ปีที่ 1 ฉบับที่ 5 (กันยายน - ตุลาคม) 2569
Vol. 1 No. 5 (September - October) 2026
วารสารศึกษาครุทรรศน์
Journal of Education and Learning Perspectives

A Corpus-Based Analysis of the Near-Synonyms *'Teenager'*, *'Minor'*, *'Youth'*, and *'Adolescent'* in COCA

Papapin Rattanarot[1], Anchana Sriraungrith[2], Natchaya Humna[3], Sapolachet Prachumchai[4]
[1-4] English Program, Phetchabun Rajabhat University
*Corresponding Author's Email: Sapolachet@gmail.com
Received: 2026-07-01; Revised: 2026-09-15; Accepted: 2026-09-15

******

Abstract

Although *Teenager*, *Minor*, *Youth*, and *Adolescent* are often treated as interchangeable terms for young people, corpus evidence suggests otherwise. This study compares the four nouns through lexicographic analysis and corpus-based methods, drawing on definitions from Oxford, Cambridge, Merriam-Webster, and American Heritage alongside frequency, genre, collocation, and prosody data from the Corpus of Contemporary American English (COCA). Roughly 200 concordance lines per noun were sampled for semantic preference and prosody analysis. *Youth* and *Adolescent* cluster heavily in academic prose (120.44 and 59.73 occurrences per million words), *Teenager* dominates news and media, and *Minor* is concentrated in legal and formal registers. Collocational patterns show minor linked to legal protection, *Adolescent* to clinical and developmental discourse, and *Teenager* to social and behavioral themes. *Youth* and *Teenager* carry mostly positive semantic prosody, while *Minor* and *Adolescent* remain largely neutral. None of the four nouns proves freely substitutable for another, a finding with implications for EFL instruction, vocabulary indexing, and language tool development.

Keywords: corpus-based study, near-synonymous nouns, Corpus of Contemporary American English (COCA), semantic prosody, collocational analysis

Introduction

Choosing the right word among near-synonyms remains one of the more persistent difficulties EFL learners face. Words sharing a core meaning can still diverge sharply in context, tone, and grammar; Alanazi (2022), for instance, found that learners routinely treat *affect* and *impact* as equivalent despite their distinct syntactic behavior. Recent scholarship continues to probe this challenge. Laosrirattanachai and Laosrirattanachai (2025) showed that even advanced learners struggle to distinguish *attempt, endeavor, strive,* and *try*, four verbs sharing a core sense of purposeful action, owing to divergent genre distributions, formality levels, and collocational patterns. Similarly, Assawawattanasuntorn and Siriganjanavong (2025) found that the synonymous adjectives *careful, cautious,* and *wary*

occupy distinct registers and carry markedly different evaluative tones, with *wary* showing a consistently negative semantic prosody. Although *teenager, minor, youth,* and *adolescent* are commonly listed as interchangeable in dictionaries, their actual distribution in authentic discourse suggests otherwise.

This distinction can be accounted for through several interconnected dimensions. Collocation reflects words' tendency to habitually co-occur (Firth, 1957; Intarajak et al., 2025). Colligation and semantic preference delineate the syntactic constraints and thematic domains a word selects, respectively (Begagić, 2013; Laosrirattanachai & Laosrirattanachai, 2025; Petcharat & Phoocharoensil, 2017). Semantic prosody, meanwhile, involves the evaluative aura embedded within a word through usage (Louw, 1993), operating alongside register and formality to determine pragmatic fitness (Chalermsook, 2014; Jarrar et al., 2024). Together, these variables explain why seemingly interchangeable words display divergent behavioral patterns in actual texts.

Recent corpus-based studies have refined these frameworks. Wei (2025) proposed a phraseological pattern analysis for discriminating English near-synonyms, showing that examining each word alongside its surrounding lexical environment—rather than context alone—yields finer-grained distinctions, complementing traditional Behavioral Profile methods. Laosrirattanachai and Phoocharoensil (2025) further showed that near-synonyms sharing all semantic preference categories—such as *booking* and *reservation*—can still differ substantially in their distribution across semantic fields, sharing only 12 of 110 collocates. These findings confirm that definitional overlap does not guarantee interchangeability in usage.

Corpus linguistics offers a robust empirical approach to synonymy. Large-scale corpora such as the BNC and COCA enable rigorous analysis of frequency, genre distribution, collocation, and semantic prosody (Davies, 2009, 2010). Harnchuwongsakul (2025) examined *complex, complicated,* and *intricate* through COCA, finding that while all three denote non-simplicity, *complex* associates with scientific/structural contexts, *complicated* with problems and confusion, and *intricate* with artistry and detailed patterns. Phumeethong (2025) applied a similar method to *resident, inhabitant,* and *population*, showing that COCA-based frequency, formality, and collocational analyses can distinguish nouns dictionaries treat as equivalents. Julul and Al-Marrani (2025) demonstrated that genre distribution within COCA strongly predicts synonym selection among English verbs denoting change.

Despite growing literature on adjectives (Intarajak et al., 2025; Assawawattanasuntorn & Siriganjanavong, 2025), verbs (Alanazi, 2022; Laosrirattanachai & Laosrirattanachai, 2025; Shi, 2024; Jarrar et al., 2024), and age-related terms (Jamet & Coupé, 2024), the semantic and pragmatic boundaries among *teenager, minor, youth,* and *adolescent* remain under-researched. While Jamet and Coupé (2024) illuminated the semantic prosody of terms denoting aging populations, comparable analyses of terms for young people are scarce—

despite the growth of youth-related corpora such as the Corpus of Turkish Youth Language (CoTY). The present study addresses this gap through a comprehensive, corpus-based analysis of these four near-synonymous nouns.

Research questions are:

1. How do dictionary definitions of the four nouns diverge or converge across Oxford, Cambridge, Merriam-Webster, and American Heritage?

2. How do these nouns vary in terms of normalized frequency and genre distribution within COCA?

3. What are their most significant collocates when measured by raw frequency versus Mutual Information (MI) scores?

4. What distinct semantic preferences and prosodic orientations characterize each of the four nouns?

The corresponding objectives are to compare dictionary definitions across the four sources, analyze frequency and genre distribution in COCA, identify significant collocates using frequency and MI scores, and categorize semantic preference and prosody from approximately 200 sampled concordance lines per noun.

Literature Review

Synonyms

Synonyms share identical or closely related meanings and play a key role in language learning and precise communication (Fitrah and Afria, 2024). Palmer (1981) argued that absolute synonyms - words interchangeable in every context without any shift in meaning or register - are rare in natural language, a point reinforced by Cruse (1986: 270), who observed that natural languages “abhor absolute synonyms just as nature abhors a vacuum.” Murphy (2003) added that lexical relations form a system shaped by both mental organization and pragmatic context. Recent corpus-based work supports this view empirically: Shi (2024) showed that synonymous verbs meaning “to start” differ in frequency, grammar, and context, and Intarajak et al. (2025) found that *expensive* and *costly*, despite sharing a core meaning, differ in the nouns they modify and the registers they appear in. The present study extends this line of work to *teenager*, *minor*, *youth*, and *adolescent*, a noun group not yet examined from this angle.

Dictionaries and Their Limits

Dictionary definitions are a useful starting point but have well-known limitations for capturing nuanced distinctions. Hartmann and James (1998) outlined the principles of dictionary compilation while noting that monolithic structures constrain semantic depth, and Yong (2022) traced lexicography's shift from prescriptive to descriptive and cognitive models, observing that even modern dictionaries struggle with contextual variation. Fuertes-Olivera and Tarp (2014) warned against over-reliance on a single general dictionary, and Łukasik

(2018) called for frameworks better suited to context-specific usage. For this study, these limitations mean dictionary comparison alone cannot distinguish the four target nouns adequately - corpus data is needed to fill in what dictionaries leave unspecified.

Corpus of Contemporary American English (COCA)

COCA is a genre-balanced monitor corpus of over 385 million words spanning spoken language, fiction, magazines, newspapers, and academic writing (Davies, 2009). Davies (2010) demonstrated its value for tracking ongoing language change, and Davies and Gardner (2010) used it to compile frequency and collocation benchmarks. More recent applications include Deliana and Anggita's (2024) collocational analysis of *explored* and Ayyat's (2024) demonstration that COCA's search tools can improve ESL writing accuracy. This study uses COCA as its primary source for frequency, genre, collocation, and prosody data.

Genre

Swales (1990) defined genre as a recognizable variety of language organized around shared communicative purposes, and Biber and Conrad (2009) argued that situational context strongly motivates a writer's lexical choices - meaning genre is a strong predictor of which synonym gets used. Julul and Al-Marrani (2025) found that synonymous change-verbs cluster in formal academic writing while the general verb change spans multiple registers, and Uba and Al-Dhahli (2026) showed that among synonymous acquisition verbs, *take* dominates informal genres such as television and speech. This study applies the same framework to examine how the four target nouns distribute across COCA's eight genre sections.

Collocation

Firth's (1957: 11) principle that a word's meaning is shaped by “the company it keeps” underlies corpus approaches to collocation (Huan, 2025). Siepmann (2005) argued for extending collocational analysis to longer spans and holistic units, Hyland (2008) showed that lexical bundles help distinguish disciplinary discourse, and Durrant and Schmitt (2010) found that second language learners can internalize collocational patterns through exposure. At the level of near-synonymy, Intarajak et al. (2025) showed that collocational data can reveal usage restrictions invisible to dictionary definitions - a finding that underpins the L4-R4 collocational analysis (four words to either side of the target noun) conducted here, using both raw frequency and MI scores.

Semantic Preference and Prosody

Sinclair (2004) argued that meaning extends beyond individual words into recurring phraseological patterns, and Stubbs (2001) showed these patterns reflect habitual lexical choices and cultural expectations. Louw (1993) demonstrated that semantic prosody can reveal a writer's implicit attitude or even irony. Applied work includes Petcharat and Phoocharoensil's (2017) study of near-synonymous adjectives, Jamet and Coupé's (2024) analysis of terms for aging people - which found strongly directional attitudinal patterns -

and Uddin's (2026) study of disciplinary variation in the prosody of *value* in academic writing. This study applies the same framework to the younger-demographic noun cluster, an area that has received little attention in the prosody literature.

Conceptual Framework

Figure 1 summarizes the analytical framework. The four target nouns (teenager, minor, youth, adolescent) are examined through five lenses - dictionary definitions, synonymy theory, COCA frequency/genre data, collocation, and semantic preference/prosody - each grounded in the theoretical sources reviewed above.

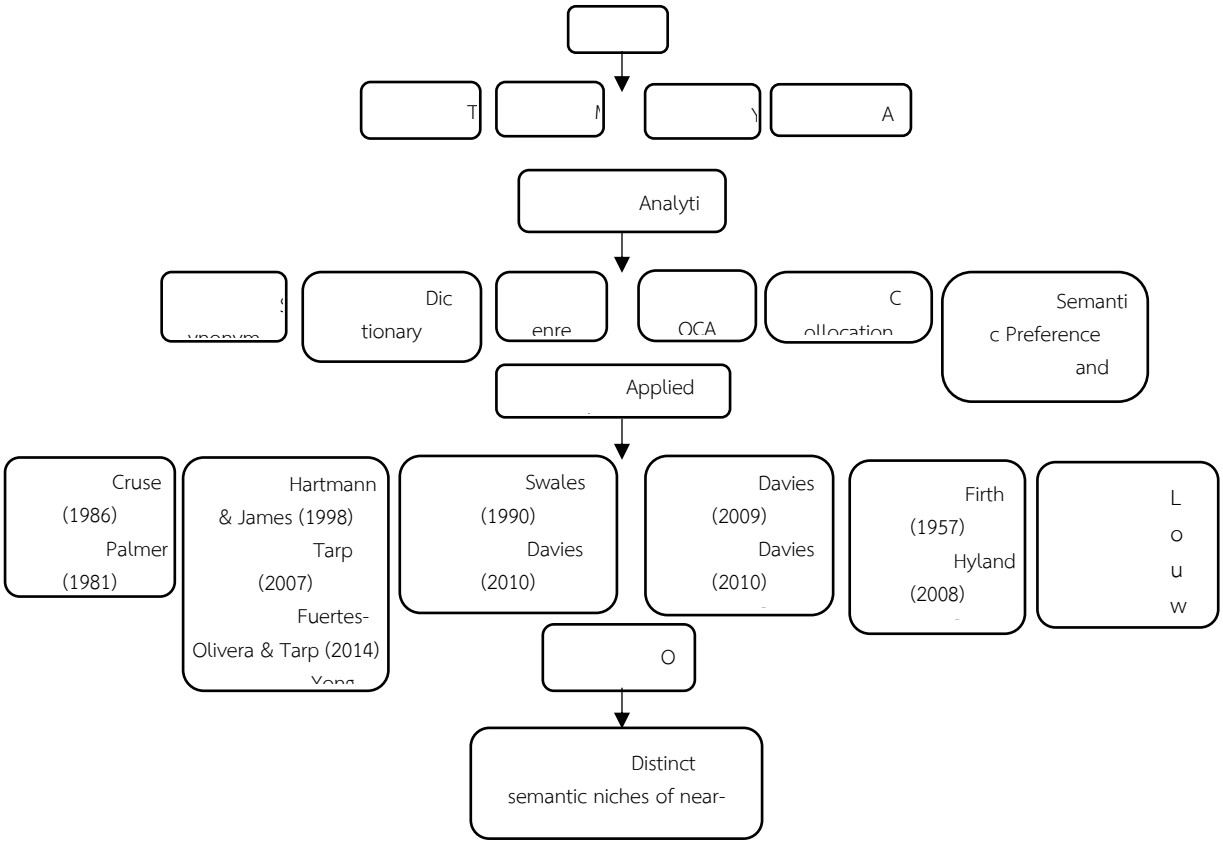


Figure 1 Conceptual framework synthesized by the authors

Methodology

This study combined lexicographic comparison with corpus-based analysis to examine the semantic similarities and differences among teenager, minor, youth, and adolescent.

Data Sources

Definitional profiles were drawn from four major online dictionaries representing both British (Oxford, Cambridge) and American (Merriam-Webster, American Heritage) traditions, chosen for their authority and broad use in pedagogy and research (Hartmann and James, 1998). All corpus data came from COCA. Although British corpora like the BNC offer valuable comparative data, COCA was selected for its large, genre-balanced representation of

contemporary American English, which allows for reliable statistical analysis across eight distinct sections (spoken, fiction, magazine, news, academic, blog, web, and TV/movies) (Davies, 2009). Its size is sufficient to yield stable results for all four nouns, leaving cross-dialectal comparisons for future research.

Dictionary Comparison

Entries for each noun were retrieved from the four platforms in January 2025. Core meanings, age boundaries, usage labels, and examples were recorded and compared to identify shared semantic cores and differences in definitional scope and disciplinary framing.

Frequency and Genre Distribution

Normalized frequencies (per million words) and genre distributions were extracted using COCA's search interface. To ensure comparability, searches were conducted for all noun forms (lemmas) of the target words. For *minor*, searches were filtered by the noun part-of-speech tag to exclude adjectival uses. Normalization follows standard register-analysis practice (Biber and Conrad, 2009) and allows fair comparison across genre sections of unequal size.

Collocational Analysis

Collocates were retrieved within a span of four words on either side of each noun (L4-R4), with both raw frequency and MI scores recorded. Frequency-ranked tables list the most frequent collocates regardless of MI; MI-ranked tables apply the standard threshold of MI ≥ 3.0 to highlight statistically distinctive associations. Collocates are grouped into nouns, adjectives, verbs, and adverbs.

Semantic Preference and Prosody

Approximately 200 concordance lines per noun (801 total) were randomly sampled from COCA. To align with the study's focus on human referents, a manual screening process was applied to the concordance lines of *minor* to exclude non-human senses (e.g., *minor key*, *minor league*). The remaining lines were classified along two dimensions using a purpose-built Python tool.

The first dimension was semantic preference (the thematic domain of surrounding collocates, e.g., Person, Group, Act, Cognition, Time). The second dimension was semantic prosody (overall evaluative orientation). Because semantic prosody is traditionally a qualitative concept, VADER sentiment scoring was employed as a supportive quantitative tool to provide an objective baseline of evaluative orientation. The automated compound scores (above 0.05 coded positive, below -0.05 negative, and otherwise neutral, following Hutto and Gilbert, 2014) served as a preliminary indicator, which was subsequently verified and refined through manual qualitative review by the coders to ensure pragmatic and contextual nuances were captured. Classifications were reviewed manually, and a 20% subsample (160 lines) was independently rated by a second coder with a linguistics

background after a one-hour orientation session; disagreements were resolved through discussion. Cohen's kappa reached 0.87, indicating strong agreement.

Tools and Ethical Considerations

COCA's online interface and built-in MI calculator served as the primary platform; the semantic analysis tool was implemented in Python 3.10 and applied offline. All search parameters (lemma searches, POS tagging for *minor*, L4-R4 span, MI ≥ 3.0 threshold) are reported to allow replication. All COCA texts are publicly available academic and journalistic materials; short excerpts are reproduced for illustration only, with full referencing, in line with fair-use and institutional research ethics guidelines.

Results

Results are presented in four subsections corresponding to the four research questions: dictionary definitions, frequency and genre distribution, collocational analysis, and semantic preference and prosody.

1. Dictionary Definitions

Table 1 Comparative dictionary definitions of the four nouns

Feature	Teenager	Minor	Youth	Adolescent
Core definition	Person aged 13-19	Person not yet a legal adult	Period of youth; young people as a group	Person developing from childhood to adulthood
Age boundary	Explicit: 13-19	Varies by jurisdiction	Not specified	Not specified; post-puberty (American Heritage)
Disciplinary frame	Demographic	Legal	Social / temporal	Developmental / biological
Shared core	Age 13-19 (all 4)	Legal underage status (all 4)	Period of being young (all 4)	Transitional developmental stage (all 4)

All four nouns share a semantic core referring to young people in a transitional life stage, but each is framed through a different disciplinary lens. *Teenager* is the only noun with an explicit numerical age boundary (13-19), making it demographically precise but contextually neutral. *Minor* is defined purely through legal status across all four dictionaries, with no fixed age - reflecting how the threshold for adulthood varies by jurisdiction - which predicts a strong link to statutory and institutional discourse. *Youth* has the broadest scope, covering a life period, a personal condition, and a social collective, and is the only noun that can denote either an individual or a group. *Adolescent* is grounded in biological and developmental discourse, with American Heritage specifying puberty as the defining threshold - a framing that anticipates its concentration in clinical and academic registers. These distinctions suggest the four nouns differ not just in meaning but in the kind of

knowledge domain from which they draw authority, a pattern tested against corpus data below.

2. Frequency and Genre Distribution

Table 2 Normalized frequencies (per million words) across COCA genre sections

Noun	Blog	Web	TV/M	Spok	Fic	Mag	News	Acad	All
Teenager	11.10	12.96	9.03	15.74	15.49	16.11	17.50	3.56	12.69
Minor	35.14	39.70	14.98	20.22	22.51	35.83	42.15	38.29	31.04
Youth	39.21	49.94	16.27	24.20	35.68	48.40	71.31	120.44	50.17
Adolescent	3.83	5.97	1.81	3.07	6.68	8.45	5.73	59.73	11.64

Teenager is distributed fairly evenly across public-facing genres, peaking in News (17.50 per million words) and Magazine (16.11), but drops sharply in Academic writing (3.56) - consistent with its demographic framing as the word journalists and general writers reach for, but one too informal for scholarly use.

Minor demonstrates a high normalized frequency overall (31.04 per million words). This frequency reflects its strong legal currency and broad reach across formal written registers, peaking in News (42.15), Web (39.70), and Academic (38.29), while its relatively lower frequency in TV/Movies (14.98) confirms its formal, written-discourse profile.

Youth shows the sharpest genre skew of the four: its Academic frequency (120.44) is nearly double its next-highest genre, News (71.31), and more than seven times its TV/Movies figure (16.27). This reflects its role as a broad social category suited to policy and analytical discussion - youth development, youth unemployment, youth culture - rather than reference to a specific individual.

Adolescent shows the sharpest academic concentration relative to its overall frequency: 59.73 per million words in Academic writing, far above its next-highest genre, Magazine (8.45), and below 7 per million words everywhere else, reaching its lowest in TV/Movies (1.81). This strongly confirms the biological and developmental framing suggested by the dictionary analysis - adolescent is primarily a term of clinical and scholarly writing.

Together, these distributions reveal a formality hierarchy. Adolescent and youth lean academic, with adolescent more narrowly confined to scholarly registers and youth more broadly spread across formal writing; minor spans formal written genres (news, academic) due to its legal function; and teenager is the only noun performing consistently across general public-facing genres while remaining scarce in academic writing.

3. Collocates by Frequency and Mutual Information

Collocates were retrieved within an L4-R4 span for each noun. Frequency-ranked lists show the most frequent collocates regardless of MI value; MI-ranked lists show only collocates meeting the threshold of MI ≥ 3.0, identifying statistically distinctive associations rather than words that simply co-occur due to general high frequency in the corpus.

Teenager

Frequent noun collocates include *child*, *adult*, and *parent*; the most MI-distinctive is *toddler* (4.59). Among adjectives, *young* and *black* are frequent, but *unarmed* (7.67) and *rebellious* (6.90) are the most distinctive - alongside *troubled* (5.23) and *pregnant* (4.06). Verb collocates center on *kill*, *shoot*, and *recruit* (3.96), with *abuse* (3.21) also MI-distinctive. Adverbs include *sexually* (4.54) and *fatally* (5.03).

Minor

The standout collocate is *unaccompanied* (MI 12.43), the highest adjective MI in the entire dataset, alongside *parental* (6.91), *harmful* (6.73), and *sexual* (5.16). Noun collocates cluster around *major* (6.92), *consent* (6.26), *abuse* (5.46), *alcohol* (5.35), and *abortion* (5.21). Verbs include *abuse* (6.08), *pitch* (4.74), *ban* (4.54), and *sell* (3.48). Adverbs include *sexually* (6.59) and *illegally* (5.60).

Youth

Youth produces the most diverse profile of the four: institutional nouns (*program*, *center*, *hostel* at MI 7.86) sit alongside metaphorical or idiomatic ones (*fountain*, MI 6.82; *misspent*, MI 9.61). Adjectives include *at-risk* (8.07), *incarcerated* (7.79), *homeless* (5.68), and *foster* (6.05). Verbs cluster around *mentor* (5.98) and *educate* (3.78).

Adolescent

Adolescent produces the most internally coherent profile, clustered around clinical and developmental terms: *literacy* (6.24), *self-esteem* (6.03), *impairment* (5.85), and *parenting* (5.77) among nouns; *at-risk* (7.16), *authoritative* (7.01), *gifted* (6.07), *overweight* (5.65), and *developmental* (5.45) among adjectives. Verbs (*report*, *perceive*, *experience*, *indicate*, *engage*) reflect the evidential, hedged language of empirical research.

4. Semantic Preference and Prosody

Teenager (201 lines, mean length 23.4 words) Dominant preference categories are Cognition (16.1%), Group (12.5%), and Attribute (11.6%), reflecting its use in describing attitudes and behaviors attributed to young people as a class, and its collective/media usage. Prosody is weakly positive (39.9% positive, 34.5% neutral, 25.6% negative; mean compound 0.017), indicating that teenager appears across a wide evaluative range without a strong directional charge.

Minor (200 lines, mean length 23.1 words)

Following the manual exclusion of non-human senses, Time remains a notable category (15.3%), reflecting procedural and statutory timeframes (e.g., age of consent laws). This is followed by Person (12.0%) and Act (10.0%), which reflect legal-protection discourse. Prosody is weakly neutral (40.5% neutral, 34.4% positive, 25.1% negative; mean compound 0.013), consistent with the factual, procedural register in which minor typically appears.

Youth (200 lines, mean length 22.5 words)

Person (26.2%) dominates, followed jointly by Group and Act (19.1% each), reflecting its dual individual/collective function and its frequent appearance in activity- and program-

related contexts. Prosody is the most positive of the four (41.1% positive, 37.0% neutral, 21.9% negative; mean compound 0.040), consistent with its institutional, community-oriented collocates (mentoring, education, participation).

Adolescent (200 lines, mean length 22.0 words)

Person (29.7%) dominates, followed by Act (17.9%) and All (11.3%), reflecting its consistent use as a human referent in academic and clinical discourse. Notably, 7.55% of lines fall into the Animal category - manual inspection shows these are developmental and neuroscientific texts using animal-model studies (e.g., rodent studies of adolescent brain development) as comparative evidence, a genuine feature of this clinical literature rather than misclassification. Prosody is weakly neutral (38.1% neutral, 37.7% positive, 24.2% negative; mean compound 0.021), consistent with the measured, evaluatively restrained language of academic writing.

Discussion

The findings demonstrate that teenager, minor, youth, and adolescent are not freely interchangeable—extending the theoretical claims of Palmer (1981) and Cruse (1986, p. 270) by providing empirical, corpus-based evidence of the precise dimensions along which these nominal near-synonyms diverge.

Lexicographic Framing and Corpus Behavior

The dictionary comparison identified four conceptual frames - demographic (teenager), legal (minor), social/temporal (youth), and developmental/biological (adolescent) - and the corpus data shows these frames actively predict real usage. Minor's legal framing accurately predicts its collocational profile (consent, unaccompanied, abuse, alcohol, abortion) and its concentration in news and academic writing. Similarly, American Heritage's developmental framing of adolescent (the only dictionary specifying puberty as a threshold) anticipates its strong association with clinical terms like self-esteem, impairment, and authoritative parenting.

The corpus also reveals what dictionaries miss. Youth's breadth across three distinct sub-corpora - institutional programs, social marginalization, and the idiomatic “life period” sense - is not reflected in any single dictionary entry. Likewise, no dictionary foregrounds the risk and victimization discourse that is statistically the most distinctive feature of teenager, evident in high-MI collocates like unarmed (7.67) and rebellious (6.90), or recruit (3.96) and abuse (3.21). These findings support Fuertes-Olivera and Tarp's (2014) argument that reliance on a single dictionary produces an incomplete picture, reinforcing the case for corpus-informed reference tools as a complement to traditional lexicography.

Genre Constraints on Synonym Selection

The genre data reveal a clear hierarchy: Adolescent is the most register-restricted (59.73 per million in academic writing vs. under 7 everywhere else); Youth shares the academic orientation but distributes more broadly (71.31 in news, plus meaningful presence in magazine, web, and blog); Minor performs strongly in both news (42.15) and academic (38.29) writing, reflecting its dual legal/journalistic function; and Teenager is the only noun with consistent frequency across public-facing genres while remaining scarce academically (3.56).

The pedagogical implication is straightforward: genre is an important guide to choosing among these four nouns. Academic writers should default to adolescent or youth depending on whether the emphasis is clinical/individual or social/collective; journalists covering legal cases should use minor when legal status matters and teenager when it doesn't. This aligns with Swales' (1990) argument that genre-specific communicative purposes constrain lexical choice, and with Uba and Al-Dhahli's (2026) finding that synonym selection is highly sensitive to register.

Collocational Profiles and Discourse Worlds

Each noun inhabits a distinct discourse world. Teenager constructs the young person as a social actor embedded in risk - distinctive adjectives (*unarmed, rebellious, troubled, pregnant*) and verbs (*recruited, abused, shot, arrested*) reflect news coverage of crime and youth-related controversy rather than teenagers themselves.

Minor constructs the young person as a legal category. The extraordinarily high MI of unaccompanied (12.43) - the highest in the entire dataset - reflects the volume of policy text generated by debates over unaccompanied minors in the U.S., while the verb cluster (*abuse, sell, ban, pitch*) frames the minor as a passive recipient of institutional action, consistent with its statutory function.

Youth has the most diverse profile, a direct consequence of its semantic breadth. Its institutional cluster (*program, mentor, educate, participate, coach*) frames young people as participants in community structures - a sharp contrast to *Teenager's* “what goes wrong” framing, reflecting instead a “what can be done to help” orientation. The presence of misspent (MI 9.61) and fountain (MI 6.82) also reflects metaphorical uses absent from the other three nouns.

Adolescent has the most internally coherent profile, clustered around psychological assessment and clinical intervention - *self-esteem, impairment, distress, disorder, gifted, overweight, authoritative, developmental* - alongside evidential verbs (*report, perceive, experience, indicate, examine*) characteristic of empirical research writing. This supports Hyland's (2008) argument that collocational patterns help distinguish and maintain disciplinary discourse. The Animal category (7.55%) in Adolescent's semantic preference profile - absent from the other three nouns - reflects the genuine use of animal-model studies (particularly rodent brain-development research) as comparative evidence in this

clinical literature, further confirming that adolescent operates closer to biological science than to general language use.

Semantic Prosody and Evaluative Orientation

The prosodic findings are partly consistent with expectations and partly surprising. *Youth* and *Teenager* are predominantly positive (41.1% and 39.9% respectively), aligning with the general intuition that terms for young people carry optimistic connotations. *Minor* and *Adolescent* are predominantly neutral (40.5% and 38.1%).

However, all four nouns show only weak prosodic strength (compound scores 0.013-0.040) - a contrast with Jamet and Coupé's (2024) finding of strongly directional, often negative or condescending prosodies for terms referring to aging people. This relative neutrality likely reflects the functional versatility of youth-related terms: unlike terms for elderly people, which cluster in evaluatively loaded social commentary, terms for young people circulate across a much wider range of text types, diluting the overall prosodic signal.

Teenager's positive orientation deserves particular care in interpretation. Despite its positive compound score (0.017), its collocational profile is dominated by risk and vulnerability themes. The explanation likely lies in the structure of news discourse: sympathetic framing of young victims - for example, a community rallying around a vulnerable teenager - generates positive sentiment from words like rallied or community, even when the underlying event involves harm. This is a methodological point that future work using automated sentiment tools should bear in mind; while VADER provides a useful quantitative proxy for semantic prosody, it may misinterpret sympathetic surrounding language as a positive prosody despite inherently negative subject matter.

The evidence across all four dimensions—lexicographic definitions, genre distributions, collocational behaviors, and prosodic orientations—converges on a single conclusion: these four target nouns are near-synonyms only insofar as they refer to young individuals. Lexically, they operate within disparate registers, fulfill distinct communicative purposes, and project different evaluative associations. Thus, substituting one for another inevitably triggers a significant shift in semantic nuance, register, or contextual appropriateness.

Conclusion

This investigation examined the synonymy of teenager, minor, youth, and adolescent through dictionary profiles and empirical corpus data from COCA. The data across all research inquiries reveals a clear negative: these nouns are by no means freely interchangeable. Although anchored in a shared conceptual core regarding young people, each noun maintains a specific functional distribution defined by genre preferences, collocational profiles, and evaluative coloring. In terms of register, teenager is prevalent in journalistic discourse and linked to social themes; minor is heavily restricted to legal and

formal texts; youth functions as a flexible, multi-purpose term spanning both singular and collective reference; and adolescent operates as a highly technical variant within psychological and medical research.

Body of New Knowledge

By integrating lexicographic analysis with corpus-based insights, this study offers a novel, multi-dimensional comparison of four nominal near-synonyms within a single framework. The empirical data reveals that standard dictionary profiles fall short of capturing nuanced real-world applications, such as the heavy association of teenager with risk and victimization or the idiomatic flexibility of youth. Additionally, this research advances demographic prosody literature by showing that terms for young people exhibit much milder evaluative loading compared to those for older populations (cf. Jamet & Coupé, 2024). The discovery that a word can project a positive prosody despite being embedded in risk-dominated collocational environments presents a critical caveat for future sentiment analysis research.

Recommendations

For Practice

The findings offer significant implications for pedagogy, lexicography, and language technology. First, EFL instruction needs to move beyond synonymous presentation, focusing instead on register-specific applications; for instance, syllabus designers should align adolescent and youth with academic genres, minor with legal texts, and teenager with journalistic narratives. Second, lexicographers and dictionary developers should supplement core definitions with empirical genre and collocational data, especially for nouns like teenager and youth where standard entries fail to reflect actual usage. Lastly, creators of language technology tools—such as writing assistants and machine translation engines—should note that automated sentiment scoring may distort the true evaluative valence of texts containing teenager, owing to the misalignment between its positive semantic prosody and its risk-dominated collocational environment.

For Future Studies

To further advance this line of inquiry, future studies could expand the current framework in three main directions. First, replicating this methodology using British English data (e.g., the BNC) would clarify whether these nominal distributions reflect variety-specific nuances or broader cross-dialectal patterns. Second, future researchers could broaden the study's lexical domain by incorporating terms like kid, child, and juvenile to sketch a more exhaustive semantic map of age-related nouns. Lastly, methodologically, there is a critical need for sentiment-analysis frameworks capable of contextual collocational decoding. Developing systems that look beyond compound sentence-level scores will refine language

technologies in handling texts where positive semantic prosody masks inherently negative or risk-dominated topics.

References

Alanazi, Z. (2022). Corpus-based analysis of near-synonymous verbs. *Asian-Pacific Journal of Second and Foreign Language Education, 7*(1), 1–25. https://doi.org/10.1186/s40862-022-00138-5

Assawawattanasuntorn, P., & Siriganjanavong, V. (2025). A corpus-based study of English near-synonyms: Careful, cautious, and wary. *LEARN Journal: Language Education and Acquisition Research Network, 18*(2), 869–892. https://doi.org/10.70730/UAXF5168

Ayyat, A. (2024). Implementing a three-stage approach to teach writing skills using the Corpus of Contemporary American English in a foundation program context. *Arab World English Journal, 15*(4), 241–254. https://doi.org/10.24093/AWEJ/VOL15NO4.15

Begagić, M. (2013). Semantic preference and semantic prosody of the collocation make sense. *Jezikoslovlje, 14*(2), 403–416. https://hrcak.srce.hr/file/165524

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Chalermsook, P. (2014). *A corpus-based study of English verbs: Increase and raise.* [Independent study paper, Thammasat University]. https://digital.library.tu.ac.th/tu_dc/frontend/Info/item/dc:142413

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Davies, M. (2009). The 385+ million-word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists*. Routledge.

Deliana, & Anggita, M. (2024). English collocations of the verb “explored” on the Corpus of Contemporary American English (COCA): A corpus-based analysis. *TALENTA Conference Series: Local Wisdom, Social, and Arts (LWSA), 7*(2), 269–276. DOI: https://doi.org/10.32734/lwsa.v7i2.2085

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*(2), 163–188. https://doi.org/10.1177/0267658309349431

Firth, J. R. (1957). *Papers in linguistics 1934–1951*. Oxford University Press.

Fitrah, Y., & Afria, R. (2024). The analysis of synonym relation meaning in Kerinci language: A semantic study. *JOLLT Journal of Languages and Language Teaching, 12*(1), 404–415. https://doi.org/10.33394/jollt.v12i1.9415

Fuertes-Olivera, P. A., & Tarp, S. (2014). *Theory and practice of specialised online dictionaries: Lexicography versus terminography*. De Gruyter Mouton.

Harnchuwongsakul, S. (2025). *A corpus-based analysis of the near-synonyms complex, complicated, and intricate* [Master's thesis, Thammasat University]. https://doi.org/10.14457/TU.the.2025.589

Hartmann, R. R. K., & James, G. (1998). *Dictionary of lexicography*. Psychology Press.

Huan, X. (2025). A literature review of corpus-based research on verb-noun collocation. *Journal of Humanities and Education Development, 7*(4), 120–126. https://doi.org/10.22161/jhed.7.4.13

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

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

Intarajak, K., Kongduang, T., & Bana, C. R. (2025). Exploring near-synonymy of the adjectives "expensive" and "costly" through corpus-based study. *Southeast Asia Language Teaching and Learning (SALTEL) Journal, 8*(1), 1–11. https://doi.org/10.35307/saltel.v8i1.127

Jamet, D., & Coupé, C. (2024). A study of the semantic preference and semantic prosody associated with the denominations of aging people. *Journal of Language and Aging Research, 2*(1), 47–84. https://doi.org/10.15460/jlar.2024.2.1.1183

Jarrar, Y. A., Al-Sayyed, S. W., Rababa'h, B. B., & Alkhawaja, L. A. (2024). A corpus-based analysis of four near-synonymous English verbs. *Forum for Linguistic Studies, 6*(6), 678–701. https://doi.org/10.30564/fls.v6i6.7378

Julul, A. A. H., & Al-Marrani, Y. M. A. (2025). The investigation of the frequency and genre distribution of the English synonymous verbs denoting change in light of the Corpus of Contemporary American English. *Sana'a University Journal of Human Sciences, 4*(4), 629–647.

Laosrirattanachai, P., & Laosrirattanachai, P. (2025). Unveiling the distinction of near synonymy: A corpus-based analysis on attempt, endeavor, strive, and try. *PASAA: Journal of Language Teaching and Learning in Thailand, 70*(1), 132–163. https://doi.org/10.58837/CHULA.PASAA.70.5

Laosrirattanachai, P., & Phoocharoensil, S. (2025). Differentiating the near-synonyms using a corpus-based approach: Booking vs. reservation. *TEFLIN Journal, 36*(2), 350–371. https://doi.org/10.15639/teflinjournal.v36i2/350-371

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and technology: In honour of John Sinclair (pp. 157–176). John Benjamins. https://doi.org/10.1075/z.64.11lou

Łukasik, M. (2018). *Theoretical underpinnings of metaterminography*. *Applied Linguistics Papers, 25*(4), 195–213. DOI:10.32612/uw.25449354.2018.4.pp.195-213

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge University Press.

Palmer, F. R. (1981). *Semantics* (2nd ed.). Cambridge University Press.

Petcharat, N., & Phoocharoensil, S. (2017). A corpus-based study of English synonyms: Appropriate, proper, and suitable. *LEARN Journal: Language Education and Acquisition Research Network Journal, 10*(2), 10–24. https://so04.tci-thaijo.org/index.php/LEARN/article/view/111700

Phumeethong, N. (2025). *A corpus-based study of English synonyms: Resident, inhabitant, and population* [Master's thesis, Thammasat University].

Shi, K. (2024). A corpus-based behavioral profile study of synonymous expressions meaning "to start." *English Language Teaching and Linguistics Studies, 6*(5), 180–195. https://doi.org/10.22158/eltls.v6n5p180

Siepmann, D. (2005). Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. *International Journal of Lexicography, 18*(4), 409–443. https://doi.org/10.1093/ijl/eci042

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. John Wiley & Sons.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tarp, S. (2007). Lexicography in the information age. *Lexikos, 17*, 170–179. DOI:10.5788/17-0-1170

Uba, S. Y., & Al-Dhahli, F. (2026). A corpus-based genre and collocational study of the near-synonyms: Grasp, capture, seize, snatch, and take. *World Journal of English Language, 16*(1), 73-85. https://doi.org/10.5430/wjel.v16n1p73

Uddin, K. M. J. (2026). A corpus-based investigation of collocation, semantic prosody, and semantic preferences of a high-frequency noun value in British Academic Written English (BAWE). *European Journal of Applied Linguistics Studies, 9*(1), 22-36. DOI: 10.46827/ejals.v9i1.679

Wei, L. (2025). Phraseological pattern analysis: A corpus-driven framework for discriminating English near-synonyms. *International Journal of Lexicography, 38*(4), 397–418. https://doi.org/10.1093/ijl/ecaf024