

TATIANA SZCZYGLÓWSKA

University of Bielsko-Biala, Poland

Institute of Modern Languages/Faculty of Humanities and Social Sciences

ORCID: 0000-0002-5051-4080

tszczyglowska@ubb.edu.pl

A CORPUS STYLISTIC APPROACH TO ERNEST HEMINGWAY'S FICTION: KEYWORDS AND RECURRENT PHRASES IN *THE SUN ALSO RISES*

Critics and readers alike revere Ernest Hemingway's novel *The Sun Also Rises* for its innovative writing style, characterised by omission and understatement, yet remaining semantically rich. However, the novel's keywords and recurrent phrases, and their respective meanings, have received little attention, probably because they are dispersed across the text. Consequently, it is difficult to analyse their narrative significance in detail within their immediate textual context. Corpus methods can address this issue by systematically identifying such repeated linguistic patterns in the text and revealing the hidden themes they convey. This study adopts this approach, using a corpus methodology to analyse the keywords and most frequent four-word n-grams and p-frames in Hemingway's debut novel, with the aim of demonstrating their contribution to the construction of its literary meanings.

Keywords: *The Sun Also Rises*, corpus stylistics, literary meaning, keywords, n-grams, p-frames

1. Introduction

Published in 1926, *The Sun Also Rises* (henceforth *TSAR*) is the debut novel of the Nobel Prize-winning 20th-century American novelist and short story writer, Ernest Hemingway. Initially deemed "skilfully written" but lacking "artistic significance" (Wagner-Martin 1990: 1), the novel soon gained recognition for its innovative writing style, characterised by lucidity, concision, understatement, and omission, yet infused with symbolism operating behind the curtain of stripped narrative (Svoboda 1983: 4; Wagner-Martin 1990: 2). While both the text and manuscript of the novel are now available to researchers, a substantial

portion of Ernest Hemingway's writing has yet to receive adequate scholarly attention (Rothman 1928: 336). One possible reason for this is that claims about style lack solid empirical support unless they are verified quantitatively (Leech and Short 2007: 57). This study therefore combines literary stylistics and corpus linguistics to analyse keywords and the most frequent four-word phrases in *TSAR*, attempting to identify their salient characteristics within their textual contexts and to reveal their respective interpretive weight, not readily accessible through close reading alone.

Previous research shows that the analysis of literary texts can be refined by examining their lexical and structural regularities. Biber (2011: 21) notes that researching literature through a corpus linguistics lens is not a new phenomenon, but that it has gained markedly in popularity in recent years and is now commonly conducted within the framework of corpus stylistics. Overviews of such analyses, as found in the works of Fischer-Starcke (2010), Mahlberg (2013) and Balossi (2014), demonstrate that corpus stylistic research in literature takes many forms, with scholars using corpus methods to analyse authorial style and authorship attribution, examine the style and discourse presentation of literary texts, and inform literary interpretation. While research aims are often pursued by focusing on frequency-based wordlists and other types of descriptive statistics, corpus methodologies for stylistic analysis also include "keyword analysis, identifying typical extended lexical phrases, and collocational analysis" (Biber 2011: 16). The latter allow to empirically identify linguistic patterns that are distinctive of the target text and divergent from those that are typical of the reference corpus. These patterns can then be interpreted for their stylistic and literary effects.

1.1. Keywords

The keywords of a text or corpus are items that occur statistically more frequently, in the case of so-called positive keywords, or less frequently, in the case of so-called negative keywords, in the text under analysis than in the reference corpus (Scott and Tribble 2006: 55). They are therefore characteristic of the text in question. Shepherd and Berber Sardinha (2013: 72) argue that keywords are crucial for interpreting literary meaning, since their unusual frequency in a text "is a useful indicator of style [...] as keywords may be indicative features of the study corpus, such as content lexical items, which in the case of narrative, may be plot explanatory". Distinctive keywords can serve specific literary purposes (Leech and Short 2007: 40), so analysing them can help to trace a text's semantic network and forms of expression. This has been demonstrated, for example, by Scott and Tribble (2006), who identify salient lexical features of *Romeo and Juliet* against Shakespeare's other plays; Culpeper (2009), who uses this method to show how character-specific keywords, such as Juliet's frequent

use of *if*, reflect underlying psychological and thematic concerns; and Fisher-Starcke (2010), who compares the keywords of *Northanger Abbey* with both Austen's other works and contemporary fiction to isolate text-specific lexis.

1.2. Extended lexical phrases

When extended lexical phrases are approached from a frequency-based perspective, they are often operationalised as n-grams – uninterrupted sequences of *n* word-forms (e.g. *and have a drink*) – and p-frames, which are n-grams with one variable lexical slot (e.g. *went out * the*) (Stubbs 2007: 166). Shepherd and Berber Sardinha (2013: 72, 84) argue that the extraction of n-grams from a corpus can reveal its phraseological or terminological characteristics, thereby uncovering either the structural patterns of a literary text or its thematic content. Additionally, focusing on the p-frames repository provides insight into the degree of phraseological fixedness and variability within the text. If such corpus-stylistic analyses of a literary text go beyond identifying its most frequent items and also consider their typical co-occurrence preferences, reflecting the collocational and colligational patterns exhibited by these items, the findings may shed light on the underlying structural and conceptual framework on which the text is based. Recurrent phraseological units in literary texts have been the subject of research in analyses such as Stubbs' (2005) study of Conrad's *Heart of Darkness*, Starcke's (2006) examination of Austen's *Persuasion*, Mahlberg's (2007, 2013) studies of Dicken's novels, and Fisher-Starcke's (2009, 2010) analyses of Jane Austen's novels.

1.3. Additional insights from corpus approaches to the study of literary texts

The applicability of corpus linguistics to the study of literary texts offers a clear advantage over more conventional methods of identifying, describing and interpreting linguistic – including stylistic – features, as it enables the extraction of actual patterns of language use. These patterns can then be analysed both quantitatively and qualitatively, in a manner that is more objective than traditional intuition-based approaches. This, in turn, allows for fresh insights into the unique textual features that are of literary significance but are difficult to detect through close reading alone. Some of these features result from the author's conscious choices, while others lie beyond their conscious control; together, however, they contribute to the distinctive linguistic imprint of an author's style. Its close analysis becomes particularly challenging in the case of the novel, which, as Short (1996: 255) observes, "is probably the most difficult genre to analyse stylistically", primarily due to its length.

Selected analyses of literary texts with the use of corpus methods have been mentioned in the previous subsections. A more comprehensive coverage of the

current state of corpus approaches to literature can be found, for instance, in Ho (2011: 8-11). Of particular relevance to the present study, which adopts the same methodological approach, is Fisher-Starcke's (2009) corpus-stylistic analysis of keywords and frequent phrases in Jane Austen's *Pride and Prejudice*. Her analysis reveals systematic linguistic patterns in the novel relating to family, emotion, communication, and social norms. These patterns illuminate literary meanings and narrative structures that cannot be easily discerned through traditional, intuitive literary criticism alone. As Shepherd and Berber Sardinha (2013: 85) suggest, "a lexical analysis of a text based on keyness, the extraction of concordancing lines for the keywords and a simple extraction of n-grams, may cast a light on the underlying (and often hidden) themes of a book".

Overall, it seems that an undeniable advantage of the frequency-based approach to literature is that it enables the identification and examination of salient lexis in a text in an objective and systematic way, within their linguistic context. The lexico-grammatical patterns in which these prominent words and phrases repeatedly occur provide information on the meaning relationships between the different aspects of the fictional world to which these lexical items refer. As Hunston and Francis (2000: 37) emphasise, the meaning of a word is determined by the words and structures with which it frequently co-occurs. Therefore, it can be assumed that the linguistic features revealed through corpus methods, owing to their recurrent nature, configure the thematic scaffolding of the narrative. Indeed, Shepherd and Berber Sardinha (2013: 71) note that the frequency of a lexical item is essential for understanding its meaning and for making wider assertions about the discourse. Thus, taking a corpus perspective to observe, describe and interpret statistically significant linguistic patterns in a literary text may help to uncover the subtle thematic threads present within it but operating below the narrative surface. This is particularly relevant in the case of *TSAR*'s famously sparse prose, characterised by Hemingway's signature 'iceberg' style of omission (Balassi 1990: 125) and "repetitive patterns" that "cohere throughout his text [...] producing meaning" (Elbaz 1984: 77).

1.4. Research questions

Taking the above into account, this study uses a corpus methodology to examine the literary text of *TSAR*, addressing the following research questions:

- 1) What literary meanings are encoded by the keywords and the linguistic patterns in which they occur throughout the novel?
- 2) What literary meanings are encoded by the novel's most frequent four-word phrases (i.e. 4-grams and 4-frames) and the linguistic patterns in which they occur?

2. The study

2.1. Materials

This study draws on Ernest Hemingway's novel *TSAR*, which forms the basis of the study, as well as two reference corpora: Hemingway's other six novels (the corpus *Hemingway*), and literary texts from roughly the same period (the corpus *ContempLit*), as summarised in Table 1. Standardised type/token ratios (i.e. STTR) are provided for comparing lexical richness across corpora, with *ContempLit* showing the lowest value. The *ContempLit* texts are listed in the Appendix.

TSAR is Hemingway's debut novel. Set in post-First World War Europe, it chronicles the lives of a group of British and American expatriates. The narrator and protagonist, Jake Barnes, is a war veteran and journalist who has been left impotent by his injuries. He is in love with Lady Brett Ashely, a woman he cannot truly have. Along with their hard-drinking, cynical friends, the couple moves between the nightlife of Paris and the bullfights of Pamplona in Spain. Jealousy, frustrated passions, and disillusionment with life after the war are prevalent themes. The novel portrays the vitality and the emptiness of the so-called 'Lost Generation' who are searching for meaning and identity in a world marked by trauma, loss, and disconnection.

The novel exemplifies modernist aesthetics, characterised by sparse prose, realism and dispassionate understatement. The narrative omits many details concerning the plot, setting and the characters' emotions, attributes and

Table 1. Corpora

Corpus	Content	Size (tokens)	STTR
<i>TSAR</i>	<i>The Sun Also Rises</i> (1926)	67,775	60.97
<i>Hemingway</i>	<i>A Farewell to Arms</i> (1929), <i>To Have and Have Not</i> (1937), <i>For Whom the Bell Tolls</i> (1940), <i>Across the River and into the Trees</i> (1950), <i>Islands in the Stream</i> (1970), ¹ <i>The Garden of Eden</i> (1986) ²	606,459	63.48
<i>ContempLit</i>	30 novels by various authors, published between 1911 and 1969 to be roughly contemporary with Hemingway's writing career	3,007,506	57.49

¹ The novel was written in 1950-1951 and published posthumously.

² The novel was written in 1946-1961 and published posthumously.

thoughts,³ encouraging readers to discover the meaning for themselves. It can also be described as a roman à clef, as Hemingway based its characters on real people from his social circle and mirrored events from his own life (Svoboda 1983: 9). *TSAR* was well received by readers and quickly became a commercial success, gradually establishing itself as Hemingway's most renowned book. Wagner-Martin (1990: 1) considers this unusual, since "most writer's first novels do not turn out to be their most important work". Using corpus-linguistic methods to delve deeper into the novel can reveal new literary meanings, particularly those that are difficult to detect with the naked eye. This is likely to be the case since, as Balassi (1990: 125) notes, "nowhere did Hemingway so thoroughly or so daringly" leave so much meaning below the surface, producing "a richly textured but elusive story that has been plagued by misreadings since its publication" (Balassi 1990: 125).

2.2. Method

This study examines the linguistic units of *TSAR*, which are identified on two hierarchical levels in language: lexis and phraseology. The lexical analysis was performed using *WordSmith Tools* v. 6. (Scott 2012) and involved examining keywords with a minimum frequency threshold of 25 that were calculated using the log-likelihood test with a probability value of 0.000001. Adopting the approach of Fisher-Starcke (2009), who studied the keywords and frequent phrases in Austen's *Pride and Prejudice*, *TSAR* was compared with two reference corpora: *Hemingway* and *ContempLit*. This revealed lexical features that are specific to the novel, as opposed to Hemingway's other books and literary works from the same period. As Baker (2004: 347) states, such quantitatively detected keywords "direct the researcher to important concepts in a text". The two lists of keywords were then compared to identify coincident semantic and functional sets, revealing the novel's "aboutness" and style (Scott and Tribble 2006: 59-60). This also helped to select the most relevant keywords for further analysis. Fisher-Starcke (2009: 501) explains that the relevance of keywords for analysis increases "when the same words are also identified by a second keyword analysis which uses another reference corpus".

The phraseological analysis was conducted using *kfNgram* (Fletcher 2012) to identify the novel's most frequent four-word sequences. This involved examining uninterrupted n-grams and their variants with one variable slot, known as p-frames. To yield a manageable number of items of each type for further

³ The writing technique of presenting only a small part of a story's deeper meaning, leaving the rest beneath the surface, is known as the "iceberg theory of omission" (Trodd 2007: 17). It was developed by Hemingway, who, as a journalist, opted for a minimalist and concise writing style due to the limited space available in newspapers.

analysis, the frequency threshold was set at six for n-grams and at four for p-frames. Four-word⁴ phrases were chosen because they are “frequent enough for good diversity” (Bestgen 2018: 206), exhibit greater phrasal, structural and functional distinctiveness than shorter units (Biber et al. 2021: 984; Goźdz-Roszkowski 2011: 110), and are particularly suitable for identifying collocational and colligational patterns (Fischer-Starcke 2010: 111), which typically span four words to the left and right of a node (Sinclair 1991). Examining the concordance lines of selected single- and multi-word units using *WordSmith Tools* provided additional insight into the important literary meanings expressed by keywords in the novel’s textual content and indicated by phrases in its textual structure.

3. Keywords

Keywords provide information about how unique the text or corpus is in terms of content, style, and structural features. Organising them into semantically and functionally related groups helps to reveal the propositional content of a text, and thus the literary meanings they encode. Examining keywords in their linguistic context further supports this intuitive process, as shown in concordance lines that highlight the items’ semantic and grammatical patterns of co-selection.

Comparing *TSAR* with *Hemingway* produced 100 keywords (85 positive) for the novel, whereas comparing *TSAR* with *ContempLit* produced 142 keywords (128 positive). Excluding proper names, the first comparison yields 71 keywords, while the second yields 120. Of the 191 positive keywords in total, 60 overlap, 60 result exclusively from comparing *TSAR* with *ContempLit*, and 11 result from comparing *TSAR* with *Hemingway*. When the raw frequencies of the keywords in both lists are compared, it emerges that the comparison with *ContempLit* generates only 1.57 times as many items as the comparison with *Hemingway* (27,245 vs. 17,256). These results can be compared with those of Fisher-Starcke (2009: 496-502), who examined the keywords in Austen’s *Pride and Prejudice*. When Austen’s other novels were used as a reference corpus, only 30 non-proper-name keywords were produced. However, when the book was compared to Austen’s contemporary literature, as many as 198 keywords were produced. In terms of raw frequency, these were nearly seven times more numerous than those resulting from the first comparison. It was concluded that *Pride and Prejudice* shares lexis and topics with Austen’s other novels, but differs significantly from its contemporary literature. In light of this, the findings of the present study suggest that *TSAR* and Hemingway’s other books are not particularly similar in terms of

⁴ Following Biber et al. (2021: 982), contracted forms were considered as single orthographic words.

topics and its respective lexis. While *TSAR* differs more sharply from its contemporary literature than from Hemingway's other works in terms of lexis, the difference is not that great.

The keywords revealed by the two comparisons are listed in Table 2. They are categorised into semantic and functional sets, which were developed intuitively based on their dominant meanings in the novel and verified through concordance-based analysis. For example, of the 28 occurrences of *dear*, the word is used as a term of address to refer to a character 24 times, and as an adjective only four times. Therefore, it was classified as a word referring to people. The middle column shows keywords that appear in both lists, sorted in decreasing order. The columns on the left and right contain keywords belonging to a given category that occur only in the list generated by comparing *TSAR* with *Hemingway* and *ContempLit*, respectively. Contractions were treated as single orthographic words, the role of which was determined by the initial element. Words marked with an asterisk appear twice in the Table because they are semantically related to different keyword categories. The numbers in brackets show how many items in a given keyword category were identified in each comparison, including overlapping items and items resulting from one comparison only.

Table 2. Keywords in *TSAR* identified by comparisons with *Hemingway* and *ContempLit*

<i>TSAR</i> vs. <i>Hemingway</i>	Overlapping keywords	<i>TSAR</i> vs. <i>ContempLit</i>
(1) <i>my</i>	personal pronouns (9) <i>I, I've, I'm, you're, he's, we, us, they're, it's</i>	(3) <i>you, me, it</i>
(1) <i>say</i>	forms of content verbs that describe the actions or events in the novel (8) <i>went, got, walked, started, came, paid, sat, stopped</i>	(13) <i>said, asked, go, get, put, come, sitting, looked, smiled, passed, know, took, eat*</i>
	referring to elements of the urban landscape (7) <i>café*, hotel, street, bus, square, taxi, town</i>	(4) <i>bar*, bridge, train, trees</i>
(3) <i>that's, around, what's</i>	functors (7) <i>a, the, out, down, up, on, all</i>	(7) <i>but, through, of, to, away, why, ahead</i>
(2) <i>lady, dear</i>	referring to (groups of) people (7) <i>crowd, count, bull-fighter*, friends, people, English, waiter*</i>	(1) <i>darling</i>
	related to bullfighting (6) <i>bull, bulls, bull-fight, bull-fighter*, ring, fiesta</i>	

Table 2. cont.

<i>TSAR</i> vs. <i>Hemingway</i>	Overlapping keywords	<i>TSAR</i> vs. <i>ContempLit</i>
(2) <i>didn't, doesn't</i>	inflectional forms of auxiliary verbs (4) <i>don't, was, were, isn't</i>	(3) <i>did, do, is</i>
	signalling intentions or desires (3) <i>let's, wanted, going</i>	(1) <i>want</i>
	related to visiting cafés and similar venues (3) <i>café*, table, tables, waiter*</i>	(8) <i>drink, wine, bottle, lunch, glasses, coffee, eat*, bar*</i>
	negatively connotated words that describe the characters' impressions of various aspects of the novel's plot (2) <i>damned, drunk</i>	(2) <i>hell, tight</i>
	interjections used to express greetings or emotions (2) <i>hello, oh</i>	
(2) <i>couldn't, wouldn't</i>	modal verbs (1) <i>can't</i>	
	positively connotated words that describe the characters' impressions of various aspects of the novel's plot (1) <i>nice</i>	(4) <i>good, fine, lovely, wonderful</i>
	intensifiers (1) <i>rather</i>	(1) <i>very</i>
	other (1) <i>hot</i>	(11) <i>white, thing, things, place, matter, open, Spanish, outside, yes, road, country</i>
	time references (0)	(4) <i>time, now, morning, years</i>

Given that “only words, and therefore topics, that are not dominant in the reference corpus are identified as key for the data analysed” (Fisher-Starcke 2009: 500), the following conclusions can be drawn about *TSAR* based on the data presented in Table 2:

- Nearly all of the keyword categories are distinctive for *TSAR* when compared with both *Hemingway* and *ContempLit*. This indicates that these topics and their respective vocabulary are particularly prevalent in the novel.

- The keywords include content-bearing terms (i.e. mainly lexical words such as those referring to people or to elements of the urban landscape) and style markers (i.e. mainly function words such as pronouns, functors or auxiliary verbs), offering insight into the novel's themes and Hemingway's linguistic choices.
- Time references are important in *TSAR* and common also in Hemingway's other novels, as indicated by their absence in the list of keywords resulting from a comparison of *TSAR* with *Hemingway*.
- The most prominent keyword categories reflecting the novel's important themes are those referring to people, bullfighting, the urban environment, and social spaces such as cafés.
- *TSAR* is strongly focused on human characters, as indicated by the occurrence of personal pronouns, words referring to people and denoting their actions (e.g. *paid*, *smiled*, *know*).
- The frequent occurrence of first and second person pronouns, along with verbs like *said* and *asked*, highlights extensive dialogues and the importance of communication.
- Keywords related to negation, intentions, and desires suggest a thematic emphasis on challenged assumptions or expectations, as well as on characters' motivations and aims.

3.1. Keywords related to bullfighting

The bullfighting keyword category was selected for further analysis because all of its terms appear as keywords in comparisons of *TSAR* with both reference corpora, confirming bullfighting as a significant theme of the novel. While not the most dominant category in terms of raw frequency, it has the highest overall keyness among content-bearing keyword categories in both comparisons. Table 3 shows the frequencies of occurrence and keyness values of the bullfighting keywords, ordered by statistical significance.

The analysis shows that keywords related to bullfighting are concentrated in the second half of the novel. As illustrated in Figure 1, which shows the dispersion plot for the six keywords listed in order of keyness, they become prominent from Chapter 10 onwards. This is where the word *bulls* first appears and becomes prevalent throughout the rest of the text, as do the other analysed terms. This concentration is preceded by an increasing frequency of the word *fiesta*, first introduced in Chapter 8, which marks the build-up to the climax of the novel: the Fiesta of San Fermin in Pamplona. The strong collocation of *fiesta* with *the* in L1 position (38 instances vs. 2 with *a*) indicates a reference to a specific event, underscoring its central role in the novel.

The concordance lines of *fiesta* with its most frequent collocates in L1 (*the*, 38 instances) and R1 (*was*, 8 instances) positions demonstrate a usage pattern that

Table 3. Keywords related to bullfighting

Keyword	Frequency in <i>TSAR</i>	Keyness	
		<i>TSAR</i> vs. <i>Hemingway</i>	<i>TSAR</i> vs. <i>ContempLit</i>
bulls	70	280.99	534.16
bull	105	258.65	463.47
fiesta	46	195.18	341.37
bull-fight	26	119.47	198.39
bull-fighter	26	119.47	189.88
ring	27	45.65	44.80
Total	300	1,019.41	1,772.07

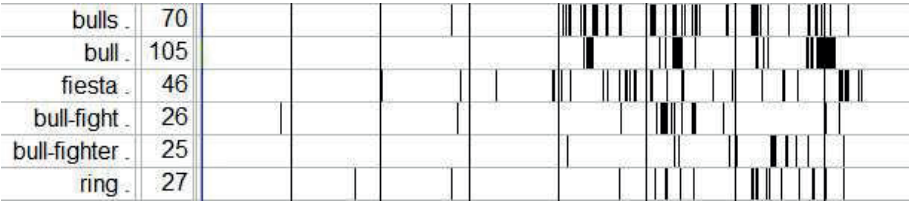


Figure 2. Proportions of distinctive feature substitutions in the sample

seems to align with a fundamental principle of Hemingway’s writing style: the iceberg theory. Examining the first six concordance lines in particular, it is evident that *fiesta* occurs in brief sentences that epitomise the writer’s sparse style, characterised by minimalism, simplicity and concision. As *fiesta* is directly linked to the novel’s climax, it is probably most susceptible to Hemingway’s avoidance of explicit explanations that would reveal its true symbolism. It is the readers who are expected to infer the significance of the fiesta for the story themselves.

t comes after a good bullfight. The fiesta was going on. The drums pounded and paper that stuck to the stone. The fiesta was over. I drank a coffee and after it stop. Outside in the square the fiesta was going on. It did not mean anythin " Cohn said. "Bring them here." The fiesta was really started. It kept up day an In the morning it was all over. The fiesta was finished. I woke about nine o'clock " It was beginning to get dark. The fiesta was going on. I began to feel drunk one of the beds in Cohn's room. The fiesta was going on outside in the night, but was the last day of the fiesta. The fiesta was solid and unbroken, but the motor

Figure 1. Keywords related to bullfighting in dispersion plot

The concordance analysis of *bull-fighter* reveals a usage pattern that is formed by the words found in its immediate co-text. In 14 instances, 11 of which are illustrated below, *bull-fighter* is directly preceded by *the* and most often followed by *chap* (6 instances)⁵ and *fellow* (5 instances) in the R1 position. This indicates that *bull-fighter* is often used attributively.⁶ While both right-hand collocates are informal and somewhat old-fashioned synonyms for *man*, *chap* sounds distinctly British, whereas *fellow*, which is used in both British and American English and carries prestige in professional or academic contexts, would more commonly be expressed as *guy*. Hemingway's alternation between *bull-fighter chap* and *bull-fighter fellow* appears to be a deliberate stylistic choice, intended to reflect the speech of British and American expatriates in the 1920s.

you know. She's gone off with the bull-fighter chap." "No." "Yes. She looked for
now. All for forgiveness. And the bull-fighter chap hit him in the face again."
e rest," Bill said. "It seems the bull-fighter chap was sitting on the floor. He
. Said it would be wicked. So the bull-fighter chap sort of rather staggered ove
table. "Brett's gone off with the bull-fighter chap. They're on their honeymoon.
e went in and found Brett and the bull-fighter chap in the bull-fighter's room,
happened finally?" "It seems the bull-fighter fellow was sitting on the bed. He
hey weren't having any of it. The bull-fighter fellow was rather good. He didn't
nd wanted to shake hands with the bull-fighter fellow. He wanted to shake hands
ohn. 'I'd be ashamed to.' "So the bull-fighter fellow hit him just as hard as he
aned down to shake hands with the bull-fighter fellow. No hard feelings, you kno

Figure 2. Example occurrences of *chap* and *fellow* in *TSAR*

Going beyond the immediate concordance lines of *chap* and *fellow*, it can be seen that British characters such as Mike Cambell or Lady Brett Ashley naturally say *chap*, while American characters such as Bill Gorton or Jake Barnes more often tend to say *fellow*. Evidence for this is provided in Figure 2. Interestingly, in Chapter 17, when Mike is recounting Robert Cohn's fight with Romero to Jake and Bill, who ask questions and encourage him to continue the story, he starts alternating between *chap* and *fellow*. One possible reason for this is that he is addressing his American friends. Alternatively, by using the casual, slightly humorous *chap*, he may be trying to reduce the aura of the bullfighter implied by the somewhat more serious and respectful *fellow*.

The novel contains 300 instances of keywords related to bullfighting, 214 of which are immediately preceded by *the*, most notably *ring* (88.88% of instances, i.e. 24 out of 27) and *fiesta* (82.6% of instances, i.e. 38 out of 46). While *the* is

⁵ In total, there are 7 instances of *chap* in the R1 position of *bull-fighter*, but in one instance, the keyword is preceded by *this*.

⁶ Of the 25 instances of *bull-fighter* in *TSAR*, 12 are attributive.

Mike speaking:

Chapter 15 - "I believe, you know, that she's falling in love with this bull-fighter chap," Mike said.

Chapter 17 - "Oh, go to hell, Cohn," Mike called from the table. "Brett's gone off with the bull-fighter chap. They're on their honeymoon."

Brett speaking:

Chapter 8 - "Chap bought it for me. Don't you like it?"

Brett and Mike speaking:

Chapter 9 - "By the way, where is Bill?" Brett asked.

"He's out at Chantilly dining with some people."

"He's a good chap."

"Splendid chap," said Mike. "He is, you know."

Bill speaking:

Chapter 8 - "You're right. Wasn't me at all. Was another fellow. Think we called him the local Harvard man. Remember him now. Studying music."

Chapter 17 - "They can't insult Mike," Bill said. "Mike is a swell fellow. They can't insult Mike. I won't stand it. Who cares if he is a damn bankrupt?" His voice broke.

Mike recounting the incident to Jake and Bill:

Chapter 17 - "It seems the bull-fighter chap was sitting on the floor. He was waiting to get strength enough to get up and hit Cohn again. Brett wasn't having any shaking hands, and Cohn was crying and telling her how much he loved her, and she was telling him not to be a ruddy ass. Then Cohn leaned down to shake hands with the bull-fighter fellow. No hard feelings, you know. All for forgiveness. And the bull-fighter chap hit him in the face again."

the most frequent word in English (Cappock and Beaver 2015: 378) and also the most frequent word in the novel, its presence in the L1 position in 71.33% of cases in this category appears deliberate. Among the other content-bearing categories, only the urban landscape category has a higher score, with its keywords collocating with *the* in the L1 position in 73.1% of instances (i.e. 424 out of 580). However, it is worth noting that this category consists of 11 keywords, almost twice as many as the bullfighting category. This repeated use of the definite article creates rhythm and precision, reflecting Hemingway's journalistic background, which is believed to have influenced his clear, "to the point writing" style (Keskesay 2021). When writing *the bull*, *the fiesta* or *the ring*, Hemingway invokes the archetypal elements of a culturally significant ritual, thereby grounding the reader in the shared reality of the characters who are direct spectators of the unfolding ritual: the fiesta.

The keywords belonging to the bullfighting category also reflect another defining feature of Hemingway's minimalist prose style: his avoidance of adjectives. He developed this technique while working as a cub reporter for the *Kansas*

City Star (Dewberry 1996: 19). Excluding 12 instances in which the six keywords are preceded by numerals acting as adjectives (e.g. *first, second, one*), only 24 other occurrences of the bullfighting keywords collocate with adjectives in the L1 and/or L2 positions. The concordance lines below provide examples of these rare cases.

en it yet. Tomorrow, if I get a **good bull**, I will try and show it to you." When h
 "I say," Mike said, "they were **fine bulls**, weren't they? Did you see their horn
 eder. So he had two small, **manageable bulls** without much horns, and when he felt th
 lls?" he asked. "Good. They were **nice bulls**." "They're all right"--Montoya shook h
 ps," Mike said. "It was a damned **fine fiesta**." "So long, Mike," Bill said. "I'll s
 hile Bill said: "Well, it was a **swell fiesta**," "Yes," I said, "something doing all
 , and went down-stairs. It was a **good bull-fight**. Bill and I were very excited abo
 . "He nearly killed the **poor, bloody bull-fighter**. Then Cohn wanted to take Brett
 after the cloth." "He's just a **young bull-fighter**. He hasn't learned it yet." "Bu

Interestingly, keywords relating to bullfighting mostly occur in narrative fragments; only 74 (24.66%) instances occur in dialogue in *TSAR*. One possible explanation for this is that Hemingway tied this motif to the narrative voice, rather than to the characters' superficial chatter, in order to emphasise the symbolic importance of bullfighting for the reader. After all, bullfighting is not simply a conversational topic, as such topics tend to lack depth. In line with Hemingway's austere writing style, the novel does not explicitly articulate the true significance of bullfighting, which must instead be decoded from Jake's narration.

3.2. Negative keywords

Analysing negative keywords "which occur significantly less often in the target corpus than in the reference one" (Evison 2010: 128) shows that, when compared with Hemingway's other novels, the top three most significantly underrepresented words in *TSAR* are *it, man* and *they*; and, when compared with his contemporary literature, *who, eyes* and *he*. Interestingly, comparing *TSAR* with each reference corpus yields items that explicitly refer to men: *he, his* and *man*. This is particularly revealing when *TSAR* is compared with Hemingway's other books: words such as *man, he, he's, him* and *his* are almost eleven times less frequent in *TSAR* than in his other works (1,977 vs. 21,047). Even considering that *he's* is among the keywords identified for *TSAR*, when both *Hemingway* and *ContempLit* are used as reference corpora, the novel still seems to focus less on a single male protagonist and more on a social group in which men are portrayed as either impotent (Jake Barnes) or weak (Robert Cohn), while women are depicted as sexually and socially dominant (Lady Brett Ashley).

4. Phrases

Phrases extracted from a text using corpus linguistic methods are word combinations of varying lengths that are identified based on their frequency. As Sinclair (2008: 408) puts it, they are the linguistic places “where structures are engineered to allow meanings to take shape”. When analysed, phrases provide insight into the grammatical, syntactic and lexical patterns of expression that dominate a text, shedding light on stylistic tendencies and recurring themes.

As outlined in Section 2.2., different frequency thresholds were assigned to the two phrase types: six for n-grams and four for p-frames. This resulted in the identification of 27 4-grams and 31 4-frames with 69 variants. It is worth noting that, when the frequency threshold for 4-frames was set above 4, there were too few recurrent units to obtain enough data for further analysis. When the minimum occurrence was set to 5, only three 4-frames were identified, and when it was increased to 7, the *kJNgram* tool returned no results. A similar pattern emerged for 4-grams: 17 units were obtained with the frequency threshold set at 7, but only 8 units were identified when it was increased to 8. For comparison, Fischer-Starcke (2009: 509) set the frequency threshold to 8 for 4-grams 17 for 4-frames in her analysis of frequent phrases in Austen's *Pride and Prejudice*, identifying 21 and 24 units respectively; she explained that this was “a restriction that has been imposed to generate a manageable amount of data”.

In this context, it is also worth comparing the lexical diversity of Hemingway's *TSAR* and Austen's *Pride and Prejudice*, which can be done by calculating the STTR. The value calculated for *TSAR* is 60.97, which is higher than the value of 57.49 calculated for Hemingway's contemporary literature, and much higher than the value of 42.25 calculated for *Pride and Prejudice*.⁷ For comparison, the STTR value for James Joyce's 1922 novel *Ulysses* is 50.91. Green (2017: 294, 296) considers this to be a high value, indicating “Joyce's lexical complexity”.

The above findings lead to the conclusion that Hemingway's *TSAR* is characterised by creative linguistic production. This is reflected in his unique word usage, as indicated by the relatively high STTR, and the low degree of phraseological recurrence reported above. The overall scarcity of recurring multi-word units and the relatively low number of variants for the 4-frames suggest that Hemingway does not rely on conventionalised, prefabricated sequences, preferring individualised, context-dependent phrasing instead. This aligns with the concise, minimalist style often attributed to the writer, who constructs meaning through expressive precision and resists formulaic repetition.

⁷ The STTR for *Pride and Prejudice* was calculated using the Project Gutenberg version of the novel, which contains 6,690 types and 124,757 tokens. This version was downloaded from: <https://www.gutenberg.org/cache/epub/1342/pg1342.txt> (accessed October 2025).

The phrases generated for the novel are listed below, alongside their raw frequencies (RF): Table 4 shows the 4-grams and Table 5 shows the 4-frames. For each 4-frame, Table 5 also lists the number of variants (V), the variant/p-frame ratio (VPR),⁸ and the predictability score (PM). V refers to the number of distinct lexical items that appear in the open slot of a p-frame, reflecting its overall productivity. The VPR, introduced by Römer (2010: 105) to express “the relation of different words that fill the blank (*) slot in a p-frame to the number of p-frame tokens”, shows internal variability in p-frames. The PM, as defined by Gray and Biber (2013: 114), represents “the percentage of all occurrences of a frame in which the most frequent filler occurred”, thereby indicating how predictable the filler word in the open slot is.

Table 4. 4-grams generated for *TSAR*

4-gram	RF	4-gram	RF	4-gram	RF
in front of the	19	and there was a	7	at the edge of	6
the edge of the	15	doesn't make any difference	7	at the far end	6
a hell of a	14	i don't want to	7	don't be an ass	6
do you want to	11	i went to the	7	i did not want	6
came out of the	9	it was a good	7	i would like to	6
went over to the	9	the other side of	7	on the other side	6
did not want to	8	we went into the	7	the centre of the	6
stopped in front of	8	you want me to	7	the door of the	6
all right i said	7	and have a drink	6	the top of the	6

The study identified 58 continuous and discontinuous sequences as *TSAR*'s most frequent phrases. 27 follow the ABCD pattern, 13 follow the *BCD pattern, 9 follow the ABC* pattern, 5 follow the A*CD pattern, and 4 follow the AB*D pattern. Among discontinuous phrases, initial- or final-slot patterns occur far more often than medial ones (22 vs. 9). This indicates that the novel's most frequent 4-frames resemble conversation, in which most frames “represent continuous sequences of fixed elements, with a preceding or following variable slot” (Biber 2009: 294).

In terms of their structural characteristics, the phrases presented in Tables 4 and 5 can be categorised into the three grammatical types proposed by Fuster-

⁸ The VPR is also known as the type-token ratio, which Gray and Biber (2013: 113) define as a measure of “internal fixedness” that “provides a way of measuring the degree to which the slot is relatively fixed or variable”.

Table 5. 4-frames generated for *TSAR*

4-frame	RF	V	VPR	PM	4-frame	RF	V	VPR	PM
the * of the	46	7	0.15	32.60	* one of the	9	2	0.22	55.55
in * of the	23	2	0.08	82.60	* on the bed	9	2	0.22	55.55
* out of the	17	3	0.17	52.94	* a good time	9	2	0.22	55.55
do you want *	15	2	0.13	73.33	* to the hotel	9	2	0.22	55.55
* the edge of	15	3	0.20	40.00	back to the *	9	2	0.22	55.55
* over to the	13	2	0.15	69.23	a bottle of *	9	2	0.22	55.55
went * to the	13	2	0.15	69.23	* came out of	8	2	0.25	50.00
* went into the	11	2	0.18	63.63	went out * the	8	2	0.25	50.00
* went to the	11	2	0.18	63.63	what do you *	8	2	0.25	50.00
i went to *	11	2	0.18	63.63	what's the mat- ter *	8	2	0.25	50.00
i * to the	11	2	0.18	63.63	in the * and	8	2	0.25	50.00
* did not want	11	2	0.18	54.54	and went * the	8	2	0.25	50.00
i did not *	10	2	0.20	60.00	in the square *	8	2	0.25	50.00
* top of the	10	2	0.20	60.00	the * it was	8	2	0.25	50.00
in the morning *	9	2	0.22	55.55	* looked at me	8	2	0.25	50.00
on the * of	9	2	0.22	55.55					

Márquez (2014: 96). Specifically, “NP-based types are noun phrases with or without a post-modifying fragment, e.g. *the * of the, the rear of the, [...]* The PP-based type [...] phrases [...] start with a preposition followed by a noun phrase with or without a post-modifying fragment, e.g. *in the * of, at the time of, [...]* the VP-based types are any phrases which contain a verb, e.g. *will be * for, we reserve the right*”. Table 6 summarises the structural types of the novel’s most frequent phrases.

Most *TSAR* phrases are VP-based, a pattern typical of conversation (Gray and Biber 2013; Biber et al. 2021), reinforced by nine subject-pronoun + VP structures (e.g. *I went to *, we went into*). In the 4-frames, slot-fillers are almost evenly split between function words (52%, 36 items, mainly prepositions and personal pronouns) and content words (47.8%, 33 items, mainly nouns and lexical verbs), further reflecting a conversational style (Biber 2009). The prevalence of nouns and prepositions among the slot-fillers indicates the importance of defining the people, objects, places and concepts that make up the novel’s fictional world, and of relating these elements to one another in terms of space, time, or logic. The strong presence of personal pronouns further emphasises the

Table 6. Structural categories of the most frequent phrases in *TSAR*

Type	4-grams (%)	4-frames (%)
NP-based	22.2	19.3
PP-based	14.8	32.3
VP-based	63.0	45.2
Other	0	3.2

central role of characters in *TSAR*, and of lexical verbs, reflects the need to bring the fictional world to life by generating energy and rhythm. Interestingly, adjectives were not used as slot-fillers and are rare, even in the fixed slots of the novel's recurring phrases. This aligns with Hemingway's austere writing style.

The 4-frames have 69 variants in total, giving an average of 2.22 per unit. Only three of the 4-frames have more than two variants, with *the * of the* being the most diverse unit with 7 different fillers. The low internal variability of the 4-frames is reflected in the VPR values, all of which are less than 0.30, indicating that the units are relatively fixed.⁹ The 4-frames also demonstrate relatively high predictability: 1 item is classified as highly predictable, 19 as predictable, 9 as almost predictable, and 2 as unpredictable; there are no highly unpredictable items.¹⁰ Based on the preceding discussion, it can be concluded that the range of recurring phrases in *TSAR*, particularly the 4-frames, is limited yet lexically and structurally stable. Hemingway tends to reproduce his few preferred multiword patterns with minimal modifications, producing strong phraseological cohesion in the novel. He uses this economical and precise phraseological core to develop the main themes of *TSAR* in a stylistically consistent and rhythmically regular way.

The dominant grammatical and functional patterns that can be identified in the novel's 4-grams and 4-frames are listed in Tables 7 and 8, respectively. Some phrases belong to multiple categories, which were developed intuitively through concordance analysis.

As the tables show, selected keywords recur within the phrases, notably the lexical words *went* (9 tokens) and *want* (7 tokens), and personal pronouns, especially *I* (8 tokens) and *you* (4 tokens). Incorporating the novel's most distinctive words into its recurrent phrases creates a coherent stylistic effect to which the reader is repeatedly exposed. This reinforces the novel's main themes in the reader's mind, eliminating the need for explicit explanations.

⁹ Values approaching 1 indicate variable slots, while values approaching 0 indicate more fixed patterns, "with fewer attested fillers" (Gray and Biber 2013: 113-114).

¹⁰ Garner (2016: 40) proposed the following categories of p-frames based on their predictability score: highly predictable with PM>075%, predictable with PM=50-75%, unpredictable with PM=25-50%, and highly unpredictable with PM<25%.

Table 7. Patterns of 4-grams in *TSAR*

Patterns of 4-grams	Realizations
Fourteen 4-grams specify location and/or movement in space	e.g. <i>at the far end, went over to the, the top of the</i>
Ten 4-grams include personal pronouns, five of which are <i>I</i>	<i>I, me, you, we, it</i> (impersonal)
Ten 4-grams contain past forms	<i>came, went, did, stopped, said, was</i>
Six 4-grams express volition, three of which negate it	<i>do you want to, did not want to, I don't want to, you want me to, I did not want to, I would like to</i>
Six 4-grams signal interactional discourse	<i>do you want to, all right I said, don't be an ass, doesn't make any difference, you want me to, and have a drink</i>
Five 4-grams contain the grammatical negation <i>not</i> or <i>n't</i>	<i>did not want to, doesn't make any difference, I don't want to, don't be an ass, I did not want</i>
Five 4-grams express evaluation or emotion	<i>a hell of a, all right I said, doesn't make any difference, it was a good, don't be an ass</i>
Three 4-grams describe familiar social spaces, moments and objects	<i>and have a drink, the door of the, the centre of the</i>

Table 8. Patterns of 4-p-frames in *TSAR*

Patterns of 4-frames	Realizations
Twenty-one 4-frames specify location, movement in space and/or temporal relationships	e.g. <i>went * to the, back to the *, in the morning *</i>
Eleven 4-frames contain past forms	<i>went, did, came, was, looked</i>
Seven 4-frames describe familiar social spaces, moments and objects	<i>* on the bed, * to the hotel, in the square *, a bottle of *, the * it was, in the morning *, * a good time</i>
Six 4-frames include personal pronouns, three of which are <i>I</i>	<i>I, me, you, it</i> (impersonal)
Four 4-frames express volition, two of which negate it	<i>do you want *, * did not want, what do you *, I did not *</i>
Three 4-frames signal interactional discourse	<i>do you want *, what do you *, what's the matter *</i>
Two 4-frames contain the grammatical negation <i>not</i>	<i>* did not want, I did not *</i>
One 4-frame expresses evaluation	<i>* a good time</i>

The tables also reveal overlaps between the dominant 4-gram and 4-frame patterns, particularly in phrases expressing spatial movement and/or location. Alongside the temporal 4-frame *in the morning* *, they build the narrative's dynamics, focusing it on the characters' restless routines. The frequent use of personal pronouns, particularly *I* and *you*, which is also implicitly present in phrases such as *and have a drink* and *don't be an ass*, foreground spoken language. While *I* emphasises the subjective experiences of the protagonists over external narration, *you* further reinforces the novel's interpersonal orientation and conversational rhythm. This spoken-style narration is marked by interpersonal friction, as evidenced by evaluative and affect-laden expressions such as *don't be an ass*, *doesn't make any difference* or *a hell of a*. Clusters of volitional and negative forms reflect the characters' disillusionment and emotional frustration, stemming from their inability to fulfil desires in the post-war reality. The negative attitude permeating the narrative is also indicated by the frequent presence of the grammatical negation *not* in the phrases, while the frequent occurrence of past-tense verbs reinforces the novel's focus on reporting lived experience rather than introspection.

4.1. A case study of selected phrases

To uncover additional literary meanings, selected frequent phrases and their co-texts were analysed to identify lexical and grammatical patterns in their usage. The phrases examined – *do you want to*, *you want me to*, *I don't want to*, *I would like to* and *I did not want* – the first and last of which are realisations of one or more frequent 4-frames, express volition and interpersonal negotiation, often in direct speech. This reinforces *TSAR*'s spoken-style narration. In two of the phrases, volition is modified by the negative particle *not*, highlighting the characters' inner conflict between desire and aversion. As *do you want to*, *I would like to* and *I did not want* have not been identified by Stubbs and Barth (2003: 73) as frequent in their fiction corpus, they may reflect Hemingway's idiolect rather than general literary usage.

The phrase *do you want to* (11 occurrences) mainly collocates with activity and communication verbs, typically referring to practical actions such as eating, going, or doing. This semantic preference is evident in the concordance lines below, in which both verb types are particularly frequent to the right of the phrase but also to the left. The phrase is a second-person interrogative and is often embedded in dialogue questions, accompanied by forms of conversational scaffolding, such as *Cohn asked*, *said Montoya*, or *I asked*. This suggests that the characters use the phrase to negotiate everyday activities and logistics, deflecting attention from the emotional emptiness they experience as members of the Lost Generation.

ave some supper?" Cohn **asked**. "Do you want to **eat**?" "Yes. Why not? I'm hungry." ard and **looked** me in the eye. "Do you want to **know** something, Jake?" "Yes." "I a tablet up. Like at Lourdes. Do you want to **hear**, Robert? I'll **tell** you. It's bout the car?" Bill **asked** me. "Do you want to keep it on?" "It doesn't make any oya. "**Come** on," **said** Montoya. "Do you want to **meet** Pedro Romero?" "Fine," **said** e. She's well **taken care of**." "Do you want to ?" "Would I **ask** you if I didn't w tside on the pavement I **said**: "Do you want to **go** to the fight?" "Sure," **said** Bi this pub in Saint Jean." "**What** do you want to **do** about the car?" Bill **asked** me. Boulevard du Port Royal. "**What** do you want to **do**?" I **asked**. "**Go** up to the café **do**?" I **asked**. Brett **turned** her he **leaned** his head around. "**Where** do you want to **go**?" I **asked**. Brett **turned** her he s looking over my desk. "**Where** do you want to **eat**?" "**How** about Wetzels? They've

The phrase *you want me to* (7 occurrences) shows similar collocational patterns, but it co-occurs more strongly with mental concepts such as *think*, *know* and *have no idea*. Evidence for this can be seen in the concordance lines below. Compared with *do you want to*, this collocational pattern is stronger, with five instances. The phrase signals interpersonal pressure between the characters, as the structure 'want someone to [verb]' involves an asymmetry of will, whereby one person expects another to act.

u've no idea, Jake." "Anything you want me to **do**?" "No, just **go** to the fight wi hen. **Think** of that." "Anything you want me to **think** about it?" "Don't be an ass r **felt** such a bitch." "**What** do you want me to **do**?" "**Come** on," Brett **said**. "Let' u'd **know** these things. **What** do you want me to **do**? **Come** over here and **tell** you e "Darling," she **said**. Then: "Do you want me to **send** him away?" "No. He's nice." k like that, darling." "**How** do you want me to **look**?" "Oh, don't be a fool. I'm let's **eat**," **said** Bill. "Unless you want me to **tell** you some more travel stories

The remaining three phrases also collocate with activity, communication, and mental concepts, as can be seen in the concordance lines below. *I would like to* (6 occurrences) is particularly associated with mental verbs and conveys a desire that is either softened or hypothetical. Jake Barnes, the novel's main protagonist, uses it three times, twice introspectively to report his thoughts. *I did not want* (6 occurrences) appears exclusively in Jake's reflective narration, signalling his inability to act on desire. The fact that these two phrases are mainly used by only one speaker suggests that they are significant in characterising that person as someone driven by inner conflict between tentative desire and reluctance. In contrast, *I don't want to* (7 occurrences) only appears in dialogue across different characters, foregrounding negation, which signals unwillingness to articulate desires.

The above findings provide interesting insights into the literary meanings of *TSAR*. As activity and communication verbs are among the novel's keywords,

out to the park?" Brett **asked**. "I don't want to **go** up yet. I fancy he's sleeping **shouted** from the wine-barrels. "I don't want to **eat** up your meal," I **said** when s
l you. I **support** that boy, but I don't want to have him around." "He is rather
never **liked** children much, but I don't want to **think** I'll never have them. I al
aw me **notice** it. "I **say**, Jake, I don't want to **ruin** your rugs. Can't you **give** a
aid. "I don't **know**," she **said**. "I don't want to **go** through that hell again." "We
d **come** back later," Bill **said**. "I don't want to **sit** here on a night like this."

a yellow, sun-baked color, and I did not want to **leave** the café. But we had to
o loosen. I was very drunk and I did not want to **shut** my eyes because the room
ething else. I **told** the barman I did not want anything to **drink** and **went** out th
e. It was really only two, but I did not want to **explain** after I had made the m
ng with my face away from her. I did not want to **see** her. "**Sent** him for champag
"**Go** on. **Go** on up and **see** him." I did not want to **climb** another flight of stairs

Is there an Englishwoman here? I would like to **see** this English lady." "Muy bue
s to **see** it," the critic **said**. "I would like to **be** convinced." "He doesn't **like**
han your life." What a **speech**! I would like to have it illuminated to hang in t
ve the bottles a twirl. "Still I would like to **hear** you talk some time." "Isn't
other high headland. I **thought** I would like to **swim** across the bay but I was af
I might **pray** for and I **thought** I would like to have some money, so I **prayed** tha

their collocation with frequent phrases confirms the importance of these themes in the text. Conversely, mental concepts are rare among the keywords, with the only examples being the verbs *know* and *want*, which are generally common in fiction and conversation (Biber et al. 2021: 367, 373-374, 378).

However, the analysed phrases themselves include mental verbs that describe emotional or attitudinal states, such as *want* and *like*. This points to an important thematic and stylistic pattern. While mental states and emotions do not emerge as a dominant theme of the novel from the list of keywords, they are granted this status implicitly by their collocation with some of the text's frequent phrases, as well as by being part of singular phrases. This is particularly evident with the verb *want*, which conveys the notions of longing, need, or lack. Its repeated use in the novel, particularly in conjunction with personal pronouns, suggests that the reality depicted in *TSAR* centres on desire rather than fulfilment. The characters seem to be constantly reaching for something that either eludes them or that they cannot realise due to restraints, as indicated by *I don't want to* and *I did not want*. The fact that some of the phrases containing *want* occur in dialogue highlights its role as a social verb, frequently involved in interpersonal interactions in the novel. The meanings and usage patterns of *want* in *TSAR* thus seem to reflect the characters' social and psychological dislocation after the war.

5. Conclusion

This study analysed the keywords and frequent 4-word phrases in *TSAR* to reveal their subtle literary meanings. The findings show that the novel differs from Hemingway's other works and his contemporary literature through its distinctive vocabulary and themes.

The keywords indicate that the novel centres on characters, their conversations, routine actions and movement through familiar urban spaces. This corroborates Elbaz's (1984: 80) assertion that "Hemingway's characters" in *TSAR* "develop functionally to the spaces they inhabit and visit, and the actions they undertake in these spaces". The frequent use of *not*-negation and negatively connotated lexis points to interpersonal tension and emotional strain among the characters, alongside a reluctance to express their desires. The presence of fewer male-referencing keywords than in other Hemingway novels suggests a shift away from a single dominant male protagonist, despite the fact that the narrator remains male. In this narrative world, keywords related to bullfighting form an important motif, gaining prominence through their concentration later in the novel, their rhythmic collocation with the definite article, their placement in narration rather than dialogue, and their appearance in concise, adjective-free sentences. These factors all serve to heighten the symbolic status of bullfighting in the novel.

Analysis of the novel's most frequent phrases corroborates these observations. The recurrent phrases form a small, lexically stable repertoire built largely on verbs and personal pronouns. This highlights the focus on characters, their spoken interactions and their concrete, everyday experiences manifested in past-tense reporting. The lexical items within the phrases and their nearest collocates reinforce these themes. This limited phraseological core shows little variation, reflecting Hemingway's linguistic minimalism and creating stylistic and thematic consistency, and a documentary rhythm throughout the novel. At the same time, the high STTR value indicates a diverse vocabulary, suggesting precise and context-sensitive lexical choices.

Overall, the analysis shows that, despite his stylistic minimalism, Hemingway's use of language in *TSAR* generates statistically significant patterns that enrich existing interpretations of his first novel. Future research could compare the novel with its translations to determine whether Hemingway's limited phraseological repertoire and distinctive stylistic features, such as sparse adjectives and stable 4-frames, survive translation. As the novel's phraseology indexes spatial movement, interpersonal negotiation, volition, and negativity, it would also be interesting to map how these patterns cluster around particular narrative fragments (e.g. before and after the fiesta) or characters, and to examine how they shape the novel's meanings. Finally, cross-novel comparisons could further identify which of *TSAR*'s keyword and phraseological features reappear elsewhere in Hemingway's works.

References

- Baker, P. 2004. Querying keywords: Questions in difference, frequency, and sense in keyword analysis. *Journal of English Linguistics* 32(4): 346-359.
- Balassi, W. 1990. Hemingway's Greatest Iceberg: The Composition of *The Sun Also Rises*. In J. Barbour and T. Quirk (eds.), *Writing the American Classics*, 125-155. Chapel Hill: North Carolina UP.
- Balossi, G. 2014. *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's 'The Waves'*. Amsterdam/Philadelphia: John Benjamins.
- Bestgen, Y. 2018. Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora* 13(2): 205-228.
- Biber, D. 2009. A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275-311.
- Biber, D. 2011. Corpus linguistics and the study of literature Back to the future? *Scientific Study of Literature* 1(1): 15-23.
- Biber, D., and S. Johansson, G. Leech, S. Conrad, E. Finegan 2021. *Grammar of Spoken and Written English*. Amsterdam/Philadelphia: John Benjamins.
- Cappock, E., and D. Beaver 2015. Definiteness and determinacy. *Linguist and Philosophy* 38: 377-435.
- Culpeper, J. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics* 14: 29-59.
- Dewberry, E. 1996. Hemingway's journalism and the realist dilemma. In S. Donaldson (ed.), *The Cambridge Companion to Hemingway*, 16-35. Cambridge: Cambridge University Press.
- Elbaz, R. 1984. The mechanics of repetition in the discourse of Ernest Hemingway. *Zagadnienia Rodzajów Literackich* 27(1): 77-90.
- Evison, J. 2010. What are the basics of analysing a corpus? In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*, 122-135. London: Routledge.
- Fisher-Starcke, B. 2009. Keywords and frequent phrases of Jane Austen's *Pride and Prejudice*: A corpus-stylistic analysis. *International Journal of Corpus Linguistics* 14(4): 492-523.
- Fischer-Starcke, B. 2010. *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. London: Continuum.
- Fletcher, W. H. 2012. *KfNgram* (Software). Annapolis: USNA. Available at: <http://www.kwicfinder.com/kfNgram/kfNgramHelp.html> (accessed 12 September 2025).
- Fuster-Márquez, M. 2014. Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction* 7(1): 84-121.
- Garner, J. 2016. A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. *International Journal of Learner Corpus Research* 2(1): 31-67.

- Goźdz-Roszkowski, S. 2011. *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study*. Frankfurt am Main: Peter Lang.
- Gray, B., D. Biber 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics* 18(1): 109-135.
- Green, C. 2017. Introducing the *Corpus of the Canon of Western Literature*: A corpus for culturomics and stylistics. *Language and Literature: International Journal of Stylistics* 26(4): 282-299.
- Ho, Y. 2011. *Corpus Stylistics in Principles and Practice: A Stylistic Exploration of John Fowles' 'The Magus'*. London: Continuum.
- Hunston, S., and G. Francis 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.
- Keskesay (2021, April 24). *Ernest Hemingway, journalism in fiction*. Medium. <https://medium.com/@keskesay productions/ernest-hemingway-journalism-in-fiction-3492d3f7a1aa>.
- Leech, G., and M. Short 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Harlow: Pearson Education.
- Mahlberg, M. 2007. Clusters, key clusters and local textual functions in Dickens. *Corpora* 2: 1-31.
- Mahlberg, M. 2013. *Corpus Stylistics and Dickens's Fiction*. New York and London: Routledge.
- Rothman, N.L. (1928, April 1). Hemingway Whistles in the Dark. *The Dial* 84: 336-338. https://upload.wikimedia.org/wikipedia/commons/e/e3/The_Dial_%28Volume_84%29.pdf.
- Römer, U. 2010. Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews. *English Text Construction* 3(1): 95-119.
- Scott, M. 2012. *WordSmith Tools (version 6.0)*. Stroud: Lexical Analysis Software.
- Scott, M., and C. Tribble 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Shepherd, T.M.G., and T. Berber Sardinha 2013. A rough guide to doing corpus stylistics. *Matraga* 20(32): 66-89.
- Short, M. 1996. *Exploring the Language of Poems, Plays and Prose*. London: Longman.
- Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J.M. 2008. The phrase, the whole phrase, and nothing but the phrase. In S. Granger and F. Meunier (eds.), *Phraseology: An Interdisciplinary Perspective*, 407-410. Amsterdam: John Benjamins.
- Starcke, B. 2006. The phraseology of Jane Austen's *Persuasion*: Phraseological units as carriers of meaning. *ICAME Journal* 30: 87-104.
- Stubbs, M. 2005. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature* 14(1): 5-24.

- Stubbs, M. 2007. Quantitative data on multi-word sequences in English: the case of the word *world*. In M. Hoey and M. Mahlberg, M. Stubbs, W. Teubert (eds.), *Text, Discourse and Corpora*, 163-190. London: Continuum.
- Stubbs, M., and I. Barth 2003. Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language* 10(1): 65-108.
- Svoboda, F.J. 1983. *Hemingway & The Sun Also Rises: The Crafting of a Style*. Lawrence, Kansas: University Press of Kansas.
- Trodd, Z. 2007. Hemingway's camera eye: The problem of language and an interwar politics of form. *The Hemingway Review* 26(2): 7-21.
- Wagner-Martin, L. 1990. Introduction. In L. Wagner-Martin (ed.), *New Essays on Sun Also Rises*, 1-18. New York: Cambridge University Press.

Appendix. The literary texts included in the corpus *ContempLit*

Jack London, *Adventure* (1911)
 Edith Wharton, *The Age of Innocence* (1920)
 Willa Cather, *A Lost Lady* (1923)
 E.M. Forster, *A Passage to India* (1924)
 F. Scott Fitzgerald, *The Great Gatsby* (1925)
 Theodore Dreiser, *An American Tragedy* (1925)
 Thornton Wilder, *The Bridge of San Luis Rey* (1927)
 William Faulkner, *As I Lay Dying* (1930)
 John Fante, *The Road to Los Angeles* (1933)
 Gertrude Stein, *The Autobiography of Alice B. Toklas* (1933)
 John Steinbeck, *Of Mice and Men* (1937)
 Margaret Mitchell, *Gone with the Wind* (1939)
 Carson McCullers, *The Heart is a Lonely Hunter* (1940)
 Sinclair Lewis, *Kingsblood Royal* (1947)
 Ralph Ellison, *Invisible Man* (1947)
 J.D. Salinger, *The Catcher in the Rye* (1951)
 James Baldwin, *Go Tell It on the Mountain* (1953)
 Ray Bradbury, *Fahrenheit 451* (1953)
 Jack Kerouac, *On the Road* (1957)
 Harper Lee, *To Kill a Mockingbird* (1960)
 John Updike, *Rabbit Run* (1960)
 Joseph Heller, *Catch-22* (1961)
 Ken Kesey, *One Flew Over the Cuckoo's Nest* (1962)
 Sylvia Plath, *The Bell Jar* (1963)
 Saul Bellow, *Hertzog* (1964)
 Cormac McCarthy, *The Orchard Keeper* (1965)
 Anne Tyler, *Tin Can Tree* (1965)
 James Salter, *A Sport and a Pastime* (1967)
 Philip Roth, *Portnoy's Complaint* (1969)
 Kurt Vonnegut, *Slaughterhouse-Five* (1969)