الأسس النظرية والمنهجية للتحليل لوحدات التعابير المركبة الجامدة حسب الأسس الإحصائية لقواعد البيانات اللغوية

د. محمد عبدا لزهرة عريبي الحسين
جامعة البصرة/ كلية الآداب/ قسم الترجمة

تاريخ الاستلام : 16-01-2022
تاريخ القبول : ٢٠٢٢-٠٣-٢٩

ملخص البحث:

يتناول هذا البحث خاصية الجمود اللغوية للتعابير المركبة التي تترشح من مجموعة أسس ومبادئ عامة لها علاقة بالتركيب العقلي والتجسيد اللغوي وصيغ الاستخدام. يكمن فهم طبيعة وتدرج الجمود اللغوي في التوزيع التكراري لمثل هذه التعابير عند المستويات النحوية والدلالية والنصية. إضافة الى المقاربات الاستطرادية النوعية عند دراسة للتعابير المركبة الجاهزة لابد من التوجه لمقاربات كمية إحصائية لدراسة التوزيع التكراري قابلة لبرهنة ثبوت الجمود اللغوي للتعابير و مقياس حالة الجمود. لتحصيل هذا الهدف يمكن الاعتماد على الأطر النظرية للدلالة التوزيعية و الدلالة الادراكية لشرح النتائج التحليلية. التحليل الاحصائي الكمي يحتاج الى قاعدة بيانات لغوية للتكون مجال للبحث و الحصول على نتائج توزيع تكراري تسجل عدد تكرار استخدام المستخدمين للغة لمثل هكذا تعابير و بنفس الصيغة في نصوص معينة او مجالات معرفية معينة. يكمن اسهام البحث في تكوين الأسس النظرية لتكوين نموذج احصائي يمكن استخدامه لبرهنة ثبوت الجمود اللغوي للتعابير و مقياس حالة الجمود واقتراح استخدام نموذج الحدود القصوى لتكرار الاستخدام كنموذج قابل للتطبيق لبلوغ اهداف البحث.

الكلمات المفتاحية: وحدات التعابير الجامدة، التسلسل المتكرر الأقصى؛ معدل التكرار، السياق اللفظي ، أسس نظرية ومنهجية ، التحليل الإحصائي ، قاعدة البيانات اللغوية

THEORETICAL & METHODOLOGICAL FOUNDATIONS FOR CORPUS-BASED ANALYSIS OF FORMULAIC EXPRESSION UNITS

Dr Mohamed A. Al-Husain
Al-Basra University College of Arts.

Receipt date: 2022-01-16
Date of acceptance: 2022-03-29

ABSTRACT

The description of the formulaic status of the linguistic behaviuor of the formulaic expression units (FEUs) arises from a set of universal principles underlying the mental organization and representation of language and conventionalized patterns of language use. The nature and gradeability of formulaicity can be clued-up by the statistical distribution of such units at all levels of linguistic analysis including the syntactic, semantic, and discourse levels. In addition to the existing discursive approach, the FEUs' formulaic status must be quantitively approached and verified by a corpus-based statistical analysis of distributional frequency. Appropriate theoretical frameworks, including distributional semantics and cognitive semantics, should undoubtedly unveil formalized semantic and cognitive parameters which could better fit for the distribution frequency statistical analysis of the linguistic data. The corpus-based method allows retrieving sets of expression units to determine their formulaic status based on the frequency of occurrences in documents/domains collection. In this research, the model of corpus-based statistical analysis of distribution, proposed here, adopts two query-based information retrieval methods: (i) *n*-gram corpus Maximum Frequent Sequences (*n*-gram-MFS) for the representation for weighing FEUs' formulaic status per n-gram corpus and (ii) Maximum Frequent Sequences (D-MFS) for weighing FEUs' formulaic status per document/domain. The use of the proposed model offers a syetematic verification tool to weigh and evaluate formulaicity status of FEUs.

Keywords: Formulaic Expression Units; Maximal Frequent Sequences; Frequency of Occurrence; Co-text, Theoretical and Methodological foundations; Corpus-based Statistical Analysis;

INTRODUCTION

In their daily communication, native speakers naturally opt for conventionalized word sequences which are produced fluently- with a level of "*creativity"* and "*conceptual fluency"* (i.e. without planning pauses) for they usually exhibit significant "*Phonological coherence*"(see Altenberg 1998), e.g., "*sound like a pound*" or articulation fluidity *"Okey Dokey", "let it go as you go*" (Bybee 2002). Erman & Warren (2000) point out such word sequences, as "*you know*?", "*one moment, please*", "*I mean..*", "*in fact...*"; "*how do you do*?", "*saddle up*", etc", made up 52.3 percent in the corpus of the spoken language they analysed and 52.3% in the corpus of written language. Users registered higher speed execution or empirical response time consume limited cognitive resources (Erman, 2007; Goodkind & Rosenberg, 2015).

Formulaic expression units (FEUs) are an integral part of the lexicon. Single words can form variably complex units, such as phrases and formulaic expressions (sentences/utterances). Complex units. with more or less formulaic status, are variably called *phraseological expressions* (Gläser,1998), *set expressions* (Safarova, 2019), *idiomatic expressions* (Titone & Connine,1999), *multi-word expressions* (Wray & Perkins, 2000)*, binomials* (Carter, 1998*), prefabricated linguistic sequences*/routines (Brown 1973); *prefabricated patterns* (Hakuta, 1974); and *prefabricated routines/prefabs* (Bolinger 1976) *lexical chunks* (Schmitt 2000*), frozen expression* (Lee, 1993).

Despite the massive literature available at hand, the subject of controversy for the present article starts at the formulaic status of FEUs: it is still not obvious how one could identify conjoined multiword units due to prespecified *formulaicity status*? How could their formulaic status be objectively formalized and quantitively verified?

The contribution of the research is mainly theoretical: it dialectically argues for a set of theoretical and methodological foundations for a formalized representation, Corpus-based statistical methods, Maximal Frequent Sequences (MFSs)- to calculate the frequency distribution and weigh the formulaicity status of words sequences (FEUs) in a corpus (set of documents/domains) and determine their categorical membership.

1.1. Terminological Dilemma: and Divergent Approaches

Mainly, scholars emphasized different properties that signal their fixedness and formulaicity. Attempts underlined their complex internal structure, such as *multi-word items* (Moon, 2015); *lexical units,* Cowie (1992); *lexical phrases*, Nattinger & DeCarrico (1992) and *lexicalized sentence stems,* Pawley & Syder (1983). Other focussed on the property structural immutability, "*morpheme equivalent units*" (Wray 2008), and non-compositionality or formulaicity "*Formula-type sequences*", (Titone & Connine, 1999). FEUs are multi-morphemic sequences that are not constructed using rules, but - like a single lexeme - are called up as a "*whole*" (Aguado 2002: 30). The terms such as *unit, sentence, routine*, seem to be conceptually unfit in this regard, as they are either too general (unit) or entail additional conceptual associations (*unit, sentence, routine*).

Biber et al (1998: referred to such variably conventionalized forms as "*lexical bundles*", i.e, "recurrent expressions, regardless of their formulaicity, and regardless of their structural status" (p. 990; see also Hyland, 2008). The term "*chunk*" refers to a words sequence that is lexically represented and conceptually stored as a whole (cf. Schmitt 2000: 101; Krishnamurthy 2002: 289), irrespective of their conversational functions or contexts (Siyanova-Chanturia, 2015). Similar to such a perspective, Lewis (2002) used *lexical chunks*/*multi-word prefabricated chunks*; Schmitt (2000) favours *lexical chunks*.

Automation arises due to their ready-made (prefabricated) status: Brown (1973) *prefabricated routines*; Hakuta (1974) *prefabricated patterns*; and Bolinger (1976) *prefabricated routines/prefab*s. Bärenfänger (2002: 120) states that "automatisms" or "prefabrication", underline "*formulaicity*" properties, (Cf. *formulaic speech* Fillmore 1976). Peters (1983) mentions "*speech formulae*" which becomes "available to a speaker as a single prefabricated item in her or his lexicon" (p. 2). Similar terms are also used: *formulae* (Raupach 1984); formulas (Hickey 1993, Ellis R. 1994); *formulaic language* (Weinert 1995); and *formulaic sequences* (Wray & Perkins 2000; Aguado 2002).

Similar to their terminological turmoil, there are different definitions available for FEUs, led by disparate theoretical constructs and perspectives[1]. The researcher accepts as generally valid the definition of a formulaic sequence proposed by Wray & Perkins (2000):

> "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray & Perkins 2000: 1).

FEUs fall into several categories based on different compositional properties of the combined elements (lexical, syntactic, semantic, pragmatic, and cognitive), their number, fixedness (freeze), and archaism. Often, FEUs' structure is the starting point for any classification (Lewis 2002: 92ff; Wray & Perkins 2000: 4). Lyons (1968:77) referred to such expressions as "ready-made utterances", defined such linguistic units as "expressions which are learned as *unanalyzable wholesand employed on particular occasions by native speakers*" (italics added; see also, Erman and Warren 2000: 31).

Ronald Carter (1998) argues that the established categories of "*Multiword expressions*" addressed the linguistic structural hierarchy of binomials, based on *what is conjoined*, e.g., (nouns: brother and sister; pronouns: this and that, propositions: in and out; adverbs: up and down; etc.), *the type of conjunction* (and, but, etc.) and/or reversibility (reversible and non-reversible).

Lewis (2002: 92ff) distinguishes between the following categories: *polywords* (taxi rank; record player; by the way; of course); *collocations* (prices fall; incomes rise; unemployment stabilised); *Meta-messages*, e.g. for that matter… (message: I just thought of a better way of making my point); …that's all (message: don't get flustered); *institutionalized expressions, (*e.g., this is to certify that; to whom it may concern; I look forward to your early reply*); grammaticalized utterances* (not yet; certainly not; just a moment, please); *Sentence heads or frames* – most

[1] It should be emphasized here that the term "Formulaic Expression Units" is used in this article with a "neutral" meaning.

typically the first words of utterances, serving a primarily pragmatic purpose, e.g., "You know, but" to interrupt; "making the long story short" to summarize (see pragmaticalization by Fox Tree & Schrock 2002); and *Full sentences* –with readily pragmatic meaning.

There are various categories of the formulaic language, including phrasal verbs, collocations, idioms, lexical phrases, lexical bundles, proverbs, etc., with archaic nature. Proverbs, provide a short practical life advice, moral truth assume an archaic identity, e.g. "*A watched pot never boils*". The archaic FEUs, for the most part, are used exclusively with highly fixed formulaic status (Svensson 2004), e.g., "*vice versa*", meaning "the other way round". Other archaic forms are borrowed from other languages such as "*vis-à-vis*", /viːzɑːˈviː, *French* vizavi/, meaning "in relation to; with regard to".

Many typological taxonomies were suggested to account for phraseological units, retaining variable formulaic status (Lamiroy, 2016). Van Lancker (1987) suggested a continuum of formulaicity status based on the degree of *fixedness* extends such continuum to cover *propositional* and *non-propositional*. Propositional speech is made up of newly-created, original, novel automatic language use, utterances. Non-propositional speech includes conventional and overlearned expressions of all kinds, including idioms, speech formulas, proverbs, expletives, serial lists, rhymes, song titles, sayings, etc. Figure (1) shows categories and properties of non-propositional language as a continuum between novel and reflexive:

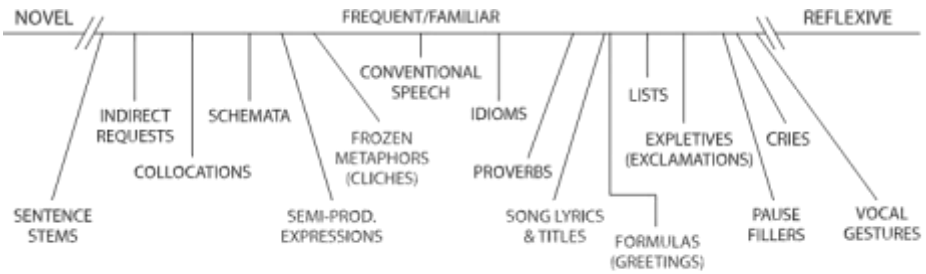

Figure 1. Van Lancker's continuum between novel and reflexive (after Van Lancker 1987: 56)

Figure 1: the continuum of PEUs (Van Lancker, 1987:56)

Yet, there is hardly any prototypical categorical name, a reference definition, or verifiable classificatory criteria for FEUs with significant mutual agreement among scholars and exclusive quantitative measures. This terminological conceptional confusion reflects the discursiveness of the data and methodology of the existing theoretical approaches. It is believed, here, that FEUs' formulaic status must be quantitively approached and verified by a corpus-based statistical analysis of their distributional frequency. Appropriate theoretical frameworks, including distributional semantics and cognitive semantics, should undoubtedly unveil formalized semantic and cognitive parameters. The corpus-based method allows retrieving sets of word sequences to determine their formulaic status based on the frequency of occurrences in documents/domains collection.

In dealing with the extraordinary nature of FEUs, we primarily commit to the belief that as a linguistic phenomenon, it is not solely attributed to the linguistic principles inherent to language, but the general principles of human cognition as well. The hereinafter principles capture the psychological plausibility and the formalized corpus-based statistical distribution model to satisfy empirical efforts. The effect of such principles methodologically necessitates enough data compilation for plausibility measure. Statistical distribution analyses will safeguard validity and reliability measures.

THEORETICAL FOUNDATIONS: GENERAL PRINCIPLES

An approach to FEUs must be based on psychologically plausible principles and empirical evidence (Evans & Green, 2006: 17). Due to their relevance, the following principles will discursively set the scenes and make up the premises for a set of formalized statements that could represent and verify the formulaic status of word sequences. They can become keywords for corpus-based queries to determine the weight, cohesion, and variability via the model of Maximal Frequent Sequences (MFS's) indicated in section (4.3).

2.1. The Principle of Compositionality

In any language, word forms [f_1, f_2, ..f_n], supported by the compositionality principle and constrained by a set of morpho-syntactic restrictions [SR] (Moon 1997), can conjoin to form more complex lexical units [F]. The composition also involves an integration of their lexical content structure [LCS], however, if [F] maintains an increase rate of frequency of occurances (Bybee & de Souza 2021), then grammaticalized, lexicalized, or pragmaticalized FEUs [F_i] evolve. The process of grammaticalization is conceived as a process during which a linguistic entity acquires a grammatical (or pragmatic) function in a particular morphosyntactic and pragmatic context (Heine, 2002; Bybee, 2003).

George Miller (1956) discovered that the process of grouping (chunking) is attributed to our brain's capability to adapt and tackle information overload. The brain fragments segments and groups the information into meaningful packages by mentally restructuring them, i.e., "*chunks*". So, overcoming the limitations of memory and managing to store more information after compartmentalizing it to improve productivity.That is why we do not remember our phone numbers, bank accounts, etc., at once, therefore need to divide them into blocks of 2, 3, or 4 objects chunks.

Steyn & Jaroongkhongdach (2016) distinguished three aspects common to most if not all formulaic sequences: a multitude of lexical constituents, no shorter than two words and no longer than five words in sequence, so $n = 2 \leq 5$; $n = [1,2,\ldots 5]$ as represented in (1) and (2). When it comes to counting the number of words in an expression, it is necessary to take into account the language in question. For instance, in Arabic "تسمعني" is equal to a 5-word sentence "are you listening to me).

$$[F] = [F_i]: n = 2 \leq 5 \quad \ldots\ldots\ldots\ldots \quad (1)$$

$$[F] = [f_1, f_2, \ldots f_5] \quad \ldots\ldots\ldots\ldots \quad (2)$$

Prototypically, the lexical content structure is the sum of the (prototypical) lexical content [LCS] of the conjoin words in a word sequence, making up a ready-made (prefabricated) spontaneous speech/writing [U].

$$[F]= [Fi] =LCS\ [f_1,f_2, f_3,..f_n] \quad \ldots\ldots\ldots\ldots\ldots \quad (3)$$

The term, formulaicity, is a clearly more discreet process. It is defined as a status whereby word sequence becomes inseparable, results in a complex unity and assumes lexical autonomy, at the expense of its conjoined elements autonomy, i.e., which in their turn assumes a loss of autonomy.

Becoming lexicalized over time, the origin of a word sequence can be found in their history of recurrent usage, i.e., grammaticalization, to arrive at a certain degree of usually referred to as (fixedness or non-compositionality) that we refer to as, "*FEU STATUS*". Word sequences, in a process of grammaticalization, undergo a loss of formal and semantic autonomy due to a set of parameters i.e. *weight, cohesion, and variability* (Lehmann 1995). Lehmann proposed parameters for gramatalized word sequences relate to syntagmatic and paradigmatic structures. The parameters make it possible to specify what the degree of grammaticalization of such units would be and aim to which linguistic mechanisms are at play in the process itself (Ibid)

The linguistic parameters underlying the autonomy of a linguistic expression, at play during the process of grammaticalization, can formally described in statement, (4):

$$[F]=\sum_{i=F}^{n=(1,2,..,5)} Fi = [f_1, f_2,\ldots, f_5] \quad \ldots\ldots\ldots\ldots\ldots \quad (4)$$

The lexical autonomy arises from the addition of the lexical content of the conjoined words to become a complex lexical unit which should meet Lehmann's parameter of *cohesion*. However, the Lexical Content Structure [LCS] of a word sequence is a function of its discourse circulation, (frequency of use). The compositionally coded meaning, formally represented in the statement (5), can be gradually integrated into the lexicon like, (safe and sound, etc.).

$$[F]=[[U[SR\ [F[\sum_{i=F}^{n=(1,2,..,5)} Fi = [f_1, f_2,\ldots, f_5)]\ LCS]]]] \quad \ldots\ldots\ldots\ldots \quad (5)$$

The loss of autonomy is correlated with increased cohesion. Cohesion refers to systematic lexico-grammatical relationships and constraints that the sign maintains with the other linguistic signs in the course of grammaticalization.

The compositionality principle, for Goldberg (2015: 419), "entail[s] that the meaning of every expression in a language must be a function of the meaning of its immediate constituents and the syntactic rule used to combine them.": a meaning that is derivable but not necessarily equivalent to the sum of the meaning in the combination of the words. For instance, the meaning of binomial, "ملح على الجرح", has a derivable emergent (non-compositional) meaning that extends beyond the sum of the meaning of the combined words, e.g., a painful cure. Most of the expressions such as idiomatic and metaphorical expressions are typically non-compositional, posing a serious semantic opacity, e.g., "سمن عل العسل", "*losing the thread*", etc.

Hopper & Traugott (1993) highlights the role of context in the emergence and evolution of certain linguistic forms. The linguistic context refers to the co-text which leads to another linguistic principle, collocation.

1.2. The Principle of Collocability and Contiguity

Much of the words' meanings are obtained principally from their co-text, i.e. the "characteristic co-occurance of patterns of words" (McEnery et al. 2006:149). The linguistic context (or co-text) therefore plays a crucial role in the emergence and the evolution of linguistic forms in the process of grammaticalization. Collocability refers to the notion of the proximity of the co-constituents in their co-textual environments (Wettler, Rapp & Sedlmeier, 2005). For word sequences to register a collocational status, we speak of constrained immediate constituency, characterized by more or less predictable lexical co-occurrences that it is specific to its Co-textual Context [Ux]. Priming experiments show that meanings are not informationally encapsulated, but are largely dependent on their co-texts (Marcel 1983).

Lexical collocability in word sequences, specific to their Co-textual Context [Ux], disambiguates their meaning, and affords semantic transparency for them. If the meaning of such units is viewed as a "process of increasing restrictions on options", seeking semantic transparency (Rieger 1991: 198), then it is essential to use voluminous data of real-life instances of language usage to quantitively specify the regular patterns of co-text cooccurrences.

Such premises make distributive analysis very relevant to the linguistic research on FEUs, where the structure of the semantic knowledge is strongly tied to the linguistic *contiguity,* Co-textual Context [Ux] as formalized in (6)

$$\mathbf{[F]=[[[[U[SR[[}\sum_{\boldsymbol{i=F}}^{\boldsymbol{n=(1,2,..,5)}} \boldsymbol{Fi} = (f_1 + f_2 + \ldots + f_5)\mathbf{]\ LCS]\ Ux]} \quad \ldots\ldots\ldots \quad (6)$$

The increase in the frequency of co-occurrence of co-text contexts makes word sequences more collocational, therefore they are more likely to be grammaticalized. Grammaticalization involving complex units is closely linked to collocation for it allows certain lexical units to form new collocations. During the grammaticalization process, the variable mobility reduction of the words on the syntagmatic axis to retain a more fixed syntactic positions is referred to as a *Syntactic Fixation Effect* (Lehmann, 1995) or *rigidification* (Croft, 2000).

Most agree that grammaticalization is not a mechanism of change in its own right, but invokes intension relying primarily on extension (Campbell, 2001: 141). Detached from the protypical lexical content [LCS] of the elements within a word sequence, the non-compositionality (idiomatic) property arises under the influence of the linguistic environment, Co-text [Ux] and situational environment, Context [X], giving peripherial Semantic Content Structuring, i.e., Sense [SNS], as formalized in statement (7)

$$\mathbf{[F]=[[[[U[SR[[}\sum_{\boldsymbol{i=F}}^{\boldsymbol{n=(1,2,..,5)}} \boldsymbol{Fi} = (f_1 + f_2 + \ldots + f_5)\mathbf{]\ LCS]\ Ux]\ SNS]\ X]} \quad \ldots\ldots\ldots \quad (7)$$

In other words, one of the properties of the process grammaticalization is the property of "*semantic generalization /reduction"* (Hopper & Traugott, 1993), which gives rise to recurrent senses over time, Semantic Content Structuring [SNS]. During the process semantic

generalization /reduction, certain components of meaning are lost and others become more generalized, leading to non-compositional meanings.

1.3. The principle of compression and Unpacking

The key to the compositional functioning of a linguistic units is the dynamic, self-developing and self-regulating nature of the conceptual unity, where the knowledge structures undergo processes of modification, fusion, compression, unpacking and restructuring, recorded and rolled up in the intrinsic physical matter of language. In the terminology of Fauconnier & Turner (2002) the semantics of compositional blending can be represented in the form of conceptual blending (back-stage process of fomulaicity), during which the initial concepts are combined into (relatively) final conceptual structures. They argue that conceptual integration, or conceptual blending, is the basis of the mental ability, leading to complex and emergent meanings (Cf. "*conceptual compression*", Heiko & Heine, 2011).

The human being is motivated through language by the tendency to obtain the maximum possible effect from the minimum effort. A significant proportion of children productions continue to rely on the use of fixed word sequencesfor they do not yet have the linguistic and cognitive skills necessary to analyze the structure of the statements (Peters 1983: 82). Even further, users compromise FEUs' formal complexity and lexical multitude with contraction and acronymization to retain their productive semantic content and the *conceptual fluency* of their production, however they. According to the principles of., Chunking and Compression, FEUs sometimes are acronymized in texting, abbreviated FEUs have been growing on texting social networking interactions and gaming websites, abbreviated variants, e.g., ASAP-As Soon Possible, W8-4-ME – Wait FOR ME; B4N – Bye for now; in business emails such as CEO- Chief executive officer, HR- Human resources, etc.

According to Zipf's law, both the speaker and the listener try to minimize their effort: the most economical compromise between the competing needs of the speaker to optimally encode and the listener's efforts to transparently decode messages is the kind of interrelationship between frequency and the linguistic hierarchy that appears in the data supporting Zipf's law(Lestrade,

2017). The principle of least effort, spending the least amount of effort to complete a task, is a fundamental principle in all human action, including verbal communication. Therefore, this tendency to minimize effort will produce a reduction at the various levels of the linguistic system, "be going to", "*gonna*". Formulaicity is a creative process that responds to the need for economy in language, i.e., the pragmatic requirement of the least cost for the maximum effect in a communication situation (Sperber & Wilson, 1986; Wray, 2008). For an FEU to become lexicalized, becomes an entry in a dictionary, it needs to achieve what we can call the *threshold point* of the mutual usage agreement*, the principle of mutual choice*, (Sinclair, 1991: 173).

1.4. The Principle of mutual choice

The principle of mutual choice illustrates the tendency of transcending the users' subjective idiosyncrasies into a mutually conventionalized/institutionalized meaning and function, via, intersubjectivity (Sinclair, 1991: 173) for the linguistic units that persistently co-occur and re-occur in usage.

It is believed here that based on the mutual agreement, the conventionalization of word sequences would provoke grammaticalization, not the other way around. Pragmaticalization is a long-term process that reflects socio-interactionally pragmatic functions: Lexical units migrate towards mutually (subjectively and intersubjectively) recognized pragmatic function across the discourse in various domains (Erman & Kotsinas, 1993; Diewald, 2011). The increase in the frequency of co-occurrence of lexical items is what makes them more conventionalized, therefore more likely to be grammaticalized, moreover to becoming lexicalized, and eventually pragmaticalized (see Figure 2).

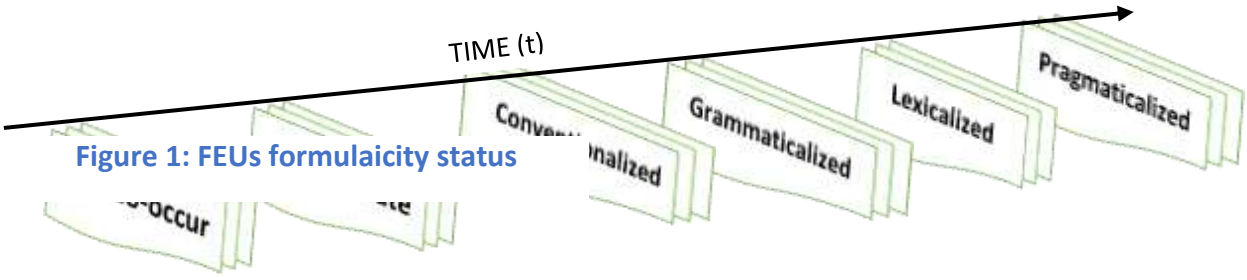

Figure 1: FEUs formulaicity status

Another level of metalinguistic abstractions, "Subjectivation or Intersubjectivation", (Traugott, 1997: 32) argues refer to when lexical units would acquire an expressive function, to serve increasingly abstract, pragmatic, interpersonal, and speaker-based functions. A complete determination of the meaning is nothing but measuring the difference in the selection of different co-texts approximately based on the set of all identifiable *environments* at a point of *time* [*t*] [Ux, X]^{t}. From such differences in the parameters, inferential processes vary subjectively and intersubjectively, leading to emergent non-compositional meaning as could be formalized in (8).

$$F=[[[[U[SR[[\sum_{i=F}^{n=(1,2,..,5)} Fi = (f_1 + f_2 + \ldots + f_5)] \text{ LCS}] \text{ Ux}] \text{ SNS}] \text{ X}] \text{ SCS}]^{t} \quad \ldots\ldots \quad (8)$$

FEUs' potential for contextualization makes them connected with intersubjective evaluations (extension and intension) and pragmaticalization, acquiring variable illocutionary forces, for instance, "I mean" which can extend to apply its forward-looking functions appropriately to diverse situations like instruct, paraphrase, repair or a less-face-threatening act, etc. (Fox Tree, & Schrock, 2002).

1.5. Formulization of FEUs Formulaicity Status:

It has become obvious that the puzzle of the formulaic status of word sequences is theoretically complete, based on some linguistic parameters (compositionality and collectability) and cognitive parameters (compression and mutual choices).

Harris defines the distribution of a sign, its relationship to all other signs of the same system- as the sum of all environments in which a language entity is located. An environment is understood to mean its respective co-competitors, i.e. elements with which it comes together in a certain combination to form an utterance (see. Harris 1954). In other words, the difference in meaning correlates with the difference in distribution. Linguistic signs never appear with one another arbitrarily, but always in very specific relationships to one another.

The Distribution Semantics approach is strongly appealing for the possibilities of strictly empirically-based corpus linguistics. This motivates for compiling large machine-readable text

corpora in Arabic but also develops special research formalizations with which they can be specifically searched and statistically processed. The linguistic corpus can statistically afford such verification but counting the frequency of co-occurrences formalized in the statements (3), (4), and (5). After all, no semantic approach can capture the meaning of a linguistic sign in its totality without reference to the qualitative or quantitative account of the contexts in which it was used. The distribution sequence's variable senses become strongly influenced by contextual and cognitive parameters at a given point of time as formalized in (8).

$$F=[[[[U[SR[[\sum_{i=F}^{n=(1,2,..,5)} Fi = (f_1 + f_2 + \ldots + f_5)] \text{ LCS}] \text{ Ux}] \text{ SNS}] \text{ X}] \text{ SCS}]^t \quad \ldots\ldots \quad (8)$$

Based on real-life language usages compiled in a corpus, this formalization is not only for identifying FEUs but to weigh their formulaic status, as well. For instrumental corpus-based analyses of distribution, the appropriate Information Retrieval Method must be incorporated as a method that supports indexing the existence, location, and frequency of occurrences of FEUs in a set of documents/domains.

METHODOLOGICAL FOUNDATIONS FOR CORPUS-BASED STATISTICAL ANALYSIS

Indexing the existence, location, and frequency of occurrences of FEUs, as a phenomenon in the whole language, requires large, manageable data that can be systematically searched, analyzed, and evaluated (Arnon & Snider, 2010). In traditional linguistic research, the data is discursively selected and the focus is mainly on peculiarities or unusual phenomena - a subjectively biased approach.

4.1. Data Corpus

A Language corpus represents a sample of real use, both spoken and written with a wide spectrum of unprecedented linguistic material that is not built selectively (Teubert & Krishnamurthy, 2007). The corpus is valuable for linguistic research, in particular for some reasons (BAAYEN, 2008). This data selection is clear, objective, and stereotype-free: texts

selection is not subjective (subjectivity is suppressed as possible). The source of language information is non-specific, it is readily available to every user and can be used to research various phenomena. Electronic corpora such as British Nation Corpus is supported with a search engine for making complex queries in several seconds.

Research in this field chooses one of mainly two approaches: The corpus-Based approach (i.e. corpus-verified research question) and the Corpus-Driven approach (i.e. corpus-inspired research objective (see Tognini-Bonelli, 2001). In the corpus-based approach, the researcher approaches linguistic data with a pre-created, introspection-based hypothesis and seeks confirmation (or refutation) for some arguments using the corpus. The corpus-driven approach creates concepts and descriptive structures to build a description of a given segment of linguistic reality, depending on the observation of a given data corpus. Both approaches could serve a more systematic and plausible description of FEUs and the proposed statistical analyses of FEUs semantic distribution. The latter will build a more transparent picture of FEUs' status individually and categorically.

1.6. Corpus Linguistics

Corpus linguistics is a branch that deals with methods of building corpora with a more or less technical toolbox for improving and supplementing language descriptions, discovering new relationships to describe linguistic reality (Teubert & Krishnamurthy, 2007). The compilation of corpora requires a set of tasks, e.g., *lemmatization, Stemming, linguistic tagging, query,* and *search* (Biber, Conrad & Seppen, 1998; Al-Omari, 1994; Abuata, Sembok & Bakar, 2011)).

Essentially responsible for solving a user's query, an IR (Information retrieval) method reports the existence, location, and frequency of occurrences of FEUs in a set of texts. Queries based on *keywords* (Kaur 2010; Wartena, Brussee, Slakhorst 2010), in our case FEUs, can afford the most important information of their formulaic status. In a query, a set of words such as "على سبيل " can be placed to find the texts "على سبيل المثال" in that order, and not all texts that contain it. Another idea also used is to represent texts/domains by terms called *n*-grams (Nie, 2000), which are sequences of *n* words or consecutive characters. This representation gives the possibility of

having representative FEUs in texts/domains formed from 1 to *n* words count, due to the frequency of occurrences.

The corpus-based statistical analysis of distribution, proposed here, adopts two query-based information retrieval methods: (i) Maximum Frequent Sequences per *n*-gram corpus (*n*-gram-MFS) for the representation for weighing FEUs' formulaic status and (ii) FEUs that are repeated within the same document/domain, are known as Maximum Frequent Sequences (D-MFS) per document/domain.

1.7. Maximal Frequent Sequences (MFS's)

The adopted IR models are the Boolean model and the vector model. In the Boolean model (Kowalski, 1997), the documents are represented by vectors, the different sequences are equal in size to the number of that appear in the index that was generated from the collection of documents. Each element of the vector represents the appearance of a sequence within an indexed document, (1 if it appears and 0 if it does not appear). The Boolean representation of the documents would be as seen in the table below.

Corpus (Documents/domains)	FEU Vector					
	X_1	X_2	X_3	X_4	X_5	X_n
Document $_1$	1	1	1	1	0	0
Document $_2$	1	0	0	0	0	0
Document $_n$	0	1	0	0	1	1

The search for the FEUs is carried out by taking the vectors of the Boolean matrix of documents vertically for the query keyword found in the index, for the example "X_1 AND/NOT X_2" where " X_1" and" X_2 would be the vectors of the sequences", as would be obtained in the table bellow:

	Corpus (Documents/domains)

FEU Vector	Document 1	Document 2	Document 3	Document 4	Document 5	Document n
X_1	1	1	1	1	0	0
X_2	1	0	0	0	0	0
X_n	0	1	0	0	1	1

Finally, it can be said that the FEUs' search and retrieval from the relevant documents/domains are based on the criterion of the Boolean matrix, where a document/domain is relevant if it has a "1" in the resulting vector, otherwise, it will be irrelevant.

The vector model (Baeza,1999) represents FEUs distribution in a set of documents/domains using vectors, where the weight of each vector is given by the number of indexed sequences in the collection. The value of each element of the vector $p_i(t_j)$ represents the weight of the FEU j in document i.

$$p_i(t_j) = \begin{cases} 1 & if\ t_j\ is\ in\ document\ \ i \\ 0 & otherwise \end{cases}$$

$$p_i(t_j) = tf_{i,j}$$

where $tf_{i,j}$ = frequency of term j in text i.

$$p_i(t_j) = idf_{i,j} = \log [\, td/d_j]$$

$IDF_{i,j}$ = Inverse Text Frequency

$IDF_{i,j}$ is the inverse proportion of the number of documents in the collection that contains a queried keyword sequence. The more documents that contain a queried keyword sequence, its IDF will be lower. On the other hand, the fewer documents that contain a queried keyword sequence, its IDF will be higher.

td = Total texts in the collection.

d_j = Number of texts containing the FEU j

The D-MFS's are used per document because they can describe the document of the documents while preserving the sequential order of the words, in n-grams. minor modifications during the process of collecting materials and their subsequent processing can be done to extend the search for other possible co-texts cooccurrences.

To describe the process of collecting, processing the corpus, and highlighting the use of queries keyword sequence, FEUs, using TF-IDF matrix in documents from electronic media materials presented in the form of document information messages for a certain period. In this case, the use of electronic media documents is quite convenient, since the materials of the publication are immediately published on the site in a certain category. In other words, the compiled texts are automatically marked as an index of their category e.g., "Society", "Economy", "Politics", "Sport", "Culture". Such collected corpus of documents allows us to refer to a sufficient sample from various domains.

The first step in text processing was *lemmatization* - the reduction of words to their vocabulary forms. The final transformation was stemming - the process of highlighting the stem of a word. After removing hyphens from the text, to identify lemmas of words morphological analysis was used for stem identification. The Arabic stemming algorithm developed by Al-Omari is studied and new versions are proposed to enhance its performance. Pioneer works on Arabic stemming have been published by researchers such as (Al-Omari, 1994) and (Abuata, Sembok & Bakar, 2011).

In the TF-IDF matrix, all unique vectors are columns, therefore, in the source text, each separate word form will have its weight in a document subcategory (Ramos, 2003). To obtain more accurate values of weights and to reduce the dimension of the matrix, the source texts of documents are pre-processed (Korenius 2004). The TF-IDF values for each term in each document are entered into a matrix where columns are represented by occurrences and rows are represented by documents. These MFS's are repeated at least a certain number of times within the same document, so they can be considered descriptive.

Once the index terms (D-MFS's) per document from a collection of documents have been extracted, the system index is built, which is based on the inverted file technique and which has the following structure of absolute values[2]:

IDSTj=|NMFS | MFS | DT | FT | DOC1 | FD1 | FDN1 | IDF1 | FI1 |…| DOCDT | FDDT | FDNDT | IDFDT | FIDT|

where:

NMFS = Number of MFSs.

MFS's = Maximal Frequent Sequence.

DT = Total documents that contain the SFM.

FT = Frequency of MFS in all documents.

DOCj = Document number j where the MFS appears.

FDj = Frequency in document j of the MFS.

NFDj = normalized NFDj.

IFDj = Inverse Frequency in document j

DSTj= Distribution of j

IDSTj = NFDj * IDFj.

This index contains all the different SFMs per document found in the documents of a collection. This structure allows you to have all the information necessary to generate the vector representation of the documents in the collection and the query, as explained in the next section. In addition to speeding up the document search process.

CONCLUSION

The status and behaviour of FEUs must be accessed systematically rather than discursively by a quantitively verifiable corpus-based statistical analysis of their distributional semantics. Appropriate models, including distributional semantics and cognitive semantics models, should undoubtedly unveil formalized semantic and cognitive parameters which could better fit for the

[2] For any real number, the absolute value or modulus is denoted as (vertical bar on each side of the value) and is always either positive or zero , but never negative. An absolute value function is a function that contains an algebraic expression within absolute value symbols.

statistical distribution analysis of the linguistic data. The nature and gradeability of formulaicity can be clued-up by the statistical distribution of such units at all levels of linguistic analysis including the syntactic, semantic, and discourse levels.

REFERENCES

Abuata, Belal & Sembok, Tengku & Bakar, Zainab. (2011). A Rule-Based Arabic Stemming Algorithm. Proceedings of the European Computing Conference, ECC '11.

Al-Kharashi, I.A. & Evens, M.W. 1994. Comparing words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System. Journal of the American Society for Information Science. 45(8): 548-560.

Al-Omari, H. 1994. ALMAS: An Arabic Language Morphological Analyzer System. National University of Malaysia. Bangi,Selangor.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), Phraseology: Theory, analysis and applications (pp. 101–22). Oxford, England: Clarendon Press.

Arnon, Inbal & Snider, Neal (2010) More than words: Frequency effects for multi-word phrases, Journal of Memory and Language,Volume 62, Issue 1,Pages 67-82,

Baayen, Hr (2008): Analyzing Linguistic Data . Cambridge: Cambridge University Press.

Baeza Y. R., Ribeiro N.B. "Modern Information Retrieval". Ed. Pearson Addison Wesley, ACM Press New York. 1999

Biber, D. - Conrad, S. - Reppen, R. (1998): Corpus Linguistics. Investigating Language Structure and Use . Cambridge: Cambridge University Press.

Brown, Roger (1973). A First Language: The Early Stages. Ma.:Harvard University Press

Bybee Joan, (2002), "Phonological evidence for exemplar storage of multiword sequences", Studies in Second Language Acquisition, 24, 215-221.

Bybee, Joan. 2003. Cognitive processes in grammaticalization. The New Psychology of Language 2. 145–167.

Bybee, Joan & de Souza, Ricardo Napoleão. (2021). Predictability and prefab status: The case of adjective + noun sequences in English. In Aleksandar Trklja & Łukasz Grabowski (eds.), Formulaic language: Theories and methods. Berlin: Language Science Press, 3–30.

Carter, R. 1998 (2nd edition). Vocabulary: Applied Linguistic Perspectives.London: Routledge.

Cowie, A. P. (1992.), Phraseology: Theory, analysis and applications. Oxford, England: Clarendon Press.

Croft (2000). Explaining Language Change: An Evolutionary Approach.

Diewald, G (2011).Pragmaticalization (defined) as grammaticalization of discourse functions. Linguistics 49–2 (2011), 365–390

Evans, V., & Green, M. (2006). Cognitive linguistics: An introduction. Lawrence Erlbaum Associates Publishers.

Erman, Britt. (2007). Cognitive processes as evidence of the idiom principle. International Journal of Corpus Linguistics, 12(1):25–53

Erman, B & Kotsinas, U. B. (1993). Pragmaticalization: The case of ba' and you know. In: J. Falk, K. Jonasson, G. Melchers, & B. Nilsson (Eds.), Stockholm Studies in Modern Philology (Vol. 10, pp. 76–93). Stockholm: Almqvist & Wiksell International.

Erman, Britt And Warren, Beatrice (2000). "The idiom principle and the open choice principle" Text & Talk, vol. 20, no. 1, , pp. 29–62.

Fox Tree, Jean E & Schrock, Josef C. (2002) Basic meanings of *you know and I mean*. Journal of Pragmatics,Volume 34, Issue 6,2002,Pages 727–747,

Gläser, Rosemarie. (1998). The stylistic potential of phraseological units in the light of genre analysis. In Anthony Paul Cowie (ed.), Phraseology: Theory, analysis, and applications (pp. 125–143). Oxford: Oxford University Press.

Goldberg, Adele E. (2015) "Compositionality". In Riemer, Nick (ed.) (2015). Routledge Handbook of Semantics. Routledge, pp 419–433

Goodkind, Adam & Rosenberg, Andrew (2015). Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production. Proceedings of NAACL–HLT, pages 87–95,

Hakuta, K. (1974), Prefabricated Patterns And The Emergence Of Structure In Second Language Acquisition. Language Learning, 24: 287–297.

Harris, Zellig Sabbatai (1954) “Distributional structure”. Word. Journal of the linguistic circle of New York. 10, 2–3, 146–162

Heine, Bernd. 2002. On the role of context in grammaticalization. In Ilse Wischer & Gabriele Diewald (eds.), New reflections on grammaticalization, 83–101. Amsterdam: John Benjamins.

Narrog, Heiko & Heine, Bernd (2011) The Oxford handbook on grammaticalization. Oxford: Oxford Univerity Press

Hickey, Tina. (1993). Identifying formulas in first language acquisition. Journal of Child Language, 20(01):27–41.

Hopper, Paul J. & Traugott, Elizabeth Closs (1993) Grammaticalization. Cambridge: Cambridge University Press,. Pp. xxi 256.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4–21.

Kaur J., Gupta V. Effective approaches for extraction of keywords // International Journal of Computer Science Issues (IJCSI). 2010. Т. 7. № 6. С. 144.

Korenius T. et al. Stemming and lemmatization in the clustering of Finnish text documents // Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004. С. 625–633.

Kowalski G. (1997). “Information Retrieval Systems Theory and Implementation”. Press Kluwer Academic Publisher..

Lamiroy, Béatrice. (2016). For a typology of phraseological expressions: how to tell an idiom from a collocation? In: Orlandi, Adriana & Giacomini, Laura (eds.) Defining collocation for lexicographic purposes. From linguistic theory to lexicographic practice. Bern: P. Lang.

Lee, Chungmin. (1993). Frozen Expressions and Semantic Representation. the Language Research Institute, Seoul National University

Lestrade, Sander. (2017). Unzipping Zipf's law. PLoS ONE. 12. 10.1371/journal.pone.0181987.

Lewis, M. (2002) [1997]. Implementing the Lexical Approach. Boston: Thomson Heinle.

McEnery, A., Xiao, R., and Tono, Y. (2006) Corpus-Based Language Studies: An Advanced Resource Book.London, U.K.: Routledge.

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. Cognitive Psychology, 15(2), 238-300.

Miller GA(1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review.;63:81-97.

Moon R., (1998), Fixed Expressions and Idioms in English: A Corpus-based Approach, Oxford: University Press, Oxford

Nattinger, James R. & DeCarrico, Jeanette S.. (1992). Lexical Phrases and Language Teaching. Studies in Second Language Acquisition, 16(2), 254-254.

Nie J. Y. (2000) "On the use of words and N-grams for Chinese Information Retrieval". IRAL-2000, Fifth International Workshop on Information Retrieval with Asian Languages. Institute of System Enginnering, pp.141-148. Hong Kong, China.

Pawley, Andrew & Syder, Frances Hodgetts (1983).Two puzzles for linguistic theory: nativelike selection and nativelike fluency. Language and communication, 191, 225.

Ramos J. et al (2003). Using TF-IDF to determine word relevance in document queries //Proceedings of the first instructional conference on machine learning..T. 242. C. 133-142.

Rieger, B. B. (1991). On Distributed Representation in Word Semantics. [ICSI-Technical Report TR-91-012], International Computer Science Institute, Berkeley, CA,

Schmit, Norbert (2000). Lexical chunks, ELT Journal, Volume 54, Issue 4, Pages 400–401.

Sinclair, J. M. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. Corpus Linguistics and Linguistic Theory, 11(2), 285-301.

Steyn, Sunee & Jaroongkhongdach, Woravut (2016).Formulaic Sequences Used by Native English Speaking Teachers in a Thai Primary School. PASAA Volume 52,105-132.

Teubert - R. Krishnamurthy (2007) Corpus Linguistics. Critical Concepts in Linguistics (vol. I) . London / New York: Routledge, pp. 93–118.

Titone, D. A. & Connine, C. M. (1999). On The Compositional and Noncompositional Nature of Idiomatic Expressions. Journal of Pragmatics, 31, 1655-1674.

Tognini-Bonelli, E. (2001): The Corpus-driven Approach. In: Corpus Linguistics at Work. Amsterdam: John Benjamins, s. 84–100.

Traugott, Elizabeth (1995)"Subjectification in grammaticalization", in Dieter Stein and Susan Wright, eds., Subjectivity and Subjectivisation. Cambridge: Cambridge University Press, 37-54.

Traugott, Elizabeth. (2009). "Lexicalization and grammaticalization", "Subjectification, intersubjectification, and grammaticalization", Studies in Historical Linguistics 2: 241-271.

Van Lancker Sidtis, Diana & Kempler, Daniel. (1987). Comprehension of Familiar Phrases by Left but not Right Hemisphere Damaged Patients. Brain and Language. 32. 265-277.

Wartena C., Brussee R., Slakhorst W. (2010). Keyword extraction using word cooccurrence/ Workshops on Database and Expert Systems Applications.IEEE, 2010. C. 54–58.

Wettler, M., Rapp, R., & Sedlmeier, P. (2005). Free Word Associations Correspond to Contiguities between Words in Texts. Journal of Quantitative Linguistics, 12 (2–3),111–122.

Wray, A. (2003). Formulaic Language and the Lexicon. Journal of Pragmatics. 35. 10.1016/S0378-2166(03)00079-1.

Wray, A. (2008). Formulaic Language: Pushing the boundaries. Oxford Applied Linguistics. Oxford: Oxford University Press. ISBN 978-0-19-442245-1. 305 pp.. 103.

Wray, A. & Perkins, M. (2000). The functions of formulaic language: An integrated model. Language & Communication - LANG COMMUN. 20. 1-28.

Zhang, Ling & Lu, Ping. (2017). Lexical Chunks Formulaic Sequences and Yukuai: Study of Terms and Definitions of English Multiword Units. English Language and Literature Studies. 7. 74.