Journal of English Teaching
e-ISSN: 2622-4224 | p-ISSN: 2087-9628
JOURNAL OF ENGLISH TEACHING http://ejournal.uki.ac.id/index.php/jet

Volume 12. Number 3, October 2026, pp, 473-490

Frequent Semantically Non-Compositional Multiword Expressions in the English Subtitles of Military and War Films and Dramas: A Corpus-Based Study

Wenhua Hsu[1]*

Article History:
Received: 09/08/2026
Revised : 18/09/2026
Accepted: 21/10/2026
Available Online: 29/19/2026

Keywords:
multiword expressions, semantic non-compositionality, English subtitles, military films and dramas

ABSTRACT
The expansion of over-the-top (OTT) streaming has popularized binge-watching television dramas. This study explores the potential for English lexical development among Taiwanese EFL college students who switch from Mandarin to English subtitles, focusing on military and war dramas as a source of extensive language exposure. Because military English features technical terms and semantically non-compositional multiword expressions (MWEs) that are underrepresented in general EFL materials, even proficient learners often struggle to comprehend military-themed audiovisual content. To address this gap, this study compiled a 60-million-word corpus from the English subtitles of 100 military drama series and 100 films. Using six selection criteria, the study identified 515 opaque MWEs of two to six words. To ensure the validity of the selected MWEs, human judgment was combined with corpus-based tools and Gemini Notebook for AI-assisted qualitative analysis. Most of these expressions consist of words from the 3,000 most frequent English word families, providing a practical, self-directed learning resource to improve comprehension of military-themed media.

[1] I-Shou University, Taiwan. Email: iwan whh@isu.edu.tw *Corresponding author
https://orcid.org/0000-0001-5168-3233

INTRODUCTION

The popularity of television drama series has grown alongside the rapid expansion of over-the-top (OTT) streaming services and Internet-connected mobile devices. In Taiwan, where Mandarin is the official language and English is taught as a compulsory foreign language (EFL), college students increasingly encounter English through digital media as well as formal instruction. The researcher-teacher has observed that some male college students frequently binge-watch military and war dramas and discuss recently viewed series and films. Although all Taiwanese men are required to complete compulsory military service, students' interest in this genre is not necessarily related to that obligation. Nevertheless, the sustained viewing encouraged by military and war dramas may provide an additional source of extensive language exposure beyond the graded readers commonly-used in EFL instruction.

The potential of audiovisual content as a source of second language (L2) input is particularly relevant in the context of OTT streaming. Streaming platforms provide access to television series and films produced in diverse languages and countries. When learners select English subtitles, they can encounter substantial amounts of written English alongside spoken dialogue, regardless of whether the original audio is in English or another language (Locher & Messerli, 2020; Pedersen, 2019). In the present study, subtitles broadly refer to L2 subtitles accompanying either L2 or L3 audio, including what Talaván Zanón (2006) calls bimodal subtitles, in which both the audio and subtitles are presented in the same foreign language.

This study is particularly concerned with the possibility that students who normally watch audiovisual content with Mandarin subtitles could switch to English subtitles and thereby increase their exposure to L2 lexical input. This possibility is especially relevant to students who regularly binge-watch television series and films for leisure. Binge-watching, which became widespread with the growth of streaming platforms, enables viewers to watch multiple episodes in a single sitting or engage in sustained "serial viewing" over several days or weeks (Rubenking & Bracken, 2021). Such viewing may provide repeated opportunities to encounter vocabulary and multiword expressions (MWEs) in meaningful contexts.

Military English is of particular interest because it differs from general English in its lexical and phraseological features. Military and war dramas and films contain technical terms, formulaic expressions, and semantically non-compositional multiword expressions (MWEs) that may be underrepresented in general EFL materials. Consequently, even learners with adequate general English proficiency may struggle to comprehend military-themed audiovisual content. Although knowledge of individual words is essential for understanding such content, recognizing the component words of an expression does not necessarily ensure comprehension of its conventional meaning. Identifying frequent semantically non-compositional MWEs may therefore provide a more comprehensive account of the lexical and phraseological input available through English subtitles.

Building on the author's (2023) examination of semantically non-compositional MWEs in the English subtitles of Korean dramas (K-dramas), the present study extends this line of inquiry to military and war dramas and films. It investigates the frequency, distribution, and characteristics of semantically non-compositional MWEs that learners may encounter through binge-watching English-subtitled audiovisual content. These expressions are treated as units that function as single semantic entities and whose meanings cannot always be derived directly from those of their individual components. By identifying and analyzing such

expressions, the study aims to contribute to a better understanding of the lexical and phraseological input available to EFL learners through audiovisual media. The study addresses the following research questions:

1. What are the most frequent semantically non-compositional MWEs occurring in the English subtitles of military and war dramas and films?
2. What vocabulary levels do these MWEs represent?
3. What semantic and linguistic characteristics are exhibited by the most frequent MWEs?

LITERATURE REVIEW

Audiovisual Input and Vocabulary Learning

The widespread availability of streaming platforms has made television programs and films increasingly accessible sources of authentic second-language (L2) input. This is particularly valuable for English as a foreign language (EFL) learners, who may have limited opportunities to encounter English outside the classroom. Audiovisual materials expose learners to vocabulary in meaningful contexts and, when subtitles are used, provide simultaneous access to spoken and written language.

Webb and Rodgers (2009a) examined the vocabulary demands of 88 American and British television programs using British National Corpus frequency lists. They found that approximately 2,000–4,000 word families were needed for 95% lexical coverage, while 5,000–9,000 were required for 98%, depending on genre. For films, Webb and Rodgers (2009b) found that about 3,000 word families, including proper nouns, provided 95% coverage and approximately 6,000 provided 98% coverage. Nation (2006) proposed 98% coverage as a useful level for inferring unfamiliar words from context. These findings indicate that audiovisual materials provide substantial lexical input but can also impose considerable vocabulary demands.

However, lexical coverage is normally calculated at the individual-word level and may overlook the difficulty of multiword expressions (MWEs). Learners may know the words within an expression but still fail to understand its conventional meaning. This issue is especially relevant to English-subtitled audiovisual content, where learners encounter large amounts of written English alongside visual and auditory information. Thus, examining MWEs in addition to individual words can provide a more complete picture of the lexical input available to EFL learners.

Narrow Viewing and Repeated Vocabulary Exposure

Repeated exposure plays an important role in incidental vocabulary learning. Research suggests that a single encounter rarely leads to strong learning, whereas multiple encounters may be necessary for meaningful uptake (Horst, 2013; Horst, Parsons, & Bryan, 2011). Nation (2014) suggested that approximately five to sixteen encounters may be needed for vocabulary learning and proposed around twelve encounters in different contexts as a useful benchmark. However, these figures should be understood as general guidelines rather than fixed requirements for learning individual lexical items.

This principle is closely related to Rodgers and Webb's (2011) concept of narrow viewing, in which learners watch related programs or multiple episodes of the same series. Their findings showed that related programs contained fewer words occurring only once and

more low-frequency words occurring repeatedly than unrelated programs. This suggests that viewing connected programs can reduce lexical burden while providing more opportunities for repeated exposure.

Narrow viewing is particularly relevant to binge-watching, as viewers repeatedly encounter the same characters, settings, topics, and storylines. For learners who regularly watch military and war dramas, such viewing may provide repeated exposure to specialized vocabulary and expressions. The present study extends this perspective from individual words to non-compositional multiword expressions (MWEs). Because the meanings of non-compositional MWEs cannot be fully predicted from the meanings of their individual components, these expressions may need to be learned as lexical units, much like single words. It is therefore reasonable to consider repeated occurrences of such expressions when identifying candidates for detailed analysis.

Following Nation's (2014) recommendation of approximately twelve encounters as a general benchmark for vocabulary learning, the present study adopts twelve occurrences as a pragmatic cut-off point for selecting frequent non-compositional MWEs from a 60-million-token corpus. This threshold is an operational choice informed by single-word vocabulary research, rather than an empirically established minimum for MWE learning. The recommendation for individual words does not directly establish an equivalent threshold for multiword expressions, and the selected cut-off should therefore not be interpreted as a universal learning requirement. Instead, it provides a consistent and transparent criterion for narrowing the corpus data to expressions that offer repeated exposure opportunities. Frequency alone, however, cannot establish whether a sequence constitutes a meaningful MWE; semantic non-compositionality and structural characteristics must also be considered.

Multiword Expressions and the Illusion of Comprehension

Language consists not only of individual words but also of recurring combinations that may function as single lexical units. Hinkel (2023) defines MWEs as recurrent word combinations that are remembered and used as single lexical items. Related terms in the literature include collocations, lexical bundles, clusters, formulaic sequences, sentence stems, prefabricated units, n-grams, and frozen phrases (Wood, 2020). Idioms, phrasal verbs, proverbs, and binomial expressions can likewise be regarded as particular types of multiword expressions.

The present study uses MWE as an umbrella term for combinations that vary in semantic compositionality and syntactic fixedness. It focuses specifically on semantically non-compositional MWEs, whose meanings cannot be readily inferred from their individual words. Such expressions may be particularly problematic for EFL learners.

Conventional lexical coverage measures may overlook this problem because they treat individual words as the basic units of analysis. Consequently, learners may recognize all the words in an expression while misunderstanding its overall meaning. Martinez and Murphy (2011) refer to this problem as an "illusion of comprehension." For example, knowing *war, effort, Oscar, Mike, no* and *copy* does not necessarily mean that a learner understands *war effort*, *Oscar Mike* or *no copy*. Similarly, expressions such as *get your six*, *about face turn*, and *house to house job* may remain opaque despite familiar component words.

Identifying meaningful MWEs therefore requires more than frequency-based extraction. Corpus tools can produce numerous n-grams, but many are syntactically

incomplete or lack independent semantic value (Simpson-Vlach & Ellis, 2010). Researchers have consequently combined frequency with measures such as mutual information, grammatical well-formedness, expert judgment, and semantic non-compositionality (Ackermann & Chen, 2013; Martinez & Schmitt, 2012; Shin & Nation, 2008).

Overall, the literature indicates that frequency is necessary but insufficient for identifying useful MWEs. The present study therefore combines corpus-based frequency analysis with semantic and structural evaluation, focusing on non-compositional expressions that may remain difficult despite familiar constituent words.

Military and War Dramas and Films as a Source of Specialized Lexical Input

Military and war dramas and films provide a distinctive form of authentic English input because they frequently depict combat, military operations, commands, institutional interactions, and conflict. Consequently, they may contain specialized vocabulary and recurring expressions that are less common in general audiovisual materials.

For frequent viewers, repeated exposure may facilitate learning of these specialized expressions through meaningful contexts. However, military language can also create comprehension difficulties because many expressions are conventionalized, context-dependent, or semantically non-compositional. A learner may understand most individual words in a military drama while still misunderstanding expressions that convey important meanings.

The present study builds on the author's (2023) investigation of semantically non-compositional MWEs in the English subtitles of K-dramas, which demonstrated that subtitle-based analysis can reveal lexical information not captured by individual-word frequency alone. Nevertheless, relatively little research has examined genre-specific MWEs in audiovisual materials, particularly in specialized genres such as military and war dramas.

This gap is significant because military and war content is increasingly accessible through streaming platforms. Existing audiovisual vocabulary research has focused largely on lexical coverage, individual-word frequency, and repeated exposure, whereas MWE research has often examined general, spoken, or academic corpora. Investigating frequent non-compositional MWEs in military and war dramas and films can therefore connect these research areas and provide a more comprehensive account of the lexical input encountered by learners.

Corpus-Based Identification of MWEs and AI-Assisted Qualitative Analysis

Corpus linguistics provides systematic methods for identifying recurrent lexical patterns through frequency, distribution, and concordance analysis. However, quantitative tools alone cannot determine whether a frequent sequence constitutes a semantically unified and pedagogically useful MWE. Human judgment is therefore necessary to evaluate semantic coherence, grammatical completeness, and potential learning value.

Recent developments in artificial intelligence (AI) offer additional support for qualitative corpus analysis. Labrador (2026) examined NotebookLM, now known as Gemini Notebook, and found it more useful for qualitative than quantitative analysis. The tool can analyze user-uploaded materials, summarize content, identify key terms and themes, and respond

interactively to questions about meanings and relationships. However, it is not designed to replace conventional corpus tools for quantitative tasks such as frequency counting.

Following this complementary approach, the present study uses corpus analysis as the primary method for identifying candidate MWEs through frequency and distributional evidence. Gemini Notebook is used only as an auxiliary tool to support qualitative examination of semantic relationships, contextual patterns, and discourse characteristics. Final decisions regarding the identification and classification of semantically non-compositional MWEs remain based on human judgment.

Overall, previous research shows that audiovisual materials provide rich and repeated lexical input, but individual-word frequency alone may not fully represent the comprehension demands of such input. Non-compositional MWEs can remain problematic even when their constituent words are familiar, and military and war dramas may contain a particularly rich set of specialized expressions. Therefore, the present study combines corpus-based frequency analysis, semantic and structural evaluation, and AI-assisted qualitative analysis to identify frequent non-compositional MWEs in the English subtitles of military and war dramas and films and to examine their potential relevance to L2 vocabulary learning and teaching.

RESEARCH METHODS

The Corpus

The present corpus comprised 100 military and war drama series and 100 military and war films with high viewership ratings on OTT media services. Table 1 presents a selection of the military and war drama series and films included in the corpus.

Table 1. A Selection of Military and War Drama Series and Films Included in the Corpus

Film titles	Release year	Length	Drama titles	Release year	Number of episode/Length
The Tiger	2025	116 min.	Boots	2025	8 eps/40-50 min.
The Last Front	2024	98 min.	Masters of the Air	2024	9 eps/ 48-77 min.
Warhorse One	2023	126 min.	A Small Light	2023	8 eps./40-56 min.
All Quiet on the Western Front	2022	148 min.	Halo S1 S2	2022-2024	17 eps/40-61 min.
Operation Mincemeat	2021	128 min.	Das Boot S1 to S4	2018-2023	32 eps/ 45-63 min.
Da 5 Bloods	2020	154 min.	Seal Teams S1 to S7	2017-2024	114 eps/ 38-49 min.
Greyhound	2020	91 min.	World on Fire S1 S2	2019-2023	13 eps/55-60 min.
The Forgotten Battle	2020	124 min.	The Liberator	2020	4 eps/45-56 min.
The Last Full Measure	2019	116 min.	Catch-22	2019	6 eps/42-47 min.
A Private War	2018	110 min.	Strike Back S1 to S8	2010-2020	76 eps/43-53 min.

The English subtitles, in SubRip Text (SRT) format, were downloaded from the Internet for research purposes. SRT files are plain-text files that contain subtitles in sequential order together with

their corresponding start and end time codes. Each subtitle segment typically consists of one or two lines, with a maximum of seven or eight words per line. Table 2 presents two excerpts in SRT format, one from the film *Warhorse One* and the other from *SEAL Team* Season 7, Episode 10.

Table 2. Two Excerpts in SRT Format from a Military Film and a Military Drama Series

The film *Warhorse One*	***Seal Team* Season 7, episode 10**
65	43
00:03:55,250 --> 00:03:56,541	00:04:57,451 --> 00:04:59,241
All right, I need that intel, Mike.	Well, congrats, Davis.
66	44
00:03:56,666 --> 00:03:57,791	00:04:59,417 --> 00:05:00,997
What do you got for me?	You just delivered a knockout punch
67	45
00:03:59,708 --> 00:04:00,708	00:05:00,998 --> 00:05:03,078
Roger that.	to China's grip in the Southern Hemisphere.
68	46
00:04:00,833 --> 00:04:02,500	00:05:03,113 --> 00:05:05,706
All told, I've got six of them traveling...	Yeah, how much more chest candy
69	is a Narco State leader worth?
00:04:02,625 --> 00:04:04,208	47
Yeah. Hit me.	00:05:05,707 --> 00:05:08,467
70	It was a total team effort,
00:04:04,333 --> 00:04:06,416	and it was a...
Logan. Make sure you get all the drives.	48
71	00:05:08,627 --> 00:05:10,887
00:04:10,458 --> 00:04:11,875	hell of a way to go out, Warrant Officer Perry.
The Reverend Mathew William Walters,	49
72	00:05:11,035 --> 00:05:12,955
00:04:12,000 --> 00:04:14,750	Gonna miss rolling with you, Ray.
white male, age 45.	

The Procedures

AntConc 4.4.2 (Anthony, 2026a) was used to retrieve multiword expressions (MWEs) from the 60-million-token English subtitle corpus. The span parameter was set to between two and six words because the frequency of MWEs drops drastically as their length increases to five words or more (Hyland, 2008). As noted earlier, learning rarely occurs after a single encounter. As the number of encounters with an unknown word or chunk increases, so does the likelihood of learning it. Following Nation (2014), the researcher adopted a minimum frequency threshold of 12 occurrences and identified expressions that occurred frequently and were widely distributed throughout the corpus.

Two initial criteria were applied to generate the candidate pool:

- Frequency criterion: At least 12 occurrences in the entire corpus.
- Range/dispersion criterion: Occurrence in at least 100 of the 200 productions.

Since one of the goals of the study was to identify MWEs commonly used in the English subtitles of military and war dramas and films, expressions occurring with very high frequency in only one or two titles were excluded. Specifically, when different inflectional forms of an expression were considered together, the expression had to occur in at least half of the military dramas and films in the corpus (i.e., at least 100 of the 200 productions, corresponding to a dispersion of at least 50%). Although this threshold was admittedly

arbitrary, it was considered appropriate because the study aimed to identify expressions widely used across the corpus rather than those concentrated in a small number of titles.

Applying the frequency and dispersion criteria yielded approximately 53,900 candidate MWEs. This initial pool was not the final list of learnable expressions. It contained many recurrent word combinations that did not convey a specific meaning as a unit, as well as combinations that crossed phrasal boundaries and were therefore less suitable for learning as single units. Further filtering was thus necessary to distinguish potentially learnable MWEs from frequent but less meaningful word combinations.

Meaningfulness was a key selection criterion: recurrent MWEs had to convey meaning and be learnable as single units, making them comparable to a list of individual words. Before manual evaluation, Mutual Information (MI) was used to filter out relatively weakly associated word combinations. According to Hunston (2002), multiword combinations with an MI value greater than 3 are considered strongly cohesive. Therefore, candidate MWEs with an MI value below 3 were excluded at this stage. The 53,900 candidate MWEs were reduced to 35,923 at this stage, representing approximately two-thirds of the original pool.

Drawing on Martinez and Schmitt (2012) and Shin and Nation (2008), the researcher formulated six questions to guide the evaluation of candidate MWEs. These questions assessed meaningfulness (Q1), phrasal boundaries (Q2), and semantic non-compositionality (Q3–Q6).

Criterion 1. Meaning: Q1. Does the candidate MWE convey a specific meaning?
Criterion 2. Phrasal boundaries: Q2. Does the candidate MWE span two phrasal boundaries?
Criterion 3. Non-compositionality: Q3. Can the meaning of the candidate MWE be derived solely from the core meanings of its individual component words?
Criterion 4. Multiple meanings: Q4. Does the candidate MWE have more than one meaning?
Criterion 5. Polysemous component: Q5. Does the candidate MWE contain at least one word with more than one meaning?
Criterion 6. Derivational morphology: Q6. Does the candidate MWE contain at least one word with a derivational affix that changes the meaning of its base form?

Criteria 1–3 were used to determine whether an expression qualified as a candidate MWE, whereas Criteria 4–6 served as separate classification criteria to further characterize the selected expressions. Using Q1–Q6, the researcher and two colleagues independently evaluated approximately 35,923 candidate MWEs that met the frequency and dispersion criteria and had an MI value greater than 3. A 'yes' answer to either Q2 or Q3 disqualified the candidate MWE. In particular, applying Criterion 2 reduced the number of candidate MWEs from 35,923 to approximately 6,000.

Responses of yes, not sure, and no were coded as 1, 0.5, and 0, respectively, to create a three-point rating scale. When no consensus was reached, a candidate MWE was provisionally retained pending further examination. Cohen's kappa coefficients were then calculated to assess inter-rater reliability for each question. All kappa values exceeded 0.80, indicating substantial agreement among the three raters.

To evaluate meaningfulness and well-formedness, the candidate MWEs were assessed using Q1 and Q2, respectively. Semantic non-compositionality and related lexical-semantic properties were then examined using Q3–Q6. Polysemous MWEs, or MWEs containing constituent words with multiple meanings, may present difficulties for learners. Examples of

the former include *dust off, get out of here*, *keep your head*, *watch your six*, *weapons free now*, *suit up, guys*, *we have visual contact*, *this is not a drill*, and *target acquired at once*. Examples of the latter include *theater* in *theater of war, magazine* in *forward magazine, bow* in *angle on bow*, *extraction* in *extraction point is secure*, *close* in *enemy is closing in*, *classified* in *classified information*, *security* and *clearance* in *security clearance*, *ass* in *move your ass*, *frosty* in *stay frosty* and *secure* in *secure the perimeter*. Questions 4 and 5 were therefore used to identify MWEs with multiple meanings, either at the level of the entire expression or within one of its constituent words.

It was also found that some candidate MWEs containing words with derivational affixes could mislead learners into forming incorrect form–meaning associations. Learners may believe they know the base form of a word without realizing that the meaning of its derived form has changed. Examples include *grind* versus *ground forces*, *arm* versus *armed forces*, *brief* versus *mission briefing*, *fall* versus *fallen soldiers*, and *close* versus *is closing in*. Question 6 was therefore included as an additional screening criterion.

Following the successive filtering and evaluation stages, 515 MWEs were retained in the final list. Thus, the initial pool of approximately 53,900 candidates was substantially narrowed through statistical filtering and manual evaluation to produce the final set of expressions aligned with the study's focus on meaningful, learnable, and widely distributed MWEs in military and war drama and film subtitles.

Data Processing

For pedagogical purposes, several modifications were made during the compilation of semantically non-compositional multiword expressions. First, expressions occurring in different inflectional forms (e.g., plural nouns; third-person singular present-tense verbs ending in *-s*; present participles/ gerunds ending in *-ing*; and simple past or past participle forms ending in *-ed*) were consolidated under a single lemma. The frequencies of all inflected forms were summed to obtain a cumulative frequency, with the lemma serving as the representative entry. For example, *bug out* (total frequency = 2,351) was derived by combining the frequencies of *bug out* (1,437 occurrences), *bugs out* (18), *bugged out* (364), and *bugging out* (532). This approach was adopted to reduce the number of lexical entries that learners need to process. Once learners become familiar with the core meaning and usage of an expression, the additional inflectional forms can be acquired through subsequent language exposure.

In addition to presenting expressions in their lemma forms, two further modifications were made. First, entries that appeared in British English were converted to their American English equivalents to ensure consistency. Second, an optional copular *be* was added to participle + preposition combinations when they can function as predicates (e.g., *(be) closing in, (be) smoked by* and *(be) advised*).

Another major revision concerned partially overlapping multiword expressions. A partial overlap refers to a situation in which a shorter word string is subsumed within a longer one, with each functioning as an independent meaningful unit, as in the case of *all ahead* in *all ahead full*. The phrase *all ahead* appeared 4,694 times, while *all ahead full* occurred 2,122 times. Although the two expressions differ in meaning, neither is semantically compositional. In order to obtain a more accurate frequency figure for *all ahead*, the frequency of *all ahead*

full was subtracted from that of *all ahead* (4,694 − 2,122 = 2,572 for *all ahead*). Since the two phrases can each stand alone as independent units, they were compiled separately in the present MWE list.

The final consideration concerned expressions with multiple senses. To identify the meanings of polysemous expressions, each entry was cross-referenced with *The Free Dictionary*, which draws on the *Cambridge Dictionary of American Idioms* and the *Cambridge International Dictionary of Idioms*. For expressions in which only one sense is non-literal, it was not feasible to determine the frequency of the non-literal sense in the 60-million-token corpus. Corpus analysis tools can calculate the total frequency of a given form, but they cannot reliably distinguish among its senses based solely on concordance lines or determine the frequency of each individual sense. For example, *a hell of* (used to intensify either a positive or a negative evaluation) occurred 3,214 times in the corpus. Because these occurrences included both positive and negative non-literal meanings, only the overall frequency of the expression was recorded rather than the frequency of each individual sense.

To sum up, the selection of semantically non-compositional multiword expressions for inclusion in the self-directed learning list followed this sequence: (1) frequency (a minimum of 12 occurrences for initial screening), (2) range (occurrence across at least half of the military and war dramas and films in the corpus), (3) lexical cohesiveness (MI > 3), (4) meaningfulness (Q1), (5) grammatical well-formedness (Q2), (6) semantic non-compositionality (Q3), and (7) non-transparency (Q4, Q5, and Q6).

Using Gemini Notebook

Following the aforementioned procedure, we identified a set of candidate multiword expressions whose overall meanings could not be derived compositionally from the meanings of their individual components. However, this candidate set still included general-purpose multiword expressions rather than expressions specifically characteristic of the military register (e.g., *here we go, a couple of, as long as, hang on, come on, what the hell is, and nothing to do with*) as opposed to military-specific expressions (e.g., *pink mist cloud, Turkey shoot, lock and load, Charlie Mike, visual contact confirmed, right full rudder, copy that, zero, roger*).

To further refine the dataset, the candidate multiword expressions were entered into Gemini Notebook with the following prompt:

"Based on the list of candidate semantically non-compositional multiword expressions, identify those that are characteristic of the military or war-related register. Classify the identified expressions according to the number of words they contain into two-word, three-word, four-word, five-word, and six-word expressions. In addition, categorize these register-specific multiword expressions according to the military contexts in which they are used."

Gemini Notebook was used as an auxiliary tool to assist in the classification of candidate expressions and their thematic grouping according to military contexts. Its suggestions were not accepted uncritically; instead, human expert judgment served as the final arbiter in determining whether each expression was semantically non-compositional and genuinely characteristic of the military or war-related register. This manual review helped ensure the linguistic validity and contextual appropriateness of the final dataset.

RESULTS AND DISCUSSION

Frequent Non-Compositional Multiword Expressions in the English Subtitles of Military and War Dramas and Films

A total of 515 semantically non-compositional expressions consisting of two to six words were ultimately selected to form the list of frequent non-compositional MWEs in military and war films and drama series (see Appendix). The list comprises 212 two-word, 145 three-word, 77 four-word, 74 five-word and 7 six-word non-compositional expressions commonly used in military films and dramas.

The Vocabulary Level of the Military-Specific Non-Compositional MWEs

The vocabulary analysis program AntWordProfiler (Anthony, 2026b) was used to examine the vocabulary levels of the individual word tokens in the military-specific non-compositional multiword expressions. The software is equipped with the 25 most frequent 1,000-word-family lists derived from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), ranked according to frequency of occurrence and dispersion in the corpora (Nation, 2017). Table 3 presents a comprehensive overview of the vocabulary levels of the military-specific non-compositional MWEs based on the BNC/COCA word-frequency scale.

The list of 515 non-compositional MWEs consists of 1,625 running words and includes 694 word families. The first 1,000 BNC/COCA word families account for 60.74% of the total words in the list, while the second 1,000 account for 15.45%, followed by the third 1000, which account for 7.02%. The combined coverage of the first 3,000 word families is 83.21%. Beyond the first 5,000 word families, the coverage provided by each additional 1,000 word families rapidly decrease to less than 2%.

Table 3. Vocabulary Levels of the 515 Military-Specific Non-Compositional MWEs

BNC/COCA vocabulary levels	Number of tokens	% coverage in tokens	Cumulative % coverage	Number of word families
1st 1,000	987	60.74%	60.74%	305
2nd 1,000	251	15.45%	76.19%	143
3rd 1,000	114	7.02%	83.21%	68
4th 1,000	57	3.51%	86.72%	40
5th 1,000	41	2.52%	89.24%	31
6th 1,000	13	0.80%	90.04%	10
7th 1,000	23	1.42%	91.46%	14
8th 1,000	18	1.11%	92.57%	12
9th 1,000	26	1.60%	94.17%	8
10th 1,000	9	0.55%	94.72%	5
11th–25th 1,000	29	1.79%	96.51%	25
Off-list	57	3.49%	100%	33
Total	1625	100%		694

As Is Evident In Table 3, A Large Number Of Military-Specific Non-Compositional Mwes Are Composed Of Very General Words, Most Of Which (Specifically, 83.21%) Belong To The First 3,000 Most Frequent Word Families In The BNC/COCA (E.G., *As You Were, Black Flag, Blue On Blue, Nine-Line, Cake Eater, Coming In Hot*). Although These Multiword

Combinations Are Composed Largely Of High-Frequency, Familiar Words, Their Meanings Are Often Difficult To Interpret Without Knowledge Of The Specialized Military Contexts In Which They Are Used.

Semantic And Linguistic Characteristics Of Frequent Military-Specific Non-Compositional Mwes

Generally, Pairings Or Strings Of Content Words (Nouns, Lexical Verbs, Adjectives, And Adverbs), As Well As Pairings Or Strings Of Content Words And Function Words (Determiners, Conjunctions, Prepositions, Pronouns, Auxiliary Verbs, Modals, And Quantifiers), Constitute Two Common Patterns In The Present List. For An Overview, Examples Of Each Pattern Are Presented In Table 4.

Table 4. Structures of Non-Compositional MWEs and Selected Examples

Structure of n-gram	Strings/pairings of content words and content words	Strings/pairings of content words and function words
2-word	how copy, fast air, fifth column, five mikes, forward magazine, Lima Charlie, friendly fire, active duty, ahead flank, ahead two-thirds	eyes on, about face, at ease, bug out, fall back, wait out, wheel up, look out, hang in, hold on
3-word	active fire control, blood spray visual, broken arrow code, danger close ordnance, dark side moon	fish and chips, drop the hammer, in a foxhole, act of war, an ied in, bodyguard of lies, embrace the suck
4-word	contact front engage now, battle ready soldiers unit, high targe value secure,	get off the X, above my pay grade, beef up regiments melted, boots on the ground, Gray wolf is hungry
5-word	Lead elements crossing phase line	We have troops in contact, full rise on bow planes, get out of the way, after the war heads roll
6-word	Slow is smooth, smooth is fast	Keep your heads on a swivel, stay in your three foot world

Most of the opaque two-word expressions (187 out of 212) are lexical collocations (e.g., *black op, zero alpha, Tango Mike, solid copy, Bingo fuel, bad intelligence,* and *soup chips*), as opposed to grammatically conditioned pairs, namely, a content word combined with a function word (e.g., *eyes on, about face*). The two-word grammatical collocations also include a few phrasal verbs (e.g., *wheel up, wait out, blow up, beef up, frag out, call off, call for, fall back, look out,* and *watch out*), as well as some idioms, such as *sour sandwich, deep end, fresh meat, goat rope, dog-faced, close shave,* and *football bat*, and slang expressions containing swear words (e.g., *shit bowl, deep shit,* and *Godforsaken shithole, all clusterfucks*). In contrast, grammatically conditioned strings occur more frequently among three- to six-word expressions than lexical strings. However, three- to six-word expressions also contain quite a few metaphors or idioms (e.g., *Gray wolf is hungry, kick a hornet's nest, lambs to slaughter, fart in wind,* and *fish in barrel*). For a complete picture, see the Appendix.

Moreover, the length of multiword expressions appears to have some influence on semantic transparency. As multiword expressions become longer, their potential for ambiguity and polysemy tends to decrease. However, the present data show that opaque multiword expressions can still be relatively long. One six-word expression, *bend over here*

it comes again (BOHICA), which is composed of the two-word expression *bend over* and the four-word expression *here it comes again*, remains semantically opaque.

Contextual Classification of Frequent Military-Specific Non-Compositional MWEs Using Gemini Notebook

Traditional N-gram extraction tools primarily focus on quantitative analysis. Although they can identify frequently-used multiword expressions in military and war films and drama series, they do not provide information about the meanings of these expressions or the contexts and situations in which they are used. For example, the expression *with all due respect* occurs frequently in military discourse. This phrase is a formal preface used when a subordinate wishes to disagree with a superior. It is often heard in military films and drama series when someone wants to express an opinion in a setting where the chain of command is strictly observed and subordinates are not supposed to overstep their rank when speaking to their superiors. Gemini Notebook can help address these limitations. With appropriate prompts, Gemini Notebook can classify the frequently-used multiword expressions identified from military and war films and drama series according to their contexts of use and provide explanations of their meanings. A few examples are presented in Table 5.

Table 5. Classification of Frequent Military-Specific Non-Compositional Multiword Expressions by Context of Use

Military contexts in which opaque MWEs are used	Non-combinational MWEs	Meaning
Combat and Tactical Operations	Hot weapon	A loaded firearm with the safety off, ready to be fired immediately.
	Watch your spacing	A tactical warning to prevent troops from huddling together, reducing the risk from explosives.
	Wheel up at 2200	The scheduled departure time for an aircraft or mission is 10:00 PM.
	Contact front engage now	An immediate order to fire upon an enemy appearing directly ahead.
	Close quarters	Combat occurring at very short range, often hand-to-hand or in buildings.
	Fire in the hole	A warning that an explosive charge is about to detonate.
Military Law, Training, and Personnel	Chalk One	The first unit or aircraft in a sequence for a mission.
	Dog-faced lifer	Slang for a career soldier who has dedicated their life to the infantry.
	On point lead	Being the lead person or unit in a formation.

Table 5. Classification of Frequent Military-Specific *…(continued)*

Military contexts in which opaque MWEs are used	Non-combinational MWEs	Meaning
Historical & Strategic Expressions	Second front	An additional theater of war opened to force the enemy to divide their resources.

	Theater of war	The entire geographic area involved in a specific conflict.
	Element of surprise	The tactical advantage of attacking an unaware enemy.
	Graveyard of empires	A historical and metaphorical reference to regions (like Afghanistan) where foreign military interventions historically fail.
Technical jargon, specialized Operations and Equipment	Bouncing betties mine	Antipersonnel mines that pop up and explode at waist height.
	Full rudder turn	An order for a vessel to use its maximum turning capability.
	Combat jump insertion	A tactical parachute drop into a contested area.
	Starboard beam area	The region directly to the right side of a naval vessel.
	Stopping power kill shot	The specific capability of a weapon to instantly drop an enemy upon impact.
	Heavy TIC at this time	Reporting "Troops In Contact" under intense fire.
Psychology and Field Slang	Conscience round	A single blank cartridge placed in one rifle of a firing squad to provide psychological doubt for the shooters.
	Charge for France now	A patriotic call to attack the enemy for the sake of the homeland.
	Cannon fodder troops	Soldiers considered expendable and sent to absorb enemy fire.
	Turkey shoot battle	An engagement so one-sided it feels like hunting stationary targets.
	Tango down call	Coded confirmation that an enemy combatant has been neutralized .
	Beef up regiments melted	Reinforcing units that have suffered extreme and unsustainable casualties.
	Contact front and gun fight	Reporting immediate engagement with the enemy.
	Broken arrow, broken arrow	Urgent signal for air help when overrun.
Field Situations and Tactical Language	No man's land	The lethal, empty ground between opposing defensive lines.
	Deep end jump	Being forced into a difficult situation without preparation.
	Phase line red	A map line used to coordinate the timing of an attack.
	Embrace the suck	A common soldier's expression for accepting miserable conditions.
	House to house job	The highly dangerous task of clearing urban buildings one by one.
	Hot weapon safety off base	A firearm that is loaded and ready to fire immediately.
	Draw tanks	To lure enemy armor into a specific trap.
	Mother of all clusterfucks	Slang for a monumental and disastrous military or logistical failure.
Naval and Aerial Specifics	Starboard beam sighting	Spotting an object at 90 degrees to the right.
	U-boat contact	Detection of a hostile submarine in the vicinity.
Counterinsurgency Terms	Regular people regular clothes	Describing insurgents who hide among civilians by not wearing uniforms.

Table 5 provides a snapshot of the occurrences of frequent MWEs across selected military registers. Although the precision of the categorization may be subject to question due to the multilayered functions of some multiword expressions, the preliminary typology generated by Gemini Notebook based on context of use may serve as a useful starting point

for raising EFL learners' awareness of the appropriate register use of these military-specific non-compositional multiword expressions and supporting self-directed learning.

Overall, the frequent MWEs exhibit variation in both semantic and linguistic characteristics. Semantically, they range from relatively transparent expressions to opaque collocations, idioms, metaphors, and military-specific expressions whose meanings depend on their contexts of use. Linguistically, they include lexical collocations, grammatically conditioned combinations, phrasal verbs, and strings of varying lengths. The contextual classification generated by Gemini Notebook further illustrates the diverse military situations and registers in which these expressions occur.

To address the question, "How important are the 515 frequent military multiword expressions in the present subtitles corpus of military and war films and drama series?", lexical coverage was calculated as the percentage of running words accounted for by these expressions in the corpus. The list comprises 515 multiword expressions (MWEs), ranging from two to six words, with a cumulative frequency of 471,029 occurrences, corresponding to 1,884,116 running words. Collectively, these expressions account for 3.14% of the tokens in the present 60-million-token subtitles corpus of military and war films and drama series.

At first glance, a lexical coverage of 3.14% may appear modest. However, this figure does not fully reflect the potential impact of these expressions on comprehension. In particular, non-compositional MWEs cannot always be understood by combining the meanings of their individual words. As a result, viewers may recognize the component words and follow the general storyline while failing to grasp the intended meanings of particular expressions. This may create an "illusion of comprehension" (Martinez & Murphy, 2011), in which apparent understanding of the overall context masks gaps in the interpretation of specific expressions. In military and war-related media, such gaps may be especially consequential because unfamiliar expressions can convey important information about military operations, commands, relationships, and events.

The potential significance of these comprehension gaps can also be considered in relation to research on lexical coverage. Carver (1994) reports that native English-speaking children experience reading difficulty when a text contains two unknown words per hundred words. Similarly, Hu and Nation (2000) and Schmitt, Jiang, and Grabe (2011) identify approximately 98% lexical coverage, equivalent to one unknown word in every 50 words, as an important benchmark for adequate text comprehension. Although these findings concern general lexical coverage rather than the specific effects of MWEs, they highlight the potential consequences of encountering unfamiliar lexical items during reading. More importantly, the impact of an unfamiliar expression depends not only on its frequency but also on its semantic opacity and its role in conveying meaning. Therefore, even though the 515 MWEs account for only 3.14% of the corpus tokens, their non-compositional nature may make them significant comprehension bottlenecks.

These findings support the pedagogical value of the proposed MWE list. Rather than assuming that viewers can infer the meanings of these expressions from their individual words or the surrounding context, the researcher proposes that the 515 frequent military MWEs be included as explicit learning targets. Teaching these expressions may help learners move beyond a general understanding of military and war-related media toward more accurate interpretation of the language used in these contexts.

CONCLUSION

The principal aim of this study was to create a self-directed learning list of frequent non-compositional MWEs for EFL learners with an interest in military dramas who regularly binge-watch military films and drama series. Based on a principled set of criteria, a total of 515 non-compositional MWEs consisting of two to six words were selected. These frequent MWEs are mostly composed of words from the first 3,000 word families. Therefore, they may partially bridge the gap in lexical coverage that general vocabulary alone cannot provide for understanding English subtitles in military and war films and drama series, thereby facilitating viewing comprehension. Although the list of military-specific non-compositional MWEs is relatively short, it may provide a viable set of expressions for EFL students to learn within a limited period of time.

This study also identified the contexts of use in which the non-compositional MWEs occur in lines of dialogue. Such information may help raise learners' awareness of how these expressions function in authentic discourse. Despite the arbitrary nature of some cutoff values used in compiling the list, the findings suggest potential advantages to the explicit instruction of these phrases. These advantages warrant further investigation but are beyond the scope of the present study. As with other individual word lists, it is hoped that the list of frequent non-compositional MWEs may serve as a reference for the teaching and learning of Military English.

In military contexts, AI-assisted language tools have received increasing attention because operational environments place a high demand on rapid comprehension, precise communication, and standardized language use (U.S. Department of War, 2025; Vasankari & Koski, 2025). However, most existing AI language tools are trained primarily on general-purpose corpora and may not adequately reflect military discourse, specialized terminology, or military formulaic language. Their sensitivity to these domain-specific linguistic features therefore remains limited (Georgieva & Marinov, 2017). This study identifies frequently used non-compositional multiword expressions in the military register as represented in military films and drama series, with the aim of addressing part of this gap.

ACKNOWLEDGMENTS

This study was supported by a Grant (NSTC 115-2410-H-214-007) from the National Science and Technology Council in Taiwan.

REFERENCES

Anthony, L. (2026a). *AntConc* (Version 4.4.2) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Anthony, L. (2026b). *AntWordProfiler* (Version 2.3.0) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antwordprofiler/

Carver, R. P. (1994). Percentage of unknown words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior, 26*, 413–437. https://doi.org/10.1080/10862969409547861

Georgieva, V., & Marinov, P. (2017). Intercultural interactions in a military context. *Land Forces Academy Review, 22*(3), 153-162. https://doi.org/10.1515/raft-2017-0021

Hinkel, E. (2023). Teaching and learning multiword expressions. In E. Hinkel (Ed.), *Handbook of practical second language teaching and learning* (pp. 435–448). Routledge. https://doi.org/10.4324/9781003106609-36

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689–696. http://hdl.handle.net/10125/67046

Horst, J. S. (2013). Context and repetition in word learning. *Frontiers in Psychology, 4*, Article 149. https://doi.org/10.3389/fpsyg.2013.00149

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology, 2*, Article 17. https://doi.org/10.3389/fpsyg.2011.00017

Hsu, W. (2023). Korean drama fever—Expanding English lexicon through watching English-subtitled K-dramas: The case of non-compositional multi-word expressions. *Journal of English Teaching, 9*(2), 269–285. https://doi.org/10.33541/jet.v9i2.4761

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. http://hdl.handle.net/10125/66973

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Labrador, B. (2026). Capitalizing on genre-based corpora with the use of the AI-powered research tool Notebook LM. *Alfinge. Revista de Filología, 37*, 51–75. https://doi.org/10.21071/arf.v37i.18449

Locher, M. A., & Messerli, T. C. (2020). Translating the other: Communal TV watching of Korean TV drama. *Journal of Pragmatics, 170*, 20–36. https://doi.org/10.1016/j.pragma.2020.07.002

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly, 45*(2), 267–290. https://doi.org/10.5054/tq.2011.247708

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16. http://hdl.handle.net/10125/66881

Nation, I. S. P. (2017). *The BNC/COCA level 6 word family lists* (Version 1.0.0) [Data file]. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Pedersen, J. (2019). Fansubbing in subtitling land: An investigation into the nature of fansubs in Sweden. *Target, 31*(1), 50–76. https://doi.org/10.1075/target.18017.ped

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly, 45*(4), 689–717. https://doi.org/10.5054/tq.2011.268062

Rubenking, B., & Bracken, C. C. (2021). Binge watching and serial viewing: Comparing new media viewing habits in 2015 and 2020. *Addictive Behaviors Reports, 14*, Article 100356. https://doi.org/10.1016/j.abrep.2021.100356

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339–348. https://doi.org/10.1093/elt/ccm091

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Talaván Zanón, N. (2006). Using subtitles to enhance foreign language learning. *Porta Linguarum, 6*, 41–52. https://doi.org/10.30827/Digibug.30659

U.S. Department of War. (2025). *The War Department unleashes AI on new GenAI.mil platform*. https://www.war.gov/News/Releases/Release/Article/4354916/the-war-department-unleashes-ai-on-new-genaimil-platform/

Vasankari, L., & Koski, A. (2025). GenAI in the military: Trends and opportunities. *Scandinavian Journal of Military Studies, 8*(1), 416–434. https://doi.org/10.31374/sjms.415

Webb, S., & Rodgers, M. P. H. (2009a). Vocabulary demands of television programs. *Language Learning, 59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics, 30*(3), 407–427. https://doi.org/10.1093/applin/amp010

Wood, D. (2020). Classifying and identifying formulaic language. In S. Webb (Ed.), *Routledge handbook of vocabulary studies* (pp. 30–45). Routledge. https://doi.org/10.4324/9780429291586-3

APPENDIX

The 32-page appendix, containing 515 military-specific non-compositional MWEs and their meanings, is available at: https://tinyurl.com/2hjprt9c