral words were included. Let us look at the other types.

As shown at the top of Table 5, derivatives with affixes, not CAs, were included in productive elements. This phenomenon is reasonable because productive elements combined with many types come to be treated as affixes over time.

Adjectives of quantity (*all-*, *no-*, *full-*, *half-*) and those of level and length (*high-*, *low-*, *long-*, *short-*) were also typically productive. Particularly, *high-* and *low-* occurred with the most significant number of types (*high-*: 104 types, 1.57%; *low-*: 71 types, 1.07%). Color terms such as *red-, white-, black-*, and *blue-* were combined with many types of elements.

The word class of *self-* (85 types, 1.29%) and *multi-* (23 types, 0.35%) must be determined by paying special attention since their tokens and types are generally numerous. They were regarded as combining forms and counted as CAs in this study. However, if they had been treated as prefixes, all the samples would have been excluded from CAs. The adverbial element, *well-* (58 types, 0.88%), can combine with a variety of past participial verbs (verb-*ed*) (e.g., *well-known*, *well-balanced*, *well-dressed*). This phenomenon was also observed in *BNC*.

One productive noun, *U.S.-,* was specific to this corpus, which was not observed in *BNC*.

Table 6 shows productive right elements.

Table 6. Productive right elements

POS		**type**	**%**	**POS**		**type**	**%**
Verb-ed	-based	143	2.16%	Noun	-level	45	0.68%
	-related	67	1.01%	(scale)	-point	23	0.35%
	-oriented	16	0.24%		-grade	20	0.30%
	-owned	14	0.21%		-size	18	0.27%
	-led	14	0.21%		-class	17	0.26%

Noun-ed	-haired	18	0.27%	Noun (scale)	-yard	15	0.23%
	-colored	15	0.23%		-story	14	0.21%
	-shaped	13	0.20%	Noun (time)	-day	31	0.47%
	-centered	12	0.18%		-year	30	0.45%
V-pp/N	-run	13	0.20%		-time	27	0.41%
Adj	-free	30	0.45%		-age	16	0.24%
	-like	24	0.36%		-term	14	0.21%
	-only	21	0.32%		-hour	13	0.20%
	-rich	18	0.27%	Noun	-style	22	0.33%
	-long	17	0.26%		-cell	15	0.23%
	-specific	14	0.21%		-game	15	0.23%
	-resistant	12	0.18%		-school	15	0.23%
Adj/N	-American	28	0.42%		-care	14	0.21%
Particles	-up	23	0.35%		-line	14	0.21%
	-in	17	0.26%		-state	14	0.21%
	-off	17	0.26%		-water	14	0.21%
					-control	13	0.20%

Note: Noun (scale, time) does not include the samples of [Arabic numeral]-nouns shown in Table 4.

The most productive right element was *-based* (e.g., *evidence-based*, *school-based, Washington-based*) (143 types, 2.16%) and the second most *-related* (e.g., *health-related*, *age-related*) (67 types, 1.01%). These two elements were the most and second-most productive in *BNC* as well. Their formal patterns were [Noun]-[past participial form of a Verb]. Other verbs of this formal pattern (e.g., *-oriented*, *-led*) were also frequent in *COCA* and *BNC*.

A second typical formation is [adjective/noun]-[Noun-ed] (e.g., *dark-haired*, *black-eyed, left-handed*). In the right element, nouns referring to physical appearances are de-nominalized with the morpheme *-ed*. The variation of this type was quite rich (e.g., *-colored, -shaped, -sleeved, -sized*).

Third, nouns referring to scale (e.g., *-level*, *-point*, *-size*) and time (e.g., *-day*, *-year*, *-time*) were highly productive, as we have already seen Arabic numerals in Table 4 (2.3). These types of nouns in Table 6 came after numeral words (e.g., *one-year*, *two-story*), other nouns (e.g., *world-class, king-size*), or adjectives (e.g., *high-level, real-time*).

Adjectival right elements, *-like* and *-free*, were worth mentioning because they were also significantly productive in *BNC*. However, the numbers of types for these elements (*-like*: 30 types, 0.45%; *-free*: 24 types, 0.35%) were not as large as those in my previous study in *BNC* (*-like*: 1326 types, 12.0%; *-free*: 319 types, 2.9%). Two facts might cause the difference. The previous study counted hapax legomena (items that occurred only once), while the present study counted the ones that occurred ten times or more. The

second reason might be the orthographic difference between American and British English. It might be possible that American orthography does not require a hyphen to combine these morphemes. A further study will be indispensable to verify that.

Another noun, *-style,* was the most productive nominal right element, except for scale and time nouns. This point corresponds to the results on *BNC*.

3. Summary

The main findings of this study are as follows. The most frequently appeared CA was *long-term,* and its opposite, *short-term*, *full-time*, and *part-time*, also appeared frequently in *COCA and BNC*. In contrast, CAs related to baseball and basketball, *all-star*, *first-round*, *regular-season*, and *3-point,* were remarkably frequent in *COCA*. Regarding productive elements, *-based* was combined with the most varied elements in *COCA* and *BNC*. Verb-ed (past participle) and Noun-ed were also productive as the right element. Numbers in the left position and words for measurements in the right position were highly productive combinations. The right elements, *-like* and *-free,* were productive in *COCA*, but the types and ratios were much less than those in *BNC*, probably because this study did not count samples that occurred less than ten times. All samples, including hapax legomena, will be necessary for further study.

References

Bauer, L. (1983) *English Word-formation*. Cambridge: Cambridge University Press.

Davies, M. (2022) Texts. *The Corpus of Contemporary American English.* https://www.english-corpora.org/*COCA*/

Nishibu, M. (2015) Fukugougo-no bunseki: Genteiyouhou-no fukugoukeiyoushi-no baai [Study on compound words: in the case of restrictive compound adjectives]. In T. Fukaya and N. Takizawa (Eds.), *Koopasu-to Eigo-no Bunpou Gohou* [Corpus and English Grammar and Usage.] (pp. 41–69). Hitsuji Shobou.

Corpora

British National Corpus (*BNC*). Mark Davies at Brigham Young University: http://corpus.byu.edu/BNC/

Corpus of Contemporary American English (*COCA*). Mark Davies at Brigham Young University: http://corpus.byu.edu/COCA/