Journal of Research in Applied Linguistics
ISSN: 2345-3303 – E-ISSN: 2588-3887 – http://rals.scu.ac.ir

Please cite this paper as follows:

Allami, H., Karlsson, M., & Shahroosvand, H. R. (2022). Conventional and nonconventional use of idioms in general vs. academic corpora of English as a lingua franca. *Journal of Research in Applied Linguistics*, *13*(1), 44-57. https://doi.org/10.22055/RALS.2022.17424

Shahid Chamran University of Ahvaz

Research Paper

Conventional and Nonconventional Use of Idioms in General vs. Academic Corpora of English as a Lingua Franca

Hamid Allami[1], Monica Karlsson[2], & Hamid Reza Shahroosvand[3]

[1]Corresponding author; English Department, Faculty of Humanities & Social Sciences, Yazd University, Yazd, Iran; *hamid_allami@yahoo.com*

[2]English Department, Halmstad University, Sweden; *monica.karlsson@hh.se*

[3]English Department, Yazd University, Yazd, Iran; *hamid.shahroosvand@gmail.com*

Received: 28/09/2020 Accepted: 08/11/2021

Abstract

The present study investigated the conventional vs. nonconventional use of idioms in general and academic English as a lingua franca (ELF) corpora taking into account the speech event type, academic domain, and discipline. ELFA and VOICE corpora were searched for idiom tokens based on *Collins COBUILD Idioms Dictionary*. Results showed that idioms were more frequent in VOICE than in ELFA, indicating a higher proportion of formulaic language in informal and interactive discourse as compared to more formal and transactional discourses. Tokens in conventional form and meaning were the most frequent in both corpora. Entirely novel idioms were small in number in both corpora. However, both corpora generated a large number of idioms with formal variations. Idiom use in the academic corpus was register sensitive. ELF speakers in both corpora used communication strategies to prevent unilateral idiomaticity. Overuse of high-frequency idioms by some speakers could be associated with idiomatic teddy bears. Results can help understand the nature of idiomaticity in ELF in general and academic settings. Findings on the academic corpus can also inform curriculum development and assessment in English for Academic Purposes.

Keywords: Idiom; ELF Corpus; Variation; Academic Domain

1. Introduction

The term English as a lingua franca (ELF) refers to "communication in English between speakers with different first languages," and thus, "most ELF interactions take place among 'non-native' speakers of English" (Seidhlofer, 2005, p. 339). ELF has some deviations from native speaker (NS) norms in various areas including lexicogrammar. Such lexicogrammatical anomalies are used in ELF without compromising meaning (Seidlhofer, 2001) because the purpose is to convey meaning and achieve communication rather than to conform to NS norms of accuracy. However, such deviations from native norms may, in some cases, lead to miscomprehension. One such area in ELF is the use of formulaic language and idioms.

Idiomaticity plays a significant role in effective use of a language by nonnative speakers (NNS), so that lexicogrammatical phrases form "the main building blocks of fluent connected speech" (Pawley & Syder, 1983, p. 214). This has been confirmed by the high presence of idiomaticity in corpus research (Prodromou, 2003). However, due to semantic and pragmatic difficulties inherent in idioms, English NNSs, even very advanced learners and users of English, often face problems in using and comprehending idioms (Prodromou, 2007). This is more serious in ELF contexts where the interlocutors are from different L1 backgrounds and are more likely to have a low pragmatic competence. This can lead to breakdown in communication or unilateral idiomaticity, where "idiomatic speech by one participant can be problematic when the expressions used are not known to the interlocutor(s)" (Seidlhofer, 2004, p. 220).

One issue that may lead to problems in use of idioms in ELF is the use of nonconventional idioms (i.e., idioms not found in native English; Pitzl, 2012, 2016; Prodromou, 2007; Seidlhofer, 2009), which result from either the literal

translation of an idiom into English or the creative construction of an idiom that is not conventional in English. Vetchinnikova (2014) states that because ELF users do not share knowledge of English as a native language (ENL) phrasology they actively go through a process of idiomatizing or active creation of idioms. Such creation may range from creating entirely novel idioms to idioms that have undergone some types of variation including formal and semantic variation with or without affecting communication of their meanings.

Recently, the need for studies that focus on how English idioms are used by people of different L1 backgrounds in their authentic daily use of the language (i.e., in ELF settings) has been acknowledged (e.g., Pitzl, 2012, 2016). However, such studies have mostly focused on either general or academic ELF, and there is scarce research comparing conventional and nonconventional use of idioms across general and academic ELF corpora. This study investigated the conventional and nonconventional use of idioms in general and academic ELF corpora (VOICE and ELFA) taking into account variables such as speech event type (conversation, workshop, doctoral defense presentation, conference discussion, seminar discussion, etc.), academic domain (behavioral sciences, economics and administration, humanities, medical sciences, natural sciences, social sciences, and technology), discipline (education, mathematics, physics, etc.; in case of the academic corpus), and variation type.

2. Literature Review

The concept of idiom is far from being clear-cut. That is perhaps the reason for various definitions suggested for an idiom in the literature. They range from inclusive definitions that consider an idiom as a multiword item that can be semantically opaque or not, to those that strictly define idioms as fixed and semantically opaque (Moon, 1998). Biber et al. (1999, p. 988) state that multiword expressions can be classified based on their 'idiomaticity' and invariability." They put idioms at one extreme and believe that idioms are relatively invariable and their meanings are not predictable from their parts. That is, an idiom is learned as a whole and knowing the constituent words does not guarantee knowing the meaning of the idiom. This definition is somehow supported in Merriam-Webster definition, which points to the semantic and/or grammatical uniqueness of idioms.

The semantic opacity mentioned above is the common point stressed in most definitions of idiom. That is, idioms contradict the principle of compositionality (also called Frege's principle), that is, the principle that the meaning of a complex expression is determined by the meanings of its constituent parts and the manner in which they are combined (Pelletier, 1994). In his idiom principle, Sinclair (1991) uses the one-choice principle to refer to noncompositionality stating that language users have a repertoire of preconstructed phrases that constitute "single choices," though they may seem to be analyzable into parts. This semantic opacity (Nunberg et al., 1994) or noncompositionality is the distinguishing feature for idioms agreed on by most researchers and is seen in most definitions. Noncompositional models like the literal first hypothesis (Bobrow & Bell, 1973), the lexical representation hypothesis (LRH; Swinney & Cutler, 1979), and the direct access hypothesis (Gibbs, 1980) posit that idioms are stored and retrieved as chunks in the lexicon and very little alteration is allowed in them. However, they have been criticized for their inability to account for the possibility of alterations in idioms, called transformational potential by Fraser (1970), leading to different compositional models.

Compositional models prefer to treat idioms as decomposable pointing to evidence showing that the relationship between form and idiomatic meaning of idioms are not always arbitrary. Some scholars thus have tried to place idioms on a compositionality continuum (Moon, 1998) or a hierarchy (Fraser, 1970). Fraser's frozenness hierarchy is a continuum of transformational potential of idioms ranging from Level 0, where no operations may apply to an idiom, that is, it permits no distortion in idioms, to Level 6 where any operations is permitted in an idiom. As is evident, Fraser's hierarchy is opposed to the noncompositional view of idioms, as it allows for some degrees of alterations in idioms without harming their figurative content. In the same line, building on the fast recognition of idioms by research subjects, Cacciari and Tabossi (1988) rejected the direct access and lexical representation hypotheses and introduced the identification of *idiom key* as the recognition point of an idiom. However, they did not present satisfactory criteria for pinpointing such a key. Another reason suggested for this fast recognition of idioms was presented in the conceptual metaphor hypothesis, whereby it is argued that conceptual metaphors facilitate the comprehension of idioms (Gibbs et al., 1997). In other words, "idioms do not need to be represented independently in the mind. They are understood via preexisting conceptual metaphor schemes" (Vega-Moreno, 2001). For example, *spill the beans* is understood through two metaphors: MIND IS A CONTAINER and IDEAS ARE PHYSICAL ENTITIES.

Compositional models have been criticized based on some empirical evidence according to which highly-familiar idioms are processed faster, that is, their comprehension does not need recourse to literal processing or conceptual metaphors. Some scholars believe that idioms, by definition, are fixed and are not subject to variations. For example, according to Eftekhari (2008), a speaker or writer is not allowed to make alterations like changing the word order, deleting or adding a word, replacing a word with another, and changing grammatical structure in an idiom, unless he or she intends to make a joke or play on words. However, this cannot be stated as a general rule because there have been cases where such alterations have not led to miscomprehension and communication breakdown as discussed below.

Some researchers have pointed to the possibility of creative idiomaticity in ELF (Pitzl, 2012, 2016; Prodromou, 2007; Seidlhofer, 2009) where L2 learners create idiomatic expressions that are nonconventional in ENL. Vetchinnikova (2014) states that because ELF users do not share knowledge of ENL phraseology they actively go through a process of idiomatizing or active creation of idioms. Seidlhofer (2009) tracks creative idiomaticity to Widdoson's cooperative and territorial imperatives. The cooperative imperative refers to fine-tuning of language by interlocutors in communication to make their intention accessible; while territorial imperative refers to interlocutors' adjustment of their language in order to enhance differences with others and reinforce their social identity in a way that is acceptable to others. A balance between these two imperatives is important for social life (Widdowson, 1983, as stated in Seidlhofer, 2009).

ELF settings where the users of English do not have a shared knowledge of ENL conventions is a rich context for creative use of idioms. In such a case, idiom use is not confined to preconstructed phrases of the ENL, but includes the innovative use of idioms actively constructed by the ELF users. That is because ELF users are not dealing with a community of users with the same L1 and are thus not concerned with being marked as an outsider as a result of their deviations from the conventional use of idioms (Seidlhofer, 2009).

Creative use of idioms can lead to problems in comprehension of creative idioms by the hearer, a phenomenon called unilateral idiomaticity (Seidlhofer, 2009). Indeed, unilateral idiomaticity can happen due to the hearer's lack of knowledge of the idiom used even if the conventional form of the idiom is used. However, assuming that the hearer's knowledge is not impaired, formal alterations in idioms, or introduction of entirely novel idioms for example via literal translation of L1 idioms can lead to unilateral idiomaticity.

Idiomatic creativity in ELF has been investigated by some researchers. Pitzel (2012) introduces the process of remetaphorization or introduction of metaphoricity into conventional idiomatic expressions. It is argued that "formal variation in ELF transforms idioms from conventional and often (seemingly) noncompositional fixed phrases to creative figurative expressions which are compositional, semantically transparent and can be interpreted as metaphors" (p. 48). She explores the general tension between creativity and convention and suggests the processes of idiomatizing and remetaphorization that indicate "an inherent tendency toward stabilization and conventionality in language use" (p. 48), instead of the idiom-metaphor dichotomy. The study by Pitzel (2012) focused on the characteristics and functions of creative idioms in Vienna-Oxford International Corpus of English (VOICE).

In another corpus study, Pitzel (2016) investigated the creative use of idioms in ELF settings. She studied some cases of novel idioms, which were transferred from or created based on ELF speaker's L1 and suggested a "shared multilingual resource pool for ELF interactions" (p. 294) as one source of such creative uses of idioms. However, her study primarily focused on transfer from L1 in the general ELF corpus (VOICE) and did not investigate or compare the results to that of academic ELF settings.

Briggs and Smith (2017) focused on idiomaticity in "English-medium instruction (EMI)" (p. 27) and ELF. They pointed to the role of EMI in preparing students for academic ELF as well as to the underpreparedness of L2-English users for ELF events, especially those involving idiomaticity that can be the result of underexposure to interactional ELF contexts. They suggested prelogical implications for improving idiomatic competence through EMI. However, their study lacked collection and analysis of data and was based only on a review of literature on the issue.

Bostanci (2017) investigated formulaic sequences in ELF corpora in Asian and European contexts and found that European ELF had a higher degree of formulaicity than Asian ELF. In terms of interaction types, social ELF interactions were found to be more formulaic compared to academic interactions in both contexts. Regarding the categorization of formulaic language, the results showed higher frequencies for "speech formulas and fixed and semifixed semantic units, whereas "situation-bound utterances and idioms" had the least frequently in both ELF contexts. Regarding

creative idiomaticity, "non-standard forms of formulaic expressions" (p. iv) were slightly more frequent in Asian ELF than in European ELF. It is to be noted that his study focused on categorization of formulaic sequences in general in two corpora of VOICE and Asian Corpus of English (ACE), rather than idioms.

In a study on creative use of idioms in ELF, Pitzl (2018) grouped formal variations in idioms used in VOICE corpus based on the categorization suggested by Langlotz (2006) who classified formal variations in idioms into morphosyntactic variations (e.g., "inflectional variants" [p. 179] of idiom constituents), syntactic variations (e.g., "passivisation and clefting" [p. 180]), and lexical substitution. Though the categorization was originally proposed for native use of idioms, it can be used for creative use of idioms in general (Pitzl, 2018). Two interpretations are suggested for idiom variations: "idiom decomposition interpretation" (Langlotz, 2006, p. 138) where a conventional idiom is decomposed and restructured and "metaphorical compositionality interpretation" (p. 138) where conventional phrases or parts thereof are used to create a metaphor, supporting a noncompositional view of idioms. Pitzl (2018) focused on formal variations as applied to the results collected from VOICE; however, an investigation of academic ELF can corroborate or, otherwise, dispute the results and a focus on other types of variations can broaden the perspective on idiomatic creativity in ELF.

Khodabandeh and Ramezani (2020) compared the use of formulaic language in ELF and ENL in academic settings. Investigating a corpus of academic lectures (Michigan Corpus of Academic Spoken English), they found that ELF users applied formulaic sequences less than NSs. Prepositional phrases were the most frequent formulaic sequence in both groups. The study, though informative on formulaic sequences in general, did not explore the use of idioms in ELF.

In a study on the use of idiomatic expressions by Dutch ELF users, Linden (2020) found no significant effects for the use of idiomatic expressions on perceived comprehensibility of business e-mails for ELF readers as well as the perceived competence of writers of those e-mails as assessed via questionnaires. The results could account for only a limited subset of academic ELF (i.e., business e-mails) and focus on perceptions of ELF writers and readers of those e-mails.

The review of literature shows a dearth of empirical studies comparing patterns of using idioms and their variations across general and academic ELF corpora. A comparison of different disciplines within the academic domain can provide a rich area of research. Therefore, the research questions investigated in this study were as follows:

1. Are the proportions of idioms in VOICE and ELFA significantly different?
2. Do different academic domains generate different proportions of idioms?
3. Do different disciplines generate different proportions of idioms?
4. What are the differences between VOICE and ELFA in terms of types of idiomatic variation?
5. What are the frequent idiom variation types in VOICE and ELFA?
6. What is the relationship between dictionary frequency and corpus frequency of idioms in VOICE and ELFA?

3. Methodology

A corpus analysis method was used for the purpose of the present study. The materials included two corpora: ELFA and VOICE.

3.1. Academic ELF Corpus

To investigate academic ELF, the corpus of English as a Lingua Franca in Academic Settings (ELFA) was used (1 million transcribed words). The speakers in ELFA are about 650 speakers with 51 different L1s from various continents. The speech events in the corpus include lectures and presentations, and seminars, thesis defenses, and conference discussions. The academic domains of ELFA are social sciences, technology, humanities, natural sciences, medicine, behavioral sciences, and economics and administration (ELFA Corpus, 2019).

3.2. General ELF Corpus

To investigate general ELF, VOICE was used (1 million words). VOICE contains transcripts of natural face-to-face ELF interactions. About 1,250 experienced ELF speakers (mainly European) with about 50 different L1s were recorded in VOICE. Speech events types in the corpus include "interviews, press conferences, service encounters, seminar discussions, working group discussions, workshop discussions, meetings, panels, question-answer-sessions, conversations" (VOICE Project, 2019, para. 4).

3.3. Procedure

The idioms in *Collins COBUILD Idioms Dictionary* were searched one by one in both corpora. In order to find variants of idioms, both the entire idioms and the single keywords in them were searched in the texts of both corpora using MS word search tool. Two criteria of compositeness ("the fact that idioms are multiword units that consist of two or more lexical constituents" [Langlotz, 2006, p. 3]) and noncompositionality (the fact that "the meaning of these constructions is not the derivational sum of the meanings of their constituents" [p. 4]) were used as criteria to select novel idioms used in the ELF corpora. Therefore, those novel formulaic expressions that were decomposable and did not have figurative content were excluded from the analysis. Such idiom tokens were found by intensively reading the transcripts.

The selected idioms were grouped based on speech event type (conversation, workshop, doctoral defense presentation, conference discussion, seminar discussion, etc.), academic domain (behavioral sciences, economics and administration, humanities, medical sciences, natural sciences, social sciences, and technology), and discipline (education, mathematics, physics, etc.; in case of the academic corpus) based on labels in the corpora texts. The tokens were also labeled for frequency based on frequency data in *Collins COBUILD Idioms Dictionary*, and for variation type ranging from No Change, Entirely Novel, Formal Variation, Semantic Variation, Formal and Semantic Variation based on the conventional and nonconventional use of idioms as distinguished based on *Collins COBUILD Idioms Dictionary*. The frequency of idioms used was compared across these groups.

This study followed a mixed-methods approach in data analysis. Therefore, thick descriptions, descriptive statistics, nonparametric tests, and cross-tabulation were used to analyze the data.

4. Results

4.1. Proportions of Idiom Types and Tokens in VOICE and ELFA

In order to compare the two corpora in terms of idiom use, frequency statistics for idiom tokens in the corpora was obtained as follows:

Table 1. *Frequency of Idioms in VOICE and ELFA*

	Frequency	Percent
ELFA	280	42.2
VOICE	384	57.8
Total	664	100.0

As seen in Table 1, a total number of 664 idiom tokens were found in both corpora (384 tokens for VOICE and 280 tokens for ELFA). The percentage of idiom tokens in VOICE was higher than that of ELFA (57.8% for VOICE and 42.2% for ELFA). In order to determine whether idioms occur with equal probability in VOICE and ELFA, a one sample binomial test was used (see Figure 1). The results showed that the null hypothesis could be rejected at the significance level 0.05 (p value = 0.00), as shown in Table 2:

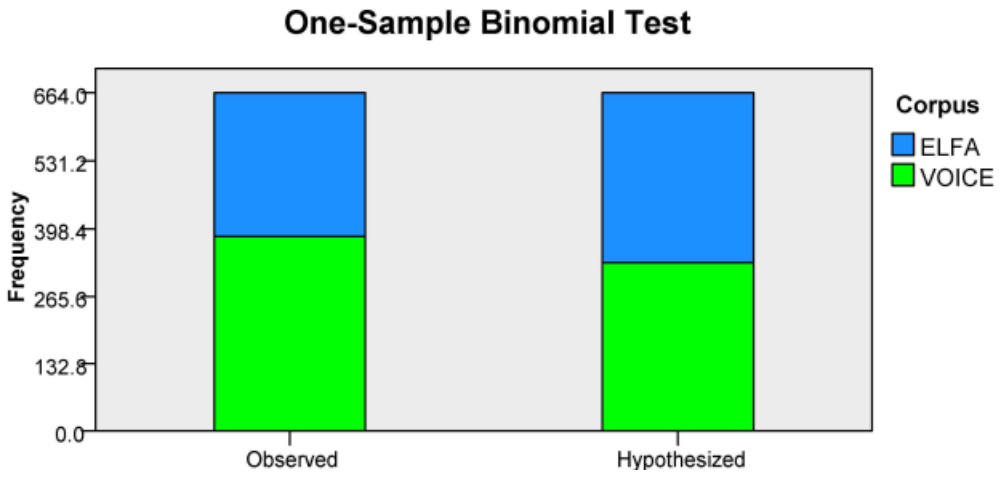

Figure 1. Equality of Proportions Hypothesis Test for Proportions of Idioms in VOICE and ELFA

Table 2. *One Sample Binomial Test Result*

Null Hypothesis	Test	*Sig.*	Decision
The categories defined by corpus (ELFA and VOICE) occur with probabilities 0.5 and 0.5.	One Sample Binomial Test	.000	Reject the Null Hypothesis

Therefore, there was no likelihood that the variation between idiom proportions in VOICE and ELFA was explained by chance, making it safe to reject the null hypothesis that equal proportions of idioms would occur in VOICE and in ELFA. That is, the proportions differ significantly from the expected equal proportions of idioms in the corpora, with VOICE generating a higher proportion of idiom tokens than ELFA.

A total number of 270 idiom types were found in both corpora, excluding entirely novel ones. The frequency of idioms in both corpora showed that a large proportion of idioms had a very low incidence in the corpora (183 idioms (67.5% of the total number of idiom types) were repeated only once and 35 idioms (13% of the total number of idiom types) were repeated only twice in both corpora (2,300,000 words). The idiom types with the highest number of tokens are listed in Table 3:

Table 3. *Most Common Idiom Types Found in Both Corpora*

Idiom Type	Frequency	Percent
at the end of the day	62	9.3
the grass roots	31	4.7
bear something in mind; keep something in mind	21	3.2
N/A (entirely novel)	17	2.6
a question mark	14	2.1
hand in hand	13	2.0
a bad apple; a rotten apple; a bad apple spoils the barrel	12	1.8
from scratch	12	1.8
in hand; take someone in hand	11	1.7
build bridges	10	1.5

These top 10 frequent idiom types were those labeled as highly frequent in *Collins COBUILD Idioms Dictionary* with a couple of exceptions. However, the most frequent idiom type (i.e., *at the end of the day*) was labeled in the dictionary as low frequent. Examining the tokens shows the overuse of the said idiom by some speakers. For example, a single speaker used the idiom 17 times and another one used it 10 times, both in VOICE corpus. This can be associated with linguistic teddy bears, that is, speakers' overuse of items they feel more comfortable with (Karlsson, 2019). Table 4 shows the idiomatic teddy bears found in the corpora:

Table 4. *Idiomatic Teddy Bears in Corpora*

Idiomatic Teddy Bear	Event Type	Corpus	No. of Times Repeated by One Speaker
at the end of the day	business meeting about coordinating a PR campaign across several countries	VOICE	17
at the end of the day	business meeting at a forwarding agency with a sales representative of an airline	VOICE	10
grass roots	doctoral defense discussion	ELFA	9
bear/keep something in mind	lecture discussion	ELFA	9
rotten apple	doctoral defense discussion	ELFA	5

4.2. Frequency of Idioms Across Different Event Types

'Seminar discussion' generated the highest number of idioms (95 tokens equal to 14.3% of the total number of tokens found in both corpora) followed by 'doctoral defense discussion' (92 tokens equal to 13.9% of the total number of tokens), 'conference discussion/presentation' (55 tokens equal to 8.3% of the total number of tokens), 'lecture' (30 tokens equal to 4.5% of the total number of tokens), and 'business meeting about coordinating a PR campaign across several countries' (24 tokens equal to 3.6% of the total number of tokens).

Although, in total, the general corpus had a significantly higher number of tokens (384 tokens in VOICE vs. 280 tokens in ELFA), the event types that generated the highest number of idiom tokens (i.e., 'seminar discussion,' 'doctoral defense discussion,' 'conference discussion/presentation,' and 'lecture') were all in the academic corpus. The tokens in the general corpus are dispersed in all event types with 'business meeting about coordinating a PR campaign across several countries' (24 tokens) generating the highest number of tokens in VOICE. However, if 'workshop discussion events' are considered as one group regardless of their discussion topics, it will include the highest number of tokens (50 tokens).

4.3. Proportions of Idioms in Different Disciplines (ELFA)

A total number of 280 tokens were found in ELFA. The disciplines that generated the highest number of idiom tokens are shown in Table 5:

Table 5. *Disciplines With High Numbers of Idioms in ELFA*

Corpus	Discipline	Frequency (per million words)	Percent
ELFA	Journalism and Mass Communication	38	13.6
	Education	31	11.1
	IT	17	6.1
	Multidisciplinary	17	6.1
	Philosophy	16	5.7
	Social Policy	15	5.4
	Political History	14	5
	Information Sciences	14	5
	Industrial Engineering	13	4.6
	Forestry	13	4.6
	Cultural Studies	12	4.3

'Journalism and mass communication' generated the highest number of idioms followed by 'education.' As can be seen, disciplines with high frequency of idioms are primarily from soft sciences. Accounting, genetics, mathematics, physics, and regional studies had only one token and hematology, neurology, translation studies, virology, and economics had only 2 tokens (per million words).

4.4. Proportions of Idioms in Different Academic Domains (ELFA)

Frequency statistics for idiom tokens in academic domains are shown in Table 6:

Table 6. *Frequency of Idioms in Different Academic Domains in ELFA*

Academic Domain	Frequency (per million words)	Percent
Behavioral Sciences	30	10.7
Economics and Administration	9	3.2
Humanities	47	16.8
Medical Sciences	9	3.2
Natural Sciences	20	7.1
Social Sciences	115	41.1
Technology	50	17.9
Total	280	100.0

Table 6 shows that 'social sciences' generated the highest number of idioms (41.1% of the total number of tokens found in ELFA) followed by 'technology' (17.9%), 'humanities' (16.8%), and 'behavioral sciences' (10.7%). 'Economics and administration,' 'medical sciences,' and 'natural sciences' generated comparatively few idioms.

4.5. Comparison of VOICE and ELFA in Terms of Idiom Variation Types

The frequency of variation types in VOICE vs. ELFA are presented in Table 7:

Table 7. *Frequency of Variation Types in VOICE vs. ELFA*

Corpus	Variation Type	Frequency (per million words)	Percent
ELFA	Entirely Novel	5	1.8
	Formal and Semantic Variation	10	3.6
	No Change	191	68.2
	Only Formal Variation	71	25.4
	Only Semantic Variation	3	1.1
	Total	280	100.0
VOICE	Entirely Novel	12	3.1
	Formal and Semantic Variation	9	2.3
	No Change	282	73.4
	Only Formal Variation	75	19.5
	Only Semantic Variation	6	1.6
	Total	384	100.0

Tokens used in conventional form and meaning, labeled No Change, were the most frequent in both corpora (68.2% of tokens in ELFA and 73.4% of tokens in VOICE), but the difference between these two proportions was not significant. A higher percentage of idioms in VOICE were used in conventional form compared to ELFA. The entirely novel idioms were significantly more frequent in VOICE. However, the frequencies of other variation types were not significantly different across VOICE and ELFA. That is, both corpora showed a similar distribution of idiom tokens across variation type categories.

A large proportion of idiom tokens had only formal variation (25.5% in ELFA and 19.5% in ELFA) ranging from inflectional change such as tense of the idiom, substitution of some words in the idiom by their synonyms (e.g., *grey zone* instead of *grey area*), deletion of noncore words in the idiom (e.g., *in nutshell* instead of *in a nutshell*), insertion of a word without affecting the meaning of the idiom (*on one's big toe* instead of *on one's toes*). Only about 7% of the tokens were either entirely novel ones or underwent variations that affected their meanings. Entirely novel idioms (those multiword expressions that were not found in idiom dictionaries and could not be interpreted based the literal meanings of their constituent parts) accounted for just 2.6% of all the tokens in both corpora.

Cross-tabulation data for Corpus and Variation Type showed that of the 27 Entirely Novel tokens, 29.4% were generated in ELFA and 70.6% in VOICE. This difference is significant showing a higher creativity in VOICE compared to ELFA. Similarly, there was a significant difference between the two corpora in terms of incidence of Only Semantic Variation. However, the total number of such tokens (9 in both corpora) was not significant. In case of Formal and Semantic Variation as well as Only Formal Variation, the difference between the two corpora was not significant.

4.6. Dictionary Frequency vs. Corpus Frequency in ELFA vs. VOICE

Collins COBUILD Idioms Dictionary includes frequency labels for idioms including High (H), Medium (M), Low (L), and Very Low (VL). Table 8 shows the frequency of tokens belonging to each frequency label:

Table 8. *Frequency of Idioms in Corpora Belonging to Each Frequency Label in Collins COBUILD Idioms Dictionary (ELFA vs. VOICE)*

Corpus	Frequency Label in Dictionary	Frequency in Corpus (per million words)	Percent
ELFA	High	143	51.1
	Low	37	13.2
	Medium	48	17.1
	N/A	5	1.8
	Not Mentioned	43	15.4
	Very Low	4	1.4
	Total	280	100.0
VOICE	High	191	49.7
	Low	82	21.4
	Medium	37	9.6
	N/A	12	3.1
	Not Mentioned	53	13.8
	Very Low	9	2.3
	Total	384	100.0

The category labeled N/A includes entirely novel idioms, which are thus not found in the dictionary. The category labeled Not Mentioned comprises those idioms for which no frequency label was mentioned in the dictionary. These idioms were not in *Collins COBUILD Idioms Dictionary*, but were included in the analysis as they had the criteria of compositeness and noncompositionality and were also found in other idioms dictionaries. A comparison of VOICE and ELFA in terms of incidence count of tokens belonging to each frequency label in *Collins COBUILD Idioms Dictionary* (i.e., High, Medium, Low, and Very Low) shows that, in both corpora, the most frequent idioms are those labeled as highly frequent in the dictionary (51.1% of the total number of tokens in ELFA and 49.7% of the total number of tokens in VOICE). However, this does not hold true for idioms labeled low frequent as they ranked second (13.2% in ELFA and 21.4% in VOICE), followed by those labeled medium (17.1% in ELFA and 9.6% in VOICE). Those idioms labeled very low in the dictionary had the lowest incidence in both corpora.

In addition, the cross-tabulation of data showed that the majority of tokens of idioms labeled as high frequency in the dictionary had undergone no change in the corpora (81.1% in ELFA and 79.6% in VOICE). The reason can be that due to their high frequency, speakers felt confident in their correct form, that is, no need for change. The opposite applies to low-frequency items.

5. Discussion

The proportions of the idioms in ELFA and VOICE were not equal in that VOICE generated a significantly higher number of idioms. That is, the percentage of idioms in the informal language was quite high. This is in line with previous findings showing a higher proportion of formulaic language in informal and interactive discourse as compared to more formal and transactional discourses (Nattinger & DeCarrico, 1992; Vilkaitė, 2016) and those finding indicating that "formulaic expressions are often associated with spontaneous ways of verbal communication" (Abdou, 2011, p. 26). The academic discourse contains a higher proportion of monologic transactional discourse like lectures as compared to VOICE, which includes more interactional discourse like informal discussions among friends and students. Therefore, the higher number of idioms in VOICE lends supports to previous findings in this area. For example, previous studies comparing academic and general discourse show that academic prose has a lower proportion of formulaic discourse in comparison to general conversations (Biber et al., 1999), which is confirmed by the results of this study. The academic corpus generally includes noncollaborative discourse, where "one speaker dominates significantly supported by back-channeling from the other speaker(s)" (Carter, 2004, p. 149). Such situations do not lend themselves to coconstruction and creative idiomaticity.

A large proportion of idioms had a very low incidence in the corpora. This confirms the finding that a majority of idioms have frequencies of 1 token or fewer per million words (Moon, 1998). The top 10 frequent idiom types were those labeled as highly frequent in the dictionary with a couple of exceptions. This supports the positive effects of frequency on idiom use. Examining the tokens shows the overuse of the said idiom by some speakers. This can be associated with linguistic teddy bears, that is, the speakers' overuse of items they feel more comfortable with (Karlsson, 2019).

A phenomenon observed in the results was the overuse of specific idioms by some speakers. Hasselgren (1994) found that even advanced L2 learners tend to use high-frequent words in order to avoid the risk of making mistakes in selecting words that are more appropriate in the intended context, but are less frequent (less familiar to them). These words are called lexical teddy bears. Cognates, transliterations, and perceived equivalence are the origins of such teddy bears. For example, lexical teddy bears are made when L2 learners select "the L2 word that seems to them most able to function like the L1 'equivalent'" (p. 256). Hasselgren (1994) relates lexical teddy bears to familiarity and calls them "dependence on the familiar" and clinging to "the familiar L1 vocabulary boundaries" and imposing them on the L2. Similarly, Karlsson (2019) states that linguistic teddy bears can be a negative effect of L2 familiarity. However, she makes a distinction between frequency and familiarity, though they appear to be used interchangeably by some researchers. Strictly speaking, whereas the former term refers to an objective opinion based on statistical evidence, the latter is a subjective approach (Baayen, 1992, 1993; Baayen & Lieber ,1997). This distinction in meaning does not, however, imply that they are disconnected, as the more frequent an item is, the more familiar the item is likely to be (Abel, 2003).

Karlsson (2019) shows very clearly that learners use some idioms more than others, that is, they function as their idiomatic teddy bears or idiomatic security blankets. The results of this study confirm her findings, that is, that ELF speakers are more likely to use and distort idioms that are familiar to them (those labeled frequent in the dictionary) than others. For example, a single speaker used the idiom *at the end of the day* 17 times in a single event in VOICE and another speaker used it 10 times in a single event ('business meeting at a forwarding agency with a sales representative of an airline') in the same corpus. The same is true for some other high-frequent idioms in the corpora as presented in Table 4.

In the majority of cases, the idioms used as teddy bears did not undergo any variation and was used properly though excessively. However, in a few cases, the dependence on idiomatic teddy bears leads to some manipulations and overgeneralization in familiar idioms rather than using a less familiar but more appropriate and native-like idiom. For example, in a doctoral defense discussion in ELFA (FILE ID: UDEFD030), the speaker uses the idiom *rotten apple* or its variation, *good apple*, 5 times in a conversation of less than 200 words. In the following example, instead of using idioms like *one swallow doesn't make* a *summer*, the ELF speaker manipulates his idiomatic teddy bear and uses the variation *good apple* as a perceived equivalence to express the same meaning. However, the response by the hearer, S4, shows that the intended message was not properly conveyed:

- </S2> . . . if they say that apples are good it doesn't mean that all apples are good and </S2>
- <S4> but if you say apples are good i i understand it it means all the apples are <S2> [yeah] </S2> [good] …<S4>

Regarding dictionary and corpus frequency, as expected, the most frequent idioms in the corpora were those labeled as highly frequent in the dictionary. The fact that idioms labeled low frequent ranked second and above those labeled medium in the dictionary confirms the difference in familiarity and frequency. That is, though such idioms are not highly frequent they seem to be familiar to users and generate a higher number of tokens than those labeled medium. This lends support to distinction made between the two by Karlsson (2019) who deems familiarity a subjective concept as opposed to frequency, which is more objective and statistical.

The entirely novel idioms were significantly more frequent in VOICE. The reason can be purported to the comparatively more scientific nature of language in ELFA, which does not allow for risks of misunderstanding.

As for patterns of variation across different academic disciplines in terms of idiom use, the results showed that humanities and social sciences generated a higher number of idioms than natural sciences and medical sciences did. This can be attributed to the use of more technical speech in the latter and more rhetorical and colorful language in the former domains. That is, the use of idioms in the academic corpus was register-sensitive and can be predicted based on the discipline. This is not consistent with results found by Simpson and Mendis (2003) who call such a conclusion stereotypical and unfounded and reject it based on their results.

Economics and administration, medicine and natural sciences generated comparatively few idioms. It is natural in that these domains already have established terminology in most areas, and it would be more hazardous of course to start tampering with such expressions, within medicine even risking that doctors misunderstand, causing people to die (Karlsson, 2019). The exception is technology, which can be attributed to the great number of new inventions. The other language domains are more of a descriptive nature, lending themselves to language that may be modified, without causing too much misunderstanding.

In terms of variation types, formal variations were the most frequent variation type in both corpora, showing a tendency among ELF users for preserving the original meaning of the idiom. One of the prevalent formal changes in the idioms was that of substituting a word or phrase in the idiom. Out of the total number of tokens (n = 664), about 113 tokens included some type of substitution, including substituting a word with its synonym, a word with close meaning but not necessarily its antonym, or its antonym, substituting prepositions and grammatical particles, and so on. An example of using a synonym to change is quoted here.

One reason can be that though the speaker knows the general form of the idiom, he or she does not remember the intended word and just substitutes it with its synonym or another word with a similar meaning to prevent communication breakdown. This can be categorized under avoidance strategy in communicative strategies. Note the following statement:

- well these are actually er two sides of the same er things (VOICE, POwsd374: 244)

The word *coin* in the conventional form has been replaced with the general word *thing*, probably to compensate for not remembering the intended word and to convey the meaning by choosing a word that can semantically fit the idiom's meaning.

The same can be seen in the following statement from VOICE where *camel* replaces *fish* in the idiom, *drink like a fish*:

- S1: i drank almost one bottle?
- S1: like a CAMEL (VOICE, PBmtg3)

The speaker here (S1) tries to find a word which can semantically fit the conventional meaning of the idiom. Therefore, S1 selects *camel* that has the ability to drink large quantities of water very quickly to describe drinking *almost one bottle*. This shows the selected lexical item is not free, but is semantically bound to the essential elements of the idiom. In this example, *fish* and *camel* are both associated with drinking large amounts of water.

However, sometimes the substitution cannot be attributed to not remembering or lack of enough knowledge of the idiom. Consider the following:

- S1: … but (.) that didn't work out (.) e::r (.) at the end of the day or at the end of the month it was (.) very clear that hh the three other organizations e:r (1) didn't wish to hh to sit in in er in in: working groups … (VOICE, POmtg546: 343)

In the above excerpt, the speaker, first, uses the conventional form of the idiom *at the end of the day*, but then repeats the idiom with the word *month* as a semantically related word replacing the word *day*. This shows that the speaker knows the conventional form, but 'plays' with changing the idiom form maybe for more communicative effect.

The cases of synonyms, antonyms, or semantically similar or related words discussed above are all semantic types of substitution. However, substitution can also happen due to phonological similarity:

- S3: . . . next time they will be voted out and then there is this er general wisdom that er it doesn't pay to implement reforms because the voters will turn a blank eye on you (VOICE, PRqas409:12).

In this example, the word *blind* in the conventional form has been replaced with *blank*, which is phonologically similar to *blind*. This is in line with psycholinguistic research on slips of the tongue. According to Karlsson (2019), there are two main collaborating networks required to know a word: (1) phonological and orthographic aspects and (2) semantic aspects. There has been some evidence indicating that low achievers in a second language rely more heavily on sound relations, which are part of the first category.

An important point in substitutions is that, in most cases, either the essential words of the idiom or their meanings are retained in the variation. The meanings are retained through using synonymous or semantically similar words in the idiom variation. Other examples are *hen and egg* instead of *chicken and egg*, *hit the street* instead of *hit the road*, and *fall off the planet* instead of *fall off the face of the earth*, wherein the essential meanings of the words *chicken*, *road*, and *planet* are preserved in idiom variations though they have been replaced with other words. In the following example, the core metaphor has been retained, but the object to which it applies has changed:

- when you ascribe the same properties to things you you carve the world of in the same er . er you you you carve the world of at its joints as people say or carve the beast at its joints okay so we want properties of that kind and disjunctive properties do not carve the world er at its joints. (ELFA, USEMD120)

Here, the "carving" metaphor used by Plato as an "analogy for the reality of forms" to indicate "identifying distinct kinds of things" (Slater & Borghini, 2011, p. 1) has been retained while the object being described, that is, *nature* has been substituted first by *world* and, then, by *beast*. In the latter case, another metaphor has been added, that is, *beast* as an analogy for *nature* or *world* (or *things* being discussed in the excerpt). The word *beast* fits in the idiom semantically as it has *joints* literally and is in congruence with *carving*. The same is seen in the use of *not wake up any dogs* instead of *let sleeping dogs lie*, where the meaning of *let . . . lie* has been retained in *not wake up . . .*, thus avoiding changes that render the idiom meaningless or incomprehensible for the hearer. Another type of substitution is replacing a word with its antonym to get the opposite meaning of the conventional idiom. Examples are *keep the thread* instead of *lose the thread* and *loosen the grip on someone* instead of *get a grip on someone*. Here again, the essential words and themes are retained, but a different aspect of the metaphor is indicated in each idiom. In this type of lexical variation, the variable items do not have the same meaning as their substitutes, but they have related meanings. That is, the lexical variation is done in a way that does not harm the metaphorical mapping.

These results show that ELF users are aware of words that are semantically essential in the idioms and using such awareness, and they try to create variations of the fixed idioms via substitution, addition, and deletion of nonkey elements. The way ELF users obtain such awareness and the way they try to keep those essential parts intact in idiom variants are areas that require further research.

6. Conclusion

This study confirms the higher prevalence of idioms in the general corpus, which is more interactive, as compared to the academic corpus, which includes more monologic discourse. Idiom use in the academic corpus seemed to be register-sensitive as it could be predicted based on discipline. This can inform English for academic purposes (EAP) courses, especially in the area of materials development. In teaching English to medical sciences students, those idioms can be taught that are congruous with robustness and objectivity of the language in such disciplines. Formal variation was the most frequent idiomatic variation type in both corpora, and in most cases, it was used by ELF users without negatively affecting communication of meaning. This challenges the noncompositional approach to defining idioms that considers an idiom as a fixed chunk, which is stored and retried in the mental lexicon just like a single word. In a similar vein, idiomatic competence in ELF settings can be defined in terms of communication of meaning in idiom use rather than sticking to fixedness criteria promoted by native norms. The frequency and degree of conventionality of idiom variants derived from corpus analysis can guide the selection of such variants to be incorporated in EFL courses. Indeed, further research can shed light on strategies used to bring about variation in idioms while minimizing negative effects on communication of meaning. This study also showed the presence of idiomatic teddy bears or overuse of idioms by ELF speakers that can be attributed to subjective factors. Further research can investigate the role of subjective individual factors affecting the use of idiomatic teddy bears and their interaction with contextual factors.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

Abdou, A. (2011). *Arabic idioms: A corpus-based study*. Routledge.

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research, 19*(4), 329-358.

Baayen, H., & Neijt, A. (1997). Productivity in context: A case study of a Dutch suffix. *Linguistics*, *35*(3), 565-587.

Baayen, H. (1993). On frequency, transparency, and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology* (pp. 181-208). Dordrecht: Kluwer.

Baayen, H. (1992). A quantitative approach to morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology* (pp. 109-149). Dordrecht: Kluwer.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers, 26*, 181-190.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bobrow, S. A., & Bell, S. M. (1973). On catching on to idiomatic expressions. *Memory & Cognition, 1*(3), 343-346.

Bostancı, T. (2017). *The use of formulaic language in Asian and European ELF contexts:* A corpus-based study. Unpublished doctoral dissertation, Bilkent University.

Briggs, J. G., & Smith, S. A. (2017). English medium instruction and idiomaticity in English as a lingua franca. *Iranian Journal of Language Teaching Research, 5*(3), 27-44.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language, 27*(6), 668-683.

Carter, R. (2004). *Language and creativity: The art of common talk*. Routledge.

Eftekhari, N. (2008). *A brief overview on idiomatic translation.* Retrieved August 9, 2020, from the World Wide Web: https://www.translationdirectory.com/articles/article1739.php.

ELFA (2008). *The corpus of English as a lingua franca in academic settings*. Director: Anna Mauranen.

ELFA Corpus. (2019). *Description of the ELFA corpus project.* Retrieved March 21, 2020, from the World Wide Web: https://www.helsinki.fi/en/researchgroups/english-as-a-lingua-franca-in-academic-settings/research/elfa-corpus

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language, 6*, 22-42.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in context. *Memory and Cognition, 8*, 149-156.

Gibbs, R. W., Bogdanovich, J. M., Sykes, J. R., & Barr, D. J. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language, 37*, 141-154.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237-258.

Karlsson, M. (2019). *Idiomatic mastery in a first and second language*. London: Multilingual Matters.

Katz, J. J., & Postal. P. M. (1964). *An integrated theory of linguistic descriptions*. Cambridge, MA: MIT Press.

Khodabandeh, F., & Ramezani, M. (2020). Comparing formulaic sequences in English as a lingua franca and English as a native language in academic lectures. *Foreign Language Research Journal, 10*(3), 558-573.

Langlotz, A. (2006). *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. Amsterdam: John Benjamins.

Merriam-Webster (n.d.). Idiom. In *Merriam-Webster.com dictionary*. Retrieved July 12, 2020, from the World Wide Web: https://www.merriam-webster.com/dictionary/idiom

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

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Nunberg, G., Sag I. A., & Wasow, T. (1994). Idioms. *Language, 70*(3), 491-538.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi, 13*(1), 11-24.

Pitzl, M. L. (2012). Creativity meets convention: Idiom variation and remetaphorization in ELF. *Journal of English as a Lingua Franca, 1*(1), 27-55.

Pitzl, M. L. (2016). World Englishes and creative idioms in English as a lingua franca. *World Englishes, 35*(2), 293-309.

Pitzl, M. L. (2018). *Creativity in English as a lingua franca: Idiom and metaphor* (Vol. 2). Berlin and Boston: Walter de Gruyter GmbH & Co KG.

Prodromou, L. (2007a). Bumping into creative idiomaticity. *English Today, 23*(1), 14-25.

Prodromou, L. (2007b). Kettles of fish: Or, does unilateral idiomaticity exist? *English Today, 23*(3-4), 34-48.

Prodromou, L. (2003). Idiomaticity and the nonnative speaker. *English Today, 19*(2), 42-48.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics, 11*(2), 133-158.

Seidlhofer, B. (2004). Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics, 24*, 209-239.

Seidlhofer, B. (2005). English as a lingua franca. *ELT Journal, 59*(4), 339-341.

Seidlhofer, B. (2009). Accommodation and the idiom principle in English as a lingua franca. *Intercultural Pragmatics, 6*(2), 195-215.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37*(3), 419-441.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Slater, M. H., & Borghini, A. (2011). Introduction: Lessons from the scientific butchery. In J. K. Campbell, M. Rourke, & M. H. Slater (Eds.), *Carving nature at its Joints*: *Natural kinds in metaphysics and science* (pp. 1-31). Cambridge, MA: MIT Press.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18*(5), 523–534.

Vega-Moreno, R. E. (2001). Representing and processing idioms. *UCL Wording Papers in Linguistics, 13*, 73-107.

Vetchinnikova, S. (2014, September 6). *Approximation, remetaphorization, and idiomatizing in ELF phraseological patterning: Looking for the point of contact*. Paper presented at the 7th International Conference of English as a Lingua Franca (ELF7), DEREE, The American College of Greece, Athens, Greece.

VOICE. (2013). *The Vienna-Oxford international corpus of English* (version 2.0 XML). Retrieved July 12, 2020, from the World Wide Web: http://voice.univie.ac.at

VOICE Project. (2019). *Corpus description*. Retrieved July 12, 2020, from the World Wide Web: https://www.univie.ac.at/voice/page/page/page/corpus_description

Vilkaitė, L. (2016). Formulaic language is not all the same: Comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji Kalbotyra, 8*, 28-54.