Vilnius University Press

Contents lists available at Vilnius University Press

Respectus Philologicus
2023, no. 44 (49), pp. 11–24

eISSN 2335-2388
DOI: http://dx.doi.org/10.15388/RESPECTUS.2023.44.49.104

Idioms in General English Corpora: on Frequency, Register, and Cross-Variety Variation

Yurii Kovaliuk

Yuriy Fedkovych Chernivtsi National University, Department of English
2 Sadova Street, Chernivtsi 58002, Ukraine
Email: y.kovalyuk@chnu.edu.ua, yuriy_kovalyuk@yahoo.de
ORCID iD: https://orcid.org/0000-0002-9379-2187
Research interests: English phraseology, Cognitive linguistics, Corpus linguistics, World Englishes

Abstract. Research into idioms and phraseology has become an established part of the corpus linguistic research agenda and has often revolved around either corpus-based or corpus-driven methodologies. At the same time, a relatively recent approach to socio-variational aspects of language in the form of Cognitive Sociolinguistics has contributed to establishing an ideal platform for the study of variation in the varieties of English. The present paper rests on these two research strands in a survey devoted to variation on the level of idioms in present-day English, namely those denoting competition. While idioms, first and foremost, are theoretically identified with the frameworks of Phraseology, Cognitive Linguistics, and Applied Linguistics, among others, this study will make use of a corpus-based method of idioms introduced by Moon and Gustawsson's idioms frequency and significance threshold, paired with Moze and Mohamed's sociolinguistic profiling of idioms. The Idioms will be examined in two national varieties of English, namely those spoken in Great Britain and the USA, which are represented in the British National Corpus and the Corpus of Contemporary American English, respectively. With the assumption that the concept of competition is variety-specific, the main questions to be answered during the analysis are: (1) To what extent can the frequency of use of idioms be regarded an element of variation? and (2) Are there any differences in the prominence of specific variables, such as frequency, register, gender, and age across the two varieties under study? The preliminary findings indicate a significant amount of similarity, but upon closer examination of the data, some important variations are emphasised. Thus, a discussion of the results provides a basis for an inter-variety comparison of the idioms denoting competition and, in so doing, adds to the universality / variation debate.

Keywords: idiom; corpus-based study; the BNC; the COCA; idiom variation.

Introduction

When studying variation of idioms and idiomaticity, two crucial aspects are important: "language user variation" and "language use variation" (Murar, 2009). These aspects are closely intertwined, with the former accounting for regional variation, such as British

Submitted 30 March 2023 / **Accepted** 18 June 2023
Įteikta 2023 03 30 / **Priimta** 2023 06 18

English (BrE) and American English (AmE), and social variation, such as social class, age, and sex. The latter aspect is commonly referred to as Pragmatics, which examines the sociocultural context of language users, or Social Cognitive Linguistics (Geeraerts, 2018; Schönefeld, 2022), which is primarily aimed at describing "the social-interactional mechanisms of how usage shapes linguistic knowledge at the level of speaker and hearer" (Divjak et al., 2016).

In her book *Colouring Meaning. Collocation and connotation in figurative language*, Gill Phillip vividly illustrates the intricacies of variation particularly in relation to corpus studies of idioms:

> Analysing phrases in context with corpus linguistics techniques provides access to the deeper layers of meaning which are easily missed, most importantly the semantic preferences, semantic associations, and semantic prosody. These are all abstract features of phrasal meaning, and as such can only be identified by looking at large numbers of examples, as in a corpus. These abstract elements of meaning are the ones which are always present, in spite of variation, anchoring the phrase to its complete, functional-pragmatic meaning. Varying the most visible of cotextual elements – the collocates and colligates – can indeed change the force and focus of the meaning, but does not change the underlying message, and novelty is merely an optional extra. It is the cherry on the cake, not the cake itself (2011, p. 12).

The state-of-the-art in the study of idioms has reached a point where empirical and data-driven ideas have started to challenge the traditional lexical approaches to the nature of idioms and idiomaticity. Thus, lexical definitions of an idiom as an overarching term for semantic (or pure) idioms, semi-idioms, metaphorical idioms, similes, proverbs, sayings, hyperboles, etc., such as Crystal's characterising them as "…grammatically and lexically fixed expressions the meaning of which cannot be deduced by examining the meanings of the constituent lexemes" (2018, p. 515), are becoming increasingly rare, giving way to more specialised adaptations and interpretations. Suffice it to say, in Corpus Linguistics, idioms are termed as "an extreme form of a prefabricated unit" (Bruckmaier, 2017, p. 283). Construction Grammar describes them as "grammatical units larger than a word which are idiosyncratic in some respect" (Croft & Cruse, 2004, p. 230). In Cognitive Linguistics, however, idioms are posited as "complex symbols with specific formal, semantic, pragmatic and sociolinguistic characteristics" (Langlotz, 2006, p. 3).

It is in this connection, and as inferred from the existing body of research, several foremost theoretical strands can be identified in the field of phraseology today: Cognitive Linguistics, Psycholinguistics, Construction Grammar, Computational Phraseology, and Corpus Linguistics (Hoffmann et al., 2016; Corpas Pastor, 2022). Particularly relevant to this paper is the Corpus Linguistic approach, which argues for "a) empiricism; b) analysis of a large and principled collection of natural texts; c) extensive automatic and interactive computer-based analysis; and d) quantitative and qualitative analytical techniques" (Biber et al., 1997).

The current study aims to contribute to the developing body of work by looking at the potential in linguistic and sociolinguistic variation in the use of the idioms denoting competition, across two national varieties of English, BrE and AmE. The hypothesis of the study consists in the following. Assuming that competition is country- and, hence, culture-specific, this paper investigates whether there are any differences across the two varieties in the prominence of specific variables, such as frequency, register, gender, and age and whether these differences can be linked to the frequency of use of the idioms denoting competition.

1. Related work

This section discusses the latest relevant developments on the research topic.

It should be noted that the majority of related studies focus on idiom variation, discursive and evaluative functions of idioms, transparency and modifications of idioms, patterns of idiom usage, the syntactic flexibility of idioms, the presentation of idioms according to a categorised thematic index, and the gender-specific lexical-pragmatic meaning of idioms under the umbrella of a corpus(-based) approach. The issue of idiom frequency in corpora underpins most of the research works discussed below. The questions of register distribution of idioms and gender differences in idiom use were attended to only in a few studies. The present review of related work aims to bridge this research gap and demonstrate the need to employ balanced corpora in tandem in the idiom investigation. In that sense, my study is complementary to existing research in this area. When combined with a frequency-based approach, idiom distribution in corpora genres and subgenres and idiom user features from the perspective of sociolinguistics can be studied more rigorously. This positions my study as a springboard for unlocking the use of idioms in the forthcoming cross-corpora inter-variety research.

In Moon's study (1998), a major corpus-based analysis of 6776 most common fixed expressions and idioms in BrE and AmE, idioms were explored in the Oxford Hector Pilot Corpus (OHPC). In the broadest sense, the findings reported on general frequencies and distributions; important insights were also yielded on variation, lexical and grammatical form, polysemy and metaphor, functions of idioms in discourse, evaluative and interactional aspects of idioms, and cohesion of idioms. In a narrower sense, as far as frequencies of idioms are concerned, the study proposes frequency bands for idiom occurrence in corpora, where less than 0.25 tokens amount to insignificant frequencies (below the significance threshold), 1-2 tokens per million words correspond to low frequencies, 2 to 50 tokens per million words correlate with medium frequencies, and 50 to 100 tokens per million words relate to high frequencies. Additionally, as regards the genre distribution of idioms, the study concludes that idioms are more prevalent in journalism, accounting for 71% of all tokens, followed by fiction and non-fiction with12% each.

Gustawsson (2006) proposes a novel model of transparency of idioms to provide an explanation as to how the modification of idioms works, based on the inquiry into semantic, lexical, and grammatical features of 300 BrE verbal idioms in the British National

Corpus (BNC). The author found that idiom modifications range from semantically and syntactically simple, such as ellipses, substitutions, modifications, derivations, and passives, to semantically and syntactically complex permutations. In her research, she analyzed the frequencies of the idioms and found that, on average, they yielded 25 matches in both their canonical and non-canonical forms in the BNC. However, unlike in Moon, the frequency labels in her study were downgraded. For example, idioms with frequencies ranging from 0.01 to 0.05 per million words were considered infrequent or low-frequent. Ca. 0.1 occurrences per million words correlated with not very frequent idioms. 0.2-0.5 times per million words were described as fairly frequent idioms. 0.5-1 tokens per million words corresponded to frequent idioms. Finally, idioms that occurred more than 1 times per million words were considered frequent.

Grant (2005) describes a corpus-based frequency approach to idioms in the BNC where she compiled a comprehensive list of idioms against two core idiom criteria, such as non-compositionality and non-figurativeness. By applying the idiom rigorousness test to the idioms included in the Cambridge International Dictionary of Idioms and Collins COBUILD Dictionary of Idioms, among others, she put together a list of 103 "core idioms" that were tested for frequency in the BNC. She found that none of the idioms occurred frequently enough, meaning they occurred less than 19 times per million words in the corpus and therefore could not be included in the list of 5000 most frequent words in English. In contrast, my approach is semantic and does not involve applying the criteria of rigorousness to idioms.

Minugh (2014) provides an additional perspective on the distribution of idioms evidenced from Coll corpus, a 3.7-million-word online corpus of university student newspapers, the BNC, and several British and American English newspaper corpora, such as the Los Angeles Times, Broadcast News, the New York Times, the Independent, and the Time Magazine corpus. In terms of distribution, it was found that, surprisingly enough, the idiom density in the Coll corpus ranged from 1.2 to 55.7 per 10,000 words. In larger corpora, such as the BNC and the Time magazine corpora, idiom frequencies are much lower and match those reported in Moon (1998), for example.

Schröder (2015), in an attempt to prove the validity of the thematic classification of idioms proposed by Horn, conducts a corpus study of base and variational forms of nine verb phrase idioms, such as *kick the bucket*, *break the ice*, and *keep tabs on*, in the BNC and COCA. Like Grant, the frequencies of the idioms were found to be very low. More specifically, only one idiom (*keep tabs on*) yielded an occurrence of 1 time per million words. Furthermore, based on statistical data from the two corpora, she concludes that the thematic classification of idioms into fixed idioms, mobile idioms, and metaphors is insufficient for predicting the syntactic behaviour of idioms.

Rafatbakhsh and Ahmadi (2019) conducted a search for 1506 idioms based on 81 semantic categories at the end of the Oxford Dictionary of Idioms to establish their frequencies in the COCA. Their method was rooted in the premise that a frequency-based thematic catalogue of most used English idioms will provide more benefits in TEFL compared to the traditional intuition-based idiom selection and teaching. Out of the 1506

idioms they examined, only 17 showed a frequency of 1 or more per million words, while 234 idioms (around 15%) were not found in the corpus. Similarly, my approach involves examining the frequencies of idioms related to a specific semantic category. In addition, in my research, I will not only conduct searches in the COCA but also utilize the BNC.

In contrast, the bottom-up survey conducted by Moze and Mohamed (2019) explores patterns of idiom use based on author-assigned demographic features, such as gender, age, profession, and education, otherwise termed "sociolexical "profiling". The authors advance this approach by investigating a set of English idioms retrieved from the Pattern Dictionary of English Verbs to account for any statistically significant differences in the way men and women use idioms in everyday communication. Their findings revealed significant differences in the way speakers of different genders use idioms.

Nevertheless, none of the existing approaches provides a comprehensive solution to the problem that is stated in this paper. This study aims to address this very problem. My own approach does not include the description of pragmatic or evaluative functions of idioms, which is thus beyond the scope of the current study.

2. Methods, Material and Corpora

This section focuses on describing the methods and the methodology, the material and the corpora employed in the present study.

In a *corpus-based* approach, a researcher relies on a set of idioms or phraseological units that are perceptually salient or theoretically relevant and are deductively selected to explore how they are actually used (Tognini-Bonelli, 2001, pp. 64–81; Gray & Biber, 2015, p. 126). The corpus is thus employed to either confirm or challenge a particular hypothesis. This approach to how corpus data is used in linguistics is underpinned by the definition of corpus linguistics as a method. Conversely, *corpus-driven* studies seek to inductively retrieve idiomatic expressions from a corpus and describe them with respect to corpus evidence solely, i.e., tables are turned in favour of corpus, and not corpus linguistics, as a theory of language, whereby theoretical knowledge about the structure and nature of idiomatic units give way to their intuitive investigation in a corpus, leading to the emergence of various hypotheses during the analysis (McEnery & Hardie, 2011, p. 6; Meyer, 2014, p. 14; Barth & Schnell, 2022, p. 126).

In the present paper, a "top-"down" (OKeeffe et al., 2007) corpus-based approach is primarily employed to establish the occurrences of the idioms in terms of their raw and absolute values. Furthermore, it is used to exemplify and juxtapose the register distribution of the idioms in the two corpora. Finally, it is applied to shed light on select sociolinguistic features of the idioms under study and in so doing to verify my hypothesis.

The material used in this survey consists of 2282 tokens (base form and variations) of 11 English idioms denoting competition selected from the Oxford Dictionary of Idioms (ODI), Cambridge Idioms Dictionary (CID), and Collins Cobuild Idioms Dictionary (CCID):

- *in pole position* (in a very strong position in a competition or competitive situation, and likely to win or be successful) (ODI);

- *ace in the hole* (an advantage which you have over an opponent or rival, and which you can use if necessary) (ODI);
- *steal a march on* (to gain an unexpected or surreptitious advantage over someone or something, as by accomplishing something before, or better than, someone else) (CID);
- *hit sb below the belt* (to target one's weakness unfairly or not in keeping with the rules) (CID);
- *throw your hat in / into the ring* (to announce that one is going to be competing with others, especially in a political election) (CCID);
- *go in for the kill* (to prepare to defeat someone in an argument or competition when that person is already in a weak position) (CCID);
- *keep up with the Joneses* (striving to achieve or own as much as the people around you) (CID);
- *carry the day* (to gain victory or be successful in a contest such as a battle, debate, or sporting competition) (ODI);
- *hit the mark* (to achieve one's aim; be successful in one's attempt) (ODI);
- *sweep the board* (to win nearly everything that it is possible to win) (CCID);
- *have / gain the upper hand* (to have more power in a competitive situation than the other side and to be able to control things) (ODI).

The idioms pertaining to this specific semantic category were chosen because of my initial assumption that the assessment of humans' competitiveness in various aspects of life is an important and frequently discussed topic in discourse. For this reason, my expectation was that the above idioms would constitute relatively frequent events in discourse, given that competition is claimed to be an important medium of manifestation of social interaction (Bardis, 1979), and thereby providing a representative sample for the corpora under investigation. All 11 idioms from the pre-selected list were present in the corpora. The minimum idiom frequency in the corpora was set at 5 to ensure that the results of the analysis were statistically significant and reliable.

Two different corpora of "General English" were used, the British National Corpus (BNC), containing over 100 million words, and the Corpus of Contemporary American English (COCA), which consists of about 560 million words. These specific corpora were selected due to their status as large, well-balanced corpora of contemporary English that are publicly available online. Besides the size, the corpora vary in genre balance. Thus, the BNC is a 90% written / 10% spoken corpus, whereas genres such as fiction, newspaper, spoken, academic, and popular magazines, are evenly represented in the COCA. Furthermore, the corpora are different with respect to their modernity. For example, the BNC was released in the 1990s and was last updated in 2014. The COCA, on the other hand, has been regularly updated since the 1990s until as recently as 2019.

The obtained information was manually verified to eliminate ambiguity and ensure the reliability and relevance of the results for analysis. Queries were run online using the search interfaces of the COCA and BNC that are available at www.english-corpora.org. Both canonical forms of idioms, i.e., their base dictionary forms, and their variations were

searched for. Efforts were made to consider all possible morphological word forms of the selected idioms if applicable. For example, in the case of the idiom *throw one's hat into the ring*, variation in the categories of verb, such as *threw one's hat into the ring*, pronoun, such as *throw your hat into the ring*, preposition *throw one's hat in the ring,* and noun number, such as *threw their hats into the ring*, were all taken into account.

Register variables were determined based on the BNC user reference guide (v. 1.1) and the actual distribution of idiom tokens in the corpora. Furthermore, they were categorised into *discourse type, text type, text subtype, text domain (written), text domain (context-governed, spoken)*, and *interaction type*. To ensure uniformity, the data was standardised across the two corpora by excluding the hits classified as Movie, Blog, Web, and TV in the COCA. The sociolinguistic variables were operationalised based on the author's / speaker's *age, gender,* and *profession* both in written and spoken discourse.

3. Results and discussion

In this section, the results will be discussed against three key parameters: the frequency of the idioms in the corpora, the register distribution of the idioms in the corpora, and the sociolinguistic features of the idioms in the corpora.

Table 1 presents the occurrences of the eleven idioms in the two corpora, first in their absolute numbers followed by their relative frequencies. The occurrence data was normalised to per million words in each corpus. Regarding the actual frequency of the idioms under analysis, the idiomatic expressions had a density ranging from 5 to 30 in the BNC and from 8 to 270 in the COCA (excluding the idiomatic expression *have / gain the upper hand*). Including *have / gain the upper hand*, the actual idiom density range extends from 5 to 122 in the BNC and from 8 to 1129 in the COCA. In terms of occurrences per million words, this corresponds to a range of 0.05-0.30 in the BNC and 0.01-0.67 in the COCA (excluding "have / gain the upper hand"). Again, were we to account for *have / gain the upper hand*, the per-million density range would have been from 0.05 to 1.22 in the BNC and from 0.01 to 2.82 in the COCA. Thus, it can be safely assumed that the numbers obtained for this particular idiom rather confirm the regularity in that, if speaking generally, the idioms under study are relatively infrequent in discourse.

Based on the above, different frequency bands of idioms in each corpus were obtained. In the BNC, for example, 30% of the idioms can be classified as low frequent, 40% as not very frequent, and 30% as fairly frequent, according to Gustawsson's terminology. In contrast, 20% of the idioms can be described as infrequent, and other 20% as not very frequent in the COCA. However, the remaining 60% can be classified as "fairly frequent" and "frequent". Therefore, it can be inferred that the idioms such as *sweep the board* and *in pole position* are not typically used in American English. However, the data per million words were clearly indicative of a trend towards a higher frequency of occurrence in the COCA. This is hardly surprising given that the COCA is significantly larger than the BNC. This view is shared in the BNC vs. COCA online compare guide, which states that for low

frequency units such as idioms, there is often a real difference between a 100-million-word corpus and a 560-million-word corpus (BNC vs. COCA, 2023).

Table 1. Occurrences of the idioms in the corpora

Idiom	tokens, BNC	per million words, BNC	tokens, COCA	per million words, COCA
ace in the hole	5	0.05	131	0.29
carry the day	27	0.27	270	0.67
steal a march on	30	0.30	36	0.09
throw your hat in / into the ring	10	0.10	121	0.30
go in for the kill	16	0.16	50	0.11
in pole position	5	0.05	8	0.01
keep up with the Joneses	8	0.08	101	0.22
hit the mark	23	0.23	156	0.39
hit sb below the belt	19	0.19	102	0.25
sweep the board have / gain the upper hand	19 122	0.19 1.22	17 1129	0.03 2.82

Table 2 represents the aggregated occurrences for the ten idioms across discourse types, text types, text subtypes, and text domains, along with interaction types in each corpus in their relative numbers, i.e., per million words. The analysis shows that the idioms are more likely to be used in written discourse at 1.21 and 1.04 tokens for the BNC and COCA, respectively. In contrast, the results for spoken discourse were relatively similar in both corpora, with 0.3 occurrences per million words in the BNC and 0.34 occurrences per million words in the COCA. This is corroborated by text types, where the "written books and periodicals" category prevails in both corpora, i.e., 1.01 for the BNC and 0.91 for the COCA. As regards text subtypes, it can be observed that the idioms have been fairly evenly distributed across "fiction", "news", and "non-academic texts" subcategories. To exemplify, 0.32 idiom tokens were registered for the "fiction" subcategory in each corpus, 0.29 and 0.32 idiom tokens were found for the "news" subcategory in the BNC and COCA, and 0.28 and 0.26 idiom tokens were ascertained for the "non-academic "texts" subcategory in the BNC and COCA, respectively. If we look at the data for "academic "prose", which is also included in this category, the numbers obtained are also relatively equal for both corpora, namely 0.07 for the BNC and 0.10 for the COCA. In the case of the "conversation" subcategory, however, the results stood somewhat in contrast to the findings above. Unlike the BNC, the COCA has demonstrated double numbers for the "conversation" subcategory, i.e., 0.9 vs. 1.9 occurrences per million words. When it comes to the "written text domain" category, a further important finding is that the results for most of the subcategories, such as "imaginative", "natural and pure science", "social science", "arts", "commerce and finance", "technology", "belief and thought", "report", and "sports", have been in agreement with those stated above. In other words, the figures

recorded have been the same or narrowly even for the BNC and COCA. Conversely, noticeable differences have been observed for the "applied science", "world affairs", and "leisure" subcategories, where the tokens were either somewhat higher (0.13 vs. 0.17 for the "leisure" subcategory) or twice as high (0.02 vs. 0.011 and 0.08 vs. 0.17 in the case of the "applied science" and "world affairs" subcategories) in the BNC and COCA, respectively. Furthermore, considerable cross-corpora disparities were observed when looking at the "educational and informative" and "public and institutional" subcategories within the "spoken context-governed" text domain, i.e., 0.02 vs. 0.27 and 0.2 vs. 0.02 tokens per million words, respectively. An interesting finding pertains to the "spoken "interaction" category. On the one hand, the idioms were about twice more likely to be used in the spoken monologues in the BNC as opposed in the COCA. Yet, tables turn when it comes to the spoken dialogues. Thus, the idioms under study were about twice more likely to be used in the COCA rather than in the BNC. Broadly speaking, however, the corpus-based register variation of the idioms has proved insignificant in terms of relative values.

Table 2. Register distribution of the idioms in the corpora

No.	Category	Subcategory	BNC	COCA
			per million words	
	Discourse type	Written	1.21	1.04
		Spoken	0.30	0.34
	Text type	Spoken, demographic	0.05	
		Spoken, context-governed	0.26	0.34
		Written books and periodicals	1.01	0.91
		Written-to-be-spoken	0.05	0.002
		Written, miscellaneous	0.16	0.12
	Text subtype	Academic prose	0.07	0.10
		Conversation	0.09	0.19
		Fiction	0.32	0.32
		News	0.29	0.32
		Non-academic texts	0.28	0.26
		Other spoken texts	0.20	0.11
		Other publications	0.26	0.03
	Text domain (written)	Imaginative	0.32	0.32
		Natural and pure science	0.02	0.01
		Applied science	0.02	0.01
		Social science	0.11	0.08
		World affairs	0.08	0.17
		Commerce and finance	0.15	0.13
		Technology	0.04	0.04
		Arts	0.09	0.10

No.	Category	Subcategory	BNC	COCA
			per million words	
		Belief and thought	0.02	0.03
		Report	0.09	0.09
		Sports	0.12	0.11
		Leisure	0.13	0.17
	Text domain (context-governed, spoken)	Educational and informative	0.02	0.27
		Business	0.08	
		Public or institutional	0.20	0.02
		Leisure		0.03
	Interaction type (spoken)	Monologue	0.21	0.11
		Dialogue	0.11	0.23

Table 3 presents the sociolinguistic features of the idioms, including the author's (speakers) gender, age, and profession. Regarding gender, the idioms were more frequently used by male authors with 1.2 tokens per million words, while female authors had 0.77 tokens per million words in the written discourse in the BNC. In contrast, male authors had slightly lower usage at 0.91 tokens in the written component of the COCA, while female authors had 0.77 tokens, which is equivalent to the BNC. Thus, the data presented points to a gender gap with regard to the usage of the idioms in favour of male authors compared to female authors, which is more pronounced in the BNC than the COCA. In terms of spoken discourse, the idioms were predominantly used by male speakers at 0.39 tokens rather than by female speakers at 0.16 tokens per million words in the BNC. Similarly to the findings for the written discourse, the numbers arrived at for male speakers in the COCA were again somewhat lower than in the BNC at 0.34 tokens per million words. Likewise, evidenced from the numbers for male and female speakers in both corpora, it can be argued that the gender disparity in the use of idioms holds for both varieties of English. Yet, the margin is wider in the BNC as opposed to the COCA. Regarding age groups, it is worth noting that either insignificant or mostly no data was obtained for the 0-14 and 15-24 age categories in both written and spoken discourse of the BNC and the COCA. At the same time, the tokens generated in the written discourse for the 25-34, 35-44, 45-59, and 60+ age categories were consistently higher in the BNC in contrast to the COCA. This, however, does not hold for the speaking discourse data given that the use of idioms is remarkably similar for the age categories in question. It is also noteworthy that the idioms under study are more frequently used by authors and speakers falling into the 35-44 and 45-59 age categories, as represented by the relative values in Table 3. Additionally, as can be seen from the table, the idioms under study were most commonly used by fiction writers, reporters, non-fiction writers, and broadcasters.

Table 3. Sociolinguistic features of the idioms in the corpora

No.	Category		BNC	COCA
			per million words	
	Sex of author (written)	Male	1.20	0.91
		Female	0.77	0.77
	Author age-group	0-14		
		15-24	0.02	
		25-34	0.27	0.33
		35-44	0.65	0.61
		45-59	0.73	0.59
		60+	0.30	0.16
	Sex of speaker (spoken)	Male	0.39	0.34
		Female	0.16	0.21
	Speaker age-group	0-14		
		15-24	0.28	0.26
		25-34	0.20	0.11
		35-44	0.26	0.03
		45-59	0.25	0.22
		60+	0.06	0.10
	Profession	Academic	0.07	0.10
		Fiction writer	0.32	0.32
		Reporter	0.29	0.32
		Non-fiction writer	0.28	0.26
		Broadcaster	0.26	0.34
		Politician	0.08	0.17
		Company executive	0.08	

Conclusion

This study has examined the linguistic and sociolinguistic variation in the use of the idioms denoting competition in British English and American English, as evidenced from the BNC and COCA, respectively. Naturally, many questions remain unanswered. As can be evidenced, the idioms are not very frequent in either corpus. Assuming that a high frequency of use correlates with their importance to the lexicographers, and hence language users, this finding is rather interesting, to say the least. Among these idioms are the ones findable in the most authoritative dictionaries of idioms, such as the Oxford Dictionary of Idioms, Cambridge Idioms Dictionary, and Collins Cobuild Idioms Dictionary, to name some, but, apparently, speakers opt to not use these idioms very often. It therefore begs the age-old question of the extent to which the idioms recorded in the dictionaries

of idioms and discussed by linguists are actually important to language users, given the discrepancy between the idioms prioritised and the idioms actually used.

In the final analysis, three important conclusions need to be drawn. Firstly, depending on the corpus, 30 to 60% of the idioms in the present study were found to be in the "fairly frequent" or "frequent" group. Secondly, slight differences emerged when comparing the register distribution of the idioms in the BNC and COCA. These differences were mainly manifest in the "text domain (context-governed, spoken)" and "interaction type (spoken)", which can be attributed to the variations in text composition between the two corpora. However, it is surprising that the search for the idioms under the categories of "written "discourse" and "spoken "discourse" has yielded nearly identical results, with similar numbers for the respective subcategories. This finding is consistent with the conclusion reached by Haagsma et al., who maintain that technical and instructional language tends to use more literal idioms, while argumentative and expressive language dealing with abstract topics tends to feature more figurative idioms (Haagsma et al., 2020, p. 285-286). Thirdly, it was found that male and female authors / speakers do tend to use the idioms somewhat differently. The idioms are more likely to be used by men than by women, as suggested by both corpora. However, neither corpus provides information on how these idioms are used by speakers aged 0-14, i.e., children and adolescents.

Despite the limitations of the study, such as the compatibility of the BNC and COCA, the statistical data can serve as a reference for theorising on the actual use of idioms in contemporary English.

Consequently, further research on how idioms are represented and used in English language corpora is positively encouraged. It is likely that corpus linguistics can reveal more insights into the study of idioms than what initially meets the eye.

Sources

BNC vs. COCA, 2023. *The Corpus of Contemporary American English and the British National Corpus*. Available at: <https://www.english-corpora.org/coca/compare-bnc.asp> [Accessed 1 March 2023].

CID, 2006. *Cambridge Idioms Dictionary*. 2nd ed. New York and Cambridge: Cambridge University Press.

CCID, 2020. *Collins Cobuild Idioms Dictionary*. 4th ed. Glasgow: HarperCollins Publishers.

ODI, 2020. *Oxford Dictionary of Idioms*. 4th ed. Oxford: Oxford University Press. https://doi: 10.1093/acref/9780198845621.001.0001.

The British National Corpus. Available at: <https://www.english-corpora.org/bnc/> [Accessed 19 March 2023].

The Corpus of Contemporary American English. Available at: <https://www.english-corpora.org/coca/> [Accessed 19 March 2023].

References

Bardis, P., 1979. Social Interaction and Social Processes. *Social Science*, 54 (3), pp. 147–167. Available at: <https://www.jstor.org/stable/41886414> [Accessed 4 February 2023].

Barth, D., Schnell, S., 2022. *Understanding Corpus Linguistics*. New York: Routledge. https://doi:10.4324/9780429269035.

Biber, D., Conrad, S., Reppen, R., 1997. *Corpus Linguistics: Investigating Language Structure and Use*. New York and Cambridge: Cambridge University Press. https://doi:10.1017/CBO9780511804489.

Bruckmaier, E., 2017. *Getting at GET in World Englishes: A Corpus-Based Semasiological-Syntactic Analysis*. Berlin: De Gruyter Mouton. https://doi:10.1515/9783110497311.

Corpas Pastor, G., 2022. *You are driving me up the wall!* A corpus-based study of a special class of resultative constructions. *Lexis: Journal in English Lexicology*, 19, pp. 1–35. https://doi:10.4000/lexis.6343. Available at: <https://journals.openedition.org/lexis/6343> [Accessed 4 February 2023].

Croft, W., Cruse, A., 2004. *Cognitive Linguistics*. New York and Cambridge: Cambridge University Press. https://doi:10.1017/CBO9780511803864.

Crystal, D., 2018. *The Cambridge Encyclopedia of the English Language*. *3rd ed*. New York and Cambridge: Cambridge University Press. https://doi:10.1017/9781108528931.

Divjak, D., Levshina, N., Klavan, J., 2016. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics*, 27 (4), pp. 447–463. Available at: https://doi:10.1515/cog-2016-0095. <https://www.degruyter.com/document/doi/10.1515/cog-2016-0095/html> [Accessed 4 February 2023].

Geeraerts, D., 2018. *Ten Lectures on Cognitive Sociolinguistics*. E-Book (PDF). Leiden: Brill. https://doi:10.1163/9789004336841_002.

Grant, L., 2005. Frequency of core idioms in the British National Corpus (BNC). *International Journal of Corpus Linguistics*, 10 (4), pp. 429–451. https://doi:10.1075/ijcl.10.4.03gra.

Gray, B., Biber, D., 2015. Phraseology. In: *The Cambridge handbook of English corpus linguistics*. Eds. D. Biber, R. Reppen. New York and Cambridge: Cambridge University Press. https://doi:10.1017/CBO9781139764377.

Gustawsson, E., 2006. *Idioms Unlimited. A study of non-canonical forms of English verbal idioms in the British National Corpus.* Gothenburg: University of Gothenburg Press.

Haagsma, H., Bos, J., Nissim, M., 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 279–287. Available at: < https://aclanthology.org/2020.lrec-1.35.pdf> [Accessed 4 February 2023].

Hoffmann, S., Fischer-Starcke, B., Sand, A., 2016. Introduction. *Current Issues in Phraseology*, pp. 1–6. https://doi:10.1075/bct.74.

Langlotz, A., 2006. *Idiomatic Creativity. A Cognitive Linguistic Model of Idiom-Representation and Idiom-Variation in English*. Amsterdam and Philadelphia: John Benjamins Publishing Company. https://doi:10.1075/hcp.17.

McEnery, T., Hardie, A. 2011. *Corpus Linguistics: Method, Theory and Practice*. New York and Cambridge: Cambridge University Press.

Meyer, Ch. F., 2014. *Development in English: Expanding Electronic Evidence*. Eds. I. Taavitsainen, M. Kytö, C. Claridge, J. Smith. New York and Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139833882.004.

Minugh, D., 2014. *Studies in Corpora and Idioms: Getting the cat out of the bag*. Stockholm: Stockholm University Press.

Moon, R., 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon Press.

Moze, S., Mohamed, E., 2019. Profiling idioms: a sociolexical approach to the study of phraseological patterns. In: Computational and Corpus-Based Phraseology: Proceedings of the Third International Europhras Conference. Ed. R. Mitkov. New York: Springer Cham. https://doi: 10.1007/978-3-030-30135-4_23. Available at: <https://wlv.openrepository.com/bitstream/handle/2436/622574/submission.pdf?sequence=2&isAllowed=y> [Accessed 4 February 2023].

Murar, I., 2009. Pragmatic and functional uses of idioms. *Analele Universităţii din Craiova. Seria Ştiinţe Filologice. Lingvistică*, 1 (2), pp. 146–156.

O'Keeffe, A., McCarthy, M., Carter, R., 2007. *From corpus to classroom: Language use and language teaching*. New York and Cambridge: Cambridge University Press. https://doi:10.1017/CBO9780511497650.

Philip, G., 2011. *Colouring Meaning: Collocation and connotation in figurative language*. Amsterdam and Philadelphia: John Benjamins Publishing Company. https://doi:10.1075/scl.43.

Rafatbakhsh, E., Ahmadi, A., 2019. A thematic corpus-based study of idioms in the Corpus of Contemporary American English. *Asian-Pacific Journal of Second and Foreign Language Education*, 4 (11), pp. 1–21. https://doi:10.1186/s40862-019-0076-4. Available at: <https://sfleducation.springeropen.com/articles/10.1186/s40862-019-0076-4> [Accessed 4 February 2023].

Schönefeld, D., 2022. Framing in American and British Governmental Discourse about COVID-19. In: *Cognitive Sociolinguistics Revisited*. Eds. G. Kristiansen, K. Franco, S. De Pascale, L. Rosseel, W. Zhang. Berlin: Walter de Gruyter. https://doi:10.1515/9783110733945.

Schröder, D., 2015. Spilling some linguistic beans: On the syntactic flexibility of idioms. In: *Within Language, Beyond Theories. Discourse Analysis, Pragmatics and Corpus-based Studies. Vol. 3*. Eds. M. Malec, M. Rusinek. Newcastle upon Tyne: Cambridge Scholars Publishing. https://doi:10.13140/RG.2.1.2955.3689.

Tognini-Bonelli, E., 2001. *Corpus linguistics at work.* Amsterdam and Philadelphia: John Benjamins Publishing Company. https://doi:10.1075/scl.6.