A Corpus-based Study of Three-element Compound Adjectives

NISHIBU Mayumi

Faculty of International Communication, Aichi University

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

要　旨

本稿ではイギリスの現代英語コーパス（British National Corpus）を用いて、3要素から成る複合形容詞の特徴を探る。分析方法は、2.1節で3要素複合形容詞のタイプ・トークンの調査を行い、2.2節では各テキストジャンル（フィクション・雑誌・新聞・学術論文、等）での分布状況を調べる。2.3節では高頻度語を特定し、その意味的構造的特徴を考察する。2.4節では、3要素形容詞の全例から多くに共通して抽出できる意味的構造的および音韻的特徴を明らかにする。分析結果として、年齢や子供の月齢などを表す数値表現**-year/month/week-old*、「名詞–前置詞–名詞」（e.g., *day-to-day*, *step-by-step*）、等位構造（e.g., *black-and-white*, *up-and-coming*）が高い頻度で現れ、強調詞・緩和詞、困難さや予定を表すto不定詞句や比較表現が全体的に顕著であることなどを含め、8種類の意味的構造的パターンを指摘することができた。

1. Introduction

1.1. Three-element Compound Adjectives

Compound adjectives (CAs) are composed of two or more stems and usually modify nouns restrictively. Most consist of two words (e.g., *ready-made*, *ever-changing*), while a smaller number contain more than two words (e.g., *up-to-date*, *step-by-step*).

This paper analyzes CAs in which three elements (stems or words) are combined by hyphens. Most of the previous literature on CAs is concerned with two-element CAs, listing examples and categorizing their formal or semantic patterns (e.g., Mizuno 1993,

Conti 2006). The features of three-element CAs (CA3s, hereafter) have not yet been fully discussed. Therefore, this study aims at clarifying the characteristics of CA3s by providing detailed qualitative and quantitative analyses of English corpus data.

Section 1.2 describes the corpus and method used in this study. The subsequent sections analyze token and type frequencies (Section 2.1) and the distribution of CA3s across text genres (Section 2.2), and describe frequently appearing examples (Section 2.3) as well as typical formal and semantic patterns (Section 2.4). The paper concludes with a summary of the research findings (Section 3).

1.2. Data and Method

This study analyzes English data from the online British National Corpus (BYU-BNC) with the assistance of the search engine created by Mark Davies. The data are written texts which consist of six text genres: academic, non-academic, fiction, magazines, miscellaneous (i.e., advertisements and brochures), and newspapers.

The analysis procedure was as follows. First, hyphenated CA3s were retrieved from the corpus by the search engine. The search results were modified in Excel for the purpose of examining the token and type frequencies of the CA3s and the nouns they modified. Fragments, single words hyphenated to exaggerate letters (e.g., o-l-d), and acronyms were excluded. Hyphenated items that are not CAs were manually excluded from the sample.

2. Analysis of Three-element Compound Adjectives based on the BNC

2.1. Token and Type Frequency

Table 1 summarizes the token and type frequencies of CAs in the BNC written texts by the number of CA elements.

Table 1. Frequencies of CAs in the BNC written texts

	Token			Type	
#elements	Raw freq	PMW	Ratio	Raw freq	Ratio
3	17012	169.5	6.2%	3642	5.9%
2	255872	2549.8	93.2%	57031	93.0%
4 +	1565	15.6	0.6%	677	1.1%
Total	274449	2734.9	100.0%	61350	100.0%

Note: PMW = Per Million Words; 4 + = four or more elements

As was expected, CA3s account for a much smaller proportion of the total CAs (6.2% of tokens; 5.9% of types) than two-element CAs (93.2% of tokens; 93.0% of

types) in the BNC written texts. However, thanks to the large volume of the corpus, 17,012 tokens representing 3642 types of CA3s were retrieved, a number large enough for an examination of their characteristics. The token frequency of all CA3s, 169.5 per million words, is equivalent to that of words ranked in 650th place in the written parts of the BNC, according to a vocabulary list (Leech et al. 2001: 185).

Some CAs appear only once in the entire corpus; these are called "hapax legomena." In previous works, they are sometimes disregarded as idiosyncratic uses, as items created in an ad-hoc way, or as being hyphenated simply to avoid syntactic ambiguity. However, hapax CAs merit close attention because they can be important manifestations of productivity, particularly when a certain element or pattern unites a variety of elements to create various CAs, as we will see in Section 2.4. Table 2 shows how many CAs are hapax legomena.

Table 2. Frequency of hapax CAs in the BNC written texts

#elements	Raw freq	PMW	Ratio	Token%*[1]	Type%*[2]
3	2453	24.4	6.5%	14.4%	67.4%
2	34972	348.5	92.2%	13.7%	61.3%
4 +	515	5.1	1.4%	32.9%	76.1%
Total	37940	378.1	100.0%	13.8%	61.8%

Note: In hapax legomena, token and type are the same in number.
*1 = Number of hapax legomena ÷ total token number of each type of CA (2, 3, 4+ elements).
*2 = Number of hapax legomena ÷ total type number of each type of CA (2, 3, 4+ elements).

Table 2 shows that 2453 CA3s (14.4% of tokens, 67.4% of types) are hapax legomena. These figures mean that, among all the CA types combined, only 33.6% of CAs appears more than once in the corpus, and a small number of types account for as much as 85.6% of the total tokens. This indicates that a small number of CA3s appear with high frequency.

Another point worth mentioning is that the percentage of hapax CA3s is similar to that of two-element CAs, whereas that of four-or more-element CAs is much higher than the others.

2.2. Genres

The distributional ratio of the CA3s across six text genres is illustrated in Figure 1. For comparison, that of two-element CAs is illustrated in Figure 2. Since the text sizes are not equal, the figures are based on the number of tokens per million words rather than raw frequency.

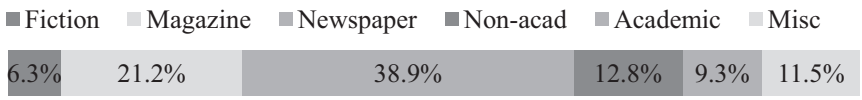

Figure 1. Three-element CAs across six text genres

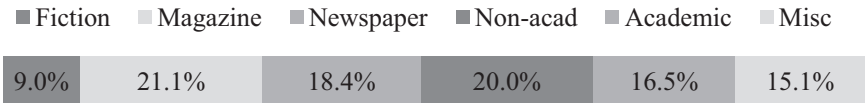

Figure 2. Two-element CAs across six text genres

Figure 1 shows that CA3s appear most frequently in Newspapers (38.9%), followed by Magazines (21.2%). By contrast, they appear least in Fiction.

In comparison with the distribution of two-element CAs, in which Magazines hold the largest amount (21.1%) and the other three genres closely follow it (Non-academic, 20.0%; Newspapers, 18.4%; Academic, 16.5%), the outstandingly higher percentage of the Newspapers category in the distribution of CA3s suggests some characteristics peculiar only to them. Why are CA3s so frequent in newspapers? The answer to this question will be revealed in Sections 2.3 and 2.4.

2.3. Frequent Three-element CAs and the Nouns they Modify

Let us examine individual CA3s which appear with high frequency. The 20 most frequent CA3s, the instances of nouns they modify, their frequencies per million words, and their percentages of the total tokens are summarized in Table 3.

Among the top 20 ranked CA3s shown in Table 3, it is evident that Numeral-*year-old* is by far the most frequent. In fact, this is the most frequently occurring CA in the BNC. Its token frequency, 57 per million words, is equivalent to that of words ranked in 2200th place in the written English texts, according to a vocabulary list (Leech et al. 2001: 195). Similarly, other CA3s expressing "age," such as -*month/week/day-old*, also occur frequently. These types of CA3s generally modify human nouns such as *boy*, *girl*, *son*, *daughter*, or people's names. Human nouns are often modified by other top 20 ranked CA3s, including *well-to-do*, *rank-and-file*, and *up-and-coming*.

In terms of form, many CA3s in Table 3 have the pivotal construction of Noun-Preposition-Noun, and the two nouns on each end are identical in almost all cases (e.g., *day-to-day*, *step-by-step*, *one-to-one*, *face-to-face*, *hand-to-hand*, *door-to-door*).

Another conspicuous formation is the coordinative construction Noun-*and*-Noun. For example, *black-and-white*, *rank-and-file*, and *up-and-coming* are listed in the table.

When we look at the middle elements of CA3s, it is important to note that second

Table 3. The 20 most frequent three-element CAs and the nouns they modify

CA3s	PMW	%	Nouns
*-year-old	56.9	32.8	girl, son, daughter, boy, *names*
day-to-day	8.8	5.1	running, basis, work, life
up-to-date	4.0	2.3	information, equipment, evidence, data
*-month-old	3.1	1.8	baby, girl, son, infant, daughter, *names*
step-by-step	2.1	1.2	guide, approach, instruction, change
one-to-one	2.0	1.2	correspondence, basis, relationship
face-to-face	2.0	1.1	contact, meeting, interview
black-and-white	1.3	0.8	image, photograph, television, print
do-it-yourself	1.2	0.7	practice, bar, enthusiast, home, system
down-to-earth	0.9	0.5	approach, attitude, advice, level
well-to-do	0.8	0.5	family, people, house
hand-to-hand	0.8	0.5	combat, fighting, weapon
rank-and-file	0.8	0.4	member, worker, soldier, police
out-of-town	0.7	0.4	site, shopping, centre, superstore
up-and-coming	0.7	0.4	artist, designer, stylist, manager
matter-of-fact	0.7	0.4	way, tone, voice, manner
behind-the-scenes	0.6	0.4	look, work, negotiation, battle
*-week-old	0.6	0.3	baby, boy, son, girl
door-to-door	0.6	0.3	sales, salesman, collection, service
on-the-spot	0.6	0.3	fine, report, treatment, advice

Note: *(wildcard) = any cardinal/ordinal numeral. *Names* = personal names.

elements in CA3s, except for the three age-related CA3s, are function words (*to, of, by, and, the*) and a less important pronoun (*it*). These words carry relatively unimportant semantic contents, and thus, they are phonologically unstressed.

In terms of meaning, it is noticeable that many frequently occurring CA3s are idiomatic expressions derived from metaphors. These include *up-to-date* (i.e., latest), *do-it-yourself*, *down-to-earth* (practical or realistic), *well-to-do* (prosperous), *rank-and-file* (ordinary, non-executive, and non-managerial), and *up-and-coming* (aspiring and promising).

Thus, the analysis shows that frequently occurring CA3s share several formal, phonological, and semantic characteristics.

2.4. Formal and Semantic Patterns

In Section 2.3, we saw that the 20 most frequently occurring CA3s tend to be idiomatic, set phrases. However, a detailed examination of all CA3s in the sample reveals that some can be classified into many more formal and semantic patterns. Table

4 shows these patterns along with examples, token frequency per million words, and percentage of the total number of tokens.

Table 4. Typical formal and semantic patterns of three-element CAs

	PMW	%
Numerals (age/size/frequency/ratio)		
Num.-year/month/week /day/hour ***-old***	61.0	36.0
Num.-cm/km/inch/foot ***-long/high/wide/deep***	1.7	1.0
Num.***-a-***week/year/night	1.0	0.6
Num.***-a-***side/share	1.2	0.7
Range/Manner		
NOUN$_1$-*to*-NOUN$_1$ day-to-day face-to-face hand-to-hand door-to-door wall-to-wall person-to-person **NOUN$_1$-*to*-NOUN$_2$** floor-to-ceiling rags-to-riches hand-to-mouth dusk-to-dawn	20.5	12.1
NOUN$_1$-PREP-NOUN$_{1/2}$ step-by-step case-by-case year-on-year matter-of-fact end-of-year/term tongue-in-cheek value-for-money	12.5	7.4
Comparative		
Comparative***-than-*** more/less/better/ higher/larger -than- average/expected	1.7	1.0
Parataxis		
NOUN$_1$-NOUN$_2$-NOUN$_3$ debtor-creditor-supplier soil-plant-water stress-strain-strength subject-verb-object	0.6	0.4
steering-wheel-type		

	PMW	%
Time/Manner/Location		
PREP-*the*-NOUN above-the-line behind-the-scene on-the-spot on-the-job over-the-top under-the-table	3.9	2.3
***off-the-*NOUN** off-the-shelf/peg/job /cuff/ball/shoulder /wall/record	1.6	0.9
PREP-PREP-NOUN up-to-date down-to-earth out-of-court/school/season down-at-heel	6.9	4.1
Coordination		
-and- -or- bed-and-breakfast cause-and-effect trial-and-error life-or-death, more-or-less make-or-break all-or-nothing	9.1	5.4
Intensifiers/Downtoners		
all-too-common/brief ***ever-so/more-***hot ***oh-so-***ordinary/clever	0.8	0.5
not-too/so/much/quite/very- not-too/so-distant not-so/very-good	1.3	0.8
Difficulty/Near-future		
ADJ-*to*-VERB ***easy/difficult-to***-use ***about-to***-open ***ready-to***-wear	0.7	0.4
ADV-*to*-VERB ***soon-to***-be ***yet-to***-come	0.2	0.1

In the classification, formal patterns are given higher priority than semantic ones, because meanings cannot be unified in some formal patterns, whereas forms give accurate clues. Let us explain the findings illustrated in Table 4 by each category.

[1] Numeral expressions: age, size, frequency, and ratio

The most significant finding shown in Table 4 is that the age-expressing CA3s account for 36% of the total CA3s, making them the large group by far. As mentioned in the previous sections, Numeral-*year-old* is the most frequent CA in English. Other numeral formations include "size" expressions and Numeral-*a*-Noun formations referring to "frequency" or "ratio." Examples of these expressions are *four-foot-high wall* and *20-a-day smoker*, *once-a-year lottery*, and *six-a-side soccer*. Together, numeral expressions account for 38.3% of all CA3s in the sample. Given that these kinds of numeral information, "age" in particular, are indispensable to reports on people or objects in newspaper articles, it follows that the newspaper genre needs many more CA3s of these types than other genres do. Accordingly, this answers the question raised in Section 2.2 about the dominance of the newspaper genre in the distribution of CA3s.

[2] Noun-Preposition-Noun constructions

The next important point is that Noun-Preposition-Noun constructions constitute the second largest percentage (19.5%) of CA3s. In particular, *to* accounts for more than 60% of all prepositions placed in the middle of three elements. Nouns at both ends combined by *to* can be identical (e.g., *day-to-day*, *face-to-face*, *door-to-door*, *wall-to-wall*, *person-to-person*) or different (e.g., *floor-to-ceiling*, *rags-to-riches*, *text-to-speech*, *hand-to-mouth*, *dusk-to-dawn*). In both cases, they generally represent a kind of "range" which is equivalent to "from one place/point/period to another." These types of CA3s are not necessarily idiomatic, closed-class expressions. Similarly, prepositions such as *by*, *on*, *of*, *in*, or *for* can connect a variety of nouns, and their right- and left-side nouns can be either identical or different.

[3] Prep-Prep-Noun and Prep-*the*-Noun constructions

Two kinds of construction headed by a preposition are also frequent formal patterns, namely, Prep-Prep-Noun (4.1%) and Prep-*the*-Noun constructions (3.2%). These patterns take a variety of prepositions and nouns. Some examples of the former pattern are *up-to-date*, *down-to-earth*, and *out-of-court/school/season*, while examples of the latter are *off-the-shelf*, *on-the-job*, *over-the-top*, and *under-the-table*. As indicated in the second column in Table 4, the *off-the*-Noun pattern occurs more frequently and takes

various nouns, which solely accounts for one-third of Prep-*the*-Noun patterns. From the semantic viewpoint, some patterns of this category are metaphoric set phrases, while others can be interpreted with literary meanings.

[4] Coordination: *-and-*, *-or-*

The third largest percentage (5.4%) of CA3s is taken by coordinate constructions, where right and left elements are connected by the conjunctions *and* or *or*. Many of these types are idiomatic expressions (e.g., *bed-and-breakfast*, *cause-and-effect*, *trial-and-error*, *question-and-answer*, *all-or-nothing*, *more-or-less*), but any kind of word combination seems possible as long as it is syntactically, semantically, and pragmatically acceptable (e.g., *per-or-better round*, *farther-and-son relationship*, *green-and-white palace*).

[5] Intensifiers and downtoners

Some fixed two-word units function to emphasize or tone-down the degree that is represented by the right-most element in the CA3. The most frequent intensifier is *all-to-*, as in *all-too-common attitude* or *all-too-brief career*. Other intensifiers are *ever-so/more-* and *oh-so-*, as in *ever-so-thin ribs* and *oh-so-ordinary melody*.

By contrast, several formal patterns headed with *not* can be subsumed into a single category that functions as a downtoner. In this formation, an intensifier such as *too*, *so*, *much*, *quite*, or *very* is negated by *not*, and the first two elements serve to tone down the degree of the subsequent adjective. Although the percentages of intensifiers and downtoners are small (0.5% and 0.8%, respectively), this category is distinctive owing to the convenient functions of the first two-element unit and the variety of the third element.

[6] Comparatives: *-than-*

About one percent of CA3s contain comparative forms, namely *more/less/better/higher/larger/smaller*, which are mostly followed by *-than-expected/average/usual*. For instance, *better-than-expected* and *higher-than-average* appear often, and *larger-than-life* and *lighter-than-air* occur less frequently.

[7] Difficulty and near-future: Adjective/Adverb-to-Infinitive Verb

This formal pattern contains a *-to*-Infinitive Verb in the last two elements. It accompanies a left-most element which falls into two types: (1) adjectives which refer to difficulty (i.e., *difficult*, *easy*), and (2) time-related adverbs or adjectives which refer to the temporal phase when something has not happened but it is expected to happen

in the near future (i.e., *about*, *ready*, *soon*, *yet*). Some examples of (1) are *easy-to-use menu*, *difficult-to-let estates*, and some of (2) are *about-to-burst buds*, *ready-to-eat food*, *soon-to-be parent*, and *yet-to-be father*. Although its percentage amounts only to 0.5% of the total CA3s, this pattern is conspicuous among CA3s as well, because many types of verbs come after these fixed patterns of initial two elements.

[8] Parataxis: $NOUN_1$-$NOUN_2$-$NOUN_3$

The nominal triplet, in which three semantically or phonologically symmetrical nouns occur in line, is another salient pattern available to CA3s. Its percentage of the total CA3s is small (0.4%), but it is productive in that many types of nouns can be arranged in this formation. Another point to note is that an example at the bottom of Table 4, *steering-wheel-type*, can be broken down into two components as [[*steering-wheel*]-*type*], though it may look like a triplet at first glance. Similar dual structures can be found in other CA3s, though they are few in number.

In this section, we have identified the eight types of formal and semantic patterns. Among all the types, numeral expressions, Noun-Prep-Noun formation, and coordination were already observed in Section 2.3, which listed the 20 most frequent CA3s. On the other hand, the other types could not have been identified without access to the large-scale English corpus.

3. Summary

The main findings of this study are as follows. First, CA3s account for a small percentage of all CAs (approximately 6%), while the frequency of total CA3s (169.5 PMW = top 650 words) is rather high. In addition, *-year-old* (57 PMW = top 2500 words) is extremely frequent, particularly in Newspapers. Given that most common 3000 words are adopted for the descriptions in major learner's English-English dictionaries, we should not regard CA3 as a rare phenomenon.

Secondly, frequently occurring CA3s are generally metaphoric idioms, as described in Section 2.3, and an overview of the large sample provided by the BNC reveals at least eight types of formal and semantic patterns. These types are (1) numeral expressions related to age, size, frequency, and ratio; (2) pivotal Noun-*to*/Prep-Noun constructions with range or manner meanings; (3) Prep-*the*/Prep-Noun constructions; (4) coordination; (5) intensifiers and downtoners such as *all-too-* or *not-so/too/very-*; (6) comparative forms such as *more/better/higher/larger-than-*; (7) Adjective/Adverb-*to*-Infinitive Verb constructions expressing difficulty or the near future such as *easy/difficult-to*-Inf. or *soon/yet-to*-Inf.; and (8) Noun-Noun-Noun parataxis. As none of

the eight patterns is a closed-class category, a variety of combinations of elements is possible, or at least one element of a CA3 is selective, and the patterns enable English speakers to construct a variety of CA3s.

As these results have been obtained by the analysis of a British English corpus, in order to generalize and verify the findings, further research needs to be conducted using other British English corpora as well as English corpora of other regions.

References

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

Conti, S. 2006. *Compound Adjectives in English: A Descriptive Approach to Their Morphology and Functions*. Doctoral Dissertation. University of Pisa.

Leech, G, P. Rayson, and A. Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow, England: Pearson Education.

Mizuno, O. 1993. "A Survey of Hyphenated Compound Adjectives of the Type 'N+A' and Their Structural Analysis: On the Business Articles in Time and Newsweek." *Fuji Phoenix Review* 1: 17–40.

Corpus

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