The Routledge Handbook of Vocabulary Studies

Stuart Webb

Classifying and Identifying Formulaic Language

Publication details

David Wood
Published online on: 13 Aug 2019

How to cite: - David Wood. 13 Aug 2019, Classifying and Identifying Formulaic Language from: The Routledge Handbook of Vocabulary Studies Routledge
Accessed on: 05 Dec 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Introduction

Formulaic language (FL) is generally defined as multiword language phenomena which holistically represent a single meaning or function, and are likely mentally stored and used as unanalyzed wholes, as are single words. The phenomenon itself is generally called formulaic language, and items themselves are referred to as formulaic sequences. FL is a rather enigmatic and elusive element of language, in itself it is a relatively recent subject of focus in linguistics and applied linguistics even though key categories and types have been subjects of scrutiny since the early to mid 1900s.

Over a long period of time the multiword units we now call formulaic language were examined more or less in isolation. This is largely due to the fact that researchers were looking at different and relatively discrete categories of multiword units, often working in quite separate areas of linguistics or other fields, including social anthropology and neurology. Over time, the existing research was examined and reinterpreted as a whole body of knowledge, but it was the late 1990s when the term FL came into common use, largely as a result of the work of Wray (e.g., 1999). The term is now standard, and a great deal of important work has been conducted into many aspects of FL and its use.

Critical Issues and Topics

There is a surprising range and scope of types of formulaic language, as seen in detail later. The categories, when examined, show quite a bit of overlap and imprecision, and are subject to interpretation. For example, determining whether a given sequence is a collocation or an idiom is sometimes a challenge. Some items do not fit comfortably in any specific category, or fall into cracks, for example sequences like and then or sooner or later are really difficult to categorize. Advances in corpus analysis technology and techniques have helped uncover new types of formulaic sequences, but all the same a sort of orthodoxy has been established over time, which can be puzzling. It is unclear, for example, how significant it is to determine a lexical bundle by means of frequency only, as compared to a sequence identified using frequency in combination with other statistical measures. As well, some categories overlap,
and, perhaps most alarmingly, there is no firm consensus that all the categories are similarly processed semantically or psycholinguistically. It is also logical to question whether the categorizations are useful to researchers or teachers. Is it possible that the classifications are really just leftovers from early studies in phraseology, and that they are largely irrelevant to concerns particular to applied research or language teaching?

The identification of formulaic language in spoken or written texts is a challenging enterprise. Formulaic sequences may be identified in corpora by frequency, and the ways in which formulaic sequences are produced also gives us clues as to what multiword combinations might be formulaic. A potentially valuable means of determining formulaicity involves expert or native speaker judgment, especially useful when applied to small or very specific data sets. Although it is often best to try to use a combination of measures, in many cases absolute certainty in identification can be elusive. Even using combinations of corpus frequency and statistical measures of co-occurrence, along with acoustical features and judges, it is common to hedge one’s claims about formulaicity. We can hope that new and more reliable or valid means of identifying formulaic language will come along in time.

Classifications

FL has been labeled in many different ways, and nearly 20 years ago it was Wray and Perkins (2000, p. 3) who identified 40 terms. The main categories of FL are collocations, idioms, lexical phrases, lexical bundles, metaphors, proverbs, phrasal verbs, n-grams, concgrams, and compounds. Some of the sequences are characterized mainly by their structural/semantic/syntactic properties, some by their pragmatic utility, and some by their distribution in particular corpora.

Sequences Distinguished by Structural, Semantic, or Syntactic Properties

Collocations

The term collocation has been around for many years, and research has been inspired by the pioneering work of Firth in the 1950s. Collocations likely come into use because of repeated context-dependent use. Such terms as *senior management*, *single parent*, and *plastic surgery* are examples of collocated pairs of words. Collocation basically refers to a syntagmatic relationship among words which co-occur. The relationship may be restricted to relationships which conform to certain syntactic and/or semantic criteria. Collocations can be in a syntactic relationship such as verb + object for example *make a decision*. Two approaches to collocation research have dominated: frequency-based and phraseological (see Granger & Paquot, 2008, for an overview of these). The frequency-based approach is in the tradition of the work of Firth (1951, 1957) and deals with the statistical probabilities of words appearing together, while the phraseological approach, dating back to very early work in Soviet phraseology, is much more concerned with restrictive descriptions of multiword units, and takes quite narrow views of what constitutes a collocation. To add to the complexity of the use of the term collocation, researchers have used it more creatively, sometimes as an umbrella term for multiword units in general. Frequency-based work on collocation was developed by Halliday, Mitchell and Greenbaum, Sinclair, and Kjellmer. These researchers extended and refined the definition to specify that a collocation is a function of the frequency of a word appearing in a certain lexical context as compared to its frequency in language as a whole.
They included syntactic and semantic aspects in descriptions of collocations, and explored the issue of what span of words to consider a collocation. Jones and Sinclair (1974) found that the span of words which is optimal for a collocation is four words to the right or left of a node, or core word. Kjellmer worked on the *Dictionary of English Collocations* (1994) defining a collocation as a continuous and recurring sequence of two or more words which are grammatically well formed. This led to the development of computer-based frequency driven study of collocations.

Unlike the frequency-focused researchers, phraseologists tend to see collocations as multi-word units whose component relations are variable and whose meaning is somewhat transparent (Nesselhauf, 2005). For example, Cowie (1994) placed collocations along a scale from composites, combinations below the sentence level with lexical or syntactic functions (e.g., *red herring*), and formulae, often sentence-length and having pragmatic functions (e.g., *how are you?*). Composites can be fully opaque and/or invariable, as in “pure idioms” (e.g., *kick the bucket*). “Figurative idioms” can have both a literal and figurative meaning (e.g., *to play a part*), and restricted collocations in which at least one element is literal and the other figurative (e.g., *explode a myth*). Cowie gives no restrictions on the number of words or the span of words in a collocation.

**Idioms**

Idioms are perhaps the archetypal formulaic sequence. Unfortunately, they are as ambiguous as collocations, and share with them a sort of dual personality, with idiom referring both to a specific type of FL, and some researchers using the term more broadly, with definitions encompassing proverbs, slang expressions, and so on. In general, however, the term is used to refer to word combinations which are, in the words of Moon (1998, p. 4), “fixed and semantically opaque or metaphorical”, for example, *kick the bucket* or *spill the beans*.

A key quality of an idiom is its semantic non-compositionality and non-productive form. Wood (1981) noted that the meaning of an idiom is not the sum of the meanings of its component parts, that is, it is not compositional, and its structure must not be transformable, that is, it is non-productive, or frozen. Examples of items which meet these two criteria are *kick the bucket* and *by and large*, which cannot be understood by means of their constituent parts, nor can they be grammatically manipulated. In fact, many types of formulaic sequences display idiomaticity to greater or lesser degrees.

The modern scholar with the most useful definitions and categories of idioms is Moon (1998), who defined idioms as “semi-transparent and opaque metaphorical expressions such as *spill the beans* and *burn one’s candle at both ends*” (p. 5). She differentiated idioms from fixed expressions such as routine expressions, sayings, similes, and so on (Moon, 1998, p. 2). Somewhat later, Grant and Bauer (2004) added the stipulation that an idiom is also non-figurative, meaning that its meaning must not be interpretable from the component words. Applying the criteria of Grant and Bauer, *kill two birds with one stone* is not an idiom because its meaning may be seen as nonliteral, and then interpreted again through examining its pragmatic intent. In contrast, applying the criteria of Grant and Bauer, *by and large* is an idiom, because it is both nonliteral and provides no indication of its figurative meaning.

In sum, there are five basic defining criteria of an idiom (see Skandera, 2004; Wood, 2015):

1. An idiom is two or more words in length.
2. Semantically opaque (the meaning of the whole is not the sum of the meanings of individual component words) – examples might be *spic and span* and *to and fro*.
in which the component words are also opaque, spic, span, or fro are never used outside of these contexts (see Allerton, 1984). Many instances of opaque idioms have historical roots, for example kick the bucket (die), relates to the slaughtering of pigs.

3 Noncompositionality – the words that make up an idiom cannot be analyzed for meaning or function. This is akin to/linked to semantic opacity.

4 Mutual expectancy – this can also be termed lexicality, and refers to the fact that the component words of an idiom co-occur in a fixed manner, giving the idiom a unitary form to accompany its holistic meaning or function. It is, in essence, operating as a single lexical item.

5 Lexicogrammatical invariability/frozenness/fixedness – the component words in an idiom are fixed and cannot be substituted by synonyms. In fact, some idioms do not even allow syntactic or morphological variation. Examples include hook line and sinker or beat around the bush; we cannot, for example, pluralize any of the nouns in these sequences, nor, for example, passivize the voice to render another appropriate idiom such as the bush is beaten around.

Metaphors

A metaphor is a sequence based on an unconventional reference in which words are used to describe something ordinarily far from its normal scope of denotation, producing a discord between a literal interpretation and a metaphoric interpretation. Metaphors have a common structure: the vehicle is the term used in an unusual manner, and the topic is the referent of the vehicle. The shared semantic content between the vehicle and topic are the grounds. An example might be time is a healer, in which healer is the vehicle, used in an unconventional sense, and time is the topic. In this case the grounds is the view of time acting like a physical remedy or medical practitioner, healing spiritual or emotional injuries and wounds in the same way as physical illnesses and injuries are healed by medicine or nurses. A metaphor can also be a simile, using is like or kind of. An example would be life is like a box of chocolates. The power of metaphor is linked to the semantic distance between vehicle and topic, and the relative explicitness of the vehicle.

Proverbs

Proverbs are generally sentence-length sequences which display an opaque relationship between literal and figurative meanings. Proverbs provide advice and warning (a stitch in time saves nine), instruction and explaining (early to bed and early to rise), and communicate common experience and observations (like death and taxes). They are taken from a store of proverbs shared by a cultural group or community. They are generally brief, direct, have simple syntax, contain elements of metaphor and sometimes dated or archaic structure or words.

Compounds

A compound is a sequence of two words (see ten Hacken, 2004), the second of which usually functions as the head or core of the compound – for example desk computer describes a type of computer and computer desk describes a type of desk (see Williams, 1981). The
head represents a type and the nonhead serves to classify the head. There are three forms of compounds:

1. **Closed form**, in which the words are written as one, such as *hardcore* or *laptop*
2. **Hyphenated form**, in which the lexical items are separated by hyphens, such as *brother-in-law* or *open-handed*
3. **Open form**, in which the two words are written separately, such as *grocery store* or *real estate*

Compounds are sometimes written as single words if the unit is strongly lexicalized. The words may be linked by a hyphen and over time become blended orthographically into a single word. Words modified by adjectives, for example, *a metal table*, are different from a compound word, for example, *a coffee table*, in the degree to which the non-headword changes the essential character of the head, or the degree to which the modifier and the noun are inseparable. In the example of *coffee table*, the compound represents a single entity, a particular type of table which is always identified in the same way, whereas the *metal table* is simply a table being described by means of the material from which it is made. The adjective slot in the sequence can be filled by any number of choices.

**Phrasal Verbs**

Phrasal verbs are distinguished largely by their distinctive structural makeup, lexical verbs combined with a preposition, particle, or both, with often nonliteral meanings, or both literal and figurative interpretations, like idioms. Three structural categories exist:

1. **Verb + preposition (prepositional phrasal verbs)**
   
   She quickly *picked up* some Portuguese on her Brazilian vacation.
   I happened to *bump into* my former boss on the street.

2. **Verb + particle (particle phrasal verbs)**
   
   You can *show that off* at the next party.
   I tried not to *cave in* under the stress of the divorce.

3. **Verb + particle + preposition (particle-prepositional phrasal verbs)**
   
   He is always *going on about* something or other.
   Jane *looks up to* her older brother.

According to Liu (2008, p. 22) there are three fundamental criteria for determining whether an item is a phrasal verb:

1. No adverb between the lexical verb and preposition or particle; for example, we cannot say *The kids loaded slowly up on chocolates before we got there*.
2. The particle cannot be at the front of a sentence; for example, we cannot say *Up with I am not putting any more outbursts*.
3. It cannot exist as only literal in meaning, but needs to have a figurative meaning, as in the preceding examples.
Classifying and Identifying Formulaic Language

Sequences Distinguished by Pragmatic Utility

Lexical Phrases

Lexical phrases are a pragmatically specialized subset of formulaic sequences first described by Nattinger and DeCarrico (1992). The phrases fall into two structural categories: strings of specific lexical items, mostly unitary and grammatically canonical, and generalized frames, category symbols and specific lexical items. The phrases display four characteristics: length and grammatical status; canonical or noncanonical shape; variability or fixedness; and discontinuousness or continuousness, the latter allowing lexical insertions (Nattinger and DeCarrico, 1992, pp. 37, 38). There are four broad categories of lexical phrases: poly-words, which function as single words, without variability or lexical insertions (e.g., “for the most part”, “so far so good”); institutionalized expressions, sentence-length, invariable, and usually continuous (e.g., “a watched pot never boils”, “nice meeting you”, “long time no see”); phrasal constraints, which have variations of lexical and phrase categories, and are generally continuous (e.g., “a ___ ago”, “the ___er the ___er”); and sentence builders, which contribute to the construction of full sentences with fillable slots (e.g., “I think that X”, “not only X but Y”) (pp. 38–45). It is clear that this particular taxonomy exhibits considerable overlap with other categories of formulaic language, such as proverbs, idioms, and collocations. The distinguishing feature of lexical phrases is that Nattinger and DeCarrico used pragmatic function as their common characteristic.

Pragmatic Formulas

In pragmatics, *formula* is used to refer to formulaic sequences employed for specific pragmatic purposes (Bardovi-Harlig, 2012). Various terms have been used for the types of sequences which have pragmatic functions in spoken interaction, including conversational routines, pragmatic idioms, speech formulas, routine formulas, situation formulas, and situation-bound utterances. These formulas are pragmalinguistic resources in spoken language and serve as the most socially appropriate means of accomplishing particular pragmatic functions. These include greetings (*how are you, what’s going on, how are things*) and turn-taking (*let me add . . . , I also think . . . , not only that, but . . . *).

Sequences Distinguished by Their Distribution in Corpora

Lexical Bundles

Lexical bundles (see Biber, Johansson, Leech, Conrad, & Finegan, 1999) are formulaic sequences distinguished by the procedures by which they are identified in corpora, and the fact that they are linked purely to functions in discourse, and are not meaning units. Lexical bundles may be considered more a type of multiword unit than strictly formulaic sequences, since there is no indication in the literature that they are stored or retrieved as wholes. They are researched using particular methods which focus exclusively on frequency and function. Research on lexical bundles tends to overwhelmingly focus on academic language, particularly written text.

Lexical bundles can be briefly defined as “combinations of three or more words which are identified in a corpus of natural language by means of corpus analysis software programs” (Wood, 2015, p. 45). Lexical bundles appear in a range of texts in a corpus. They have been
shown to be essential to the construction of academic writing, with particular bundles used more in specific disciplines (Cortes, Jones, & Stoller, 2002).

The pioneer of lexical bundles research is Biber (2006), who discovered that academic disciplines use specific lexical bundles, and created a categorization of functions of bundles – referential bundles – which refer to real or abstract entities or to textual content or attributes, for example, “the size of the . . .”, “one of the things”; stance bundles, which express attitudes or assessments of certainty, for example, “. . . are likely to be . . .”, “what do you think . . .”; discourse organizers, which indicate connections between previous and subsequent discourse, for example, “on the other hand”, “as well as . . .”.

Concgrams

A concgram is a combination of two or more words, but is distinctive in that it is a noncontinuous sequence, with the constituent words separated by others. The COBUILD team at the University of Birmingham in the 1980s were the first to use computer software to search corpora for noncontiguous word sequences. According to Sinclair (2005), it is likely that researchers will similarly uncover new patterns of word sequences, with “intercollocatability” and “interparaphrasability”, already entering the picture.

Clearly there is a surprising range and scope of types of formulaic language. The phenomenon is not a unitary construct, and classifications overlap and require considerable interpretation. Looking at a sequence and determining whether it is, for example, a collocation or an idiom, one may experience quite a bit of indecision. Some items may appear to fit with no particular category, for example, sequences like and then or sooner or later seem to defy labeling. Thanks to corpus analysis technology and techniques we have discovered new types of formulaic sequences.

In any case, formulaic sequences fall into various categories based on their features or usage. The descriptions of categories have evolved over time, and the classifications are somewhat fluid, with plenty of overlap and outliers.

Identification

It is useful to have a definition of formulaic language, and a sense of what the major categories are, because this takes us one step toward being able to handle it in research and in education. However, sooner or later any researcher or educator will come up against an obstacle which is tough to deal with: how can one go about identifying formulaic sequences in texts, spoken or written?

To understand the challenge of identification, examine the first two sentences in this section, and attempt to identify the formulaic sequences. Issues crop up immediately. Several multiword sequences stand out as more or less idiomatic, for example, one step toward, handle it, tough to deal with, sooner or later. How confident can one be with these decisions, and what features of the sequences lead us to decide they are formulaic or idiomatic? More importantly, what other elements of the sentences are formulaic but are not readily accessible to our intuitions and perceptions? How can they be uncovered? This question has been a preoccupation in the study of formulaic language. Perhaps one might decide that some sequences are more formulaic than others, but even then, what can guide the decisions? How frequently they are used in a given register? Prosodic features of the production of the sequence? Their frequency in a large corpus?
Fortunately, there are a number of reliable and well-developed means of identifying formulaic sequences. Some are more reliable than others, and they lend themselves to particular purposes and texts.

**Frequency and Statistical Measures**

It is axiomatic that particular formulaic sequences are generally recurring in a particular register, and a word sequence which sees frequent use is probably formulaic, provided it is also a more or less unitary meaning or function unit. Maybe they will also be mentally stored and retrieved as a single unit.

A distributional or frequency-based approach to identifying formulaic sequences is quite common in research (Durrant & Mathews-Aydinli, 2011). Statistical identification of formulaic language is a very productive method. In this type of identification procedure a set of parameters is established, marked by minimum lengths of sequences, minimum criteria for frequency, usually expressed as occurrences per million words in a corpus. The corpus is scanned for word combinations that fit within the parameters. Frequency cutoffs can range from 10 to 40 occurrences per million words (Biber et al., 1999; Simpson-Vlach & Ellis, 2010). The sequences which are uncovered by this type of corpus-based, statistically driven procedure are often not complete structural units (Cortes, 2004), and the majority of research of this type has uncovered units labeled *lexical bundles* (e.g., Biber et al., 1999) or *multiword constructions* (Liu, 2012; Wood & Appel, 2014). In some cases, researchers have used these parameters as part of a more elaborate process of identification, and have employed the term *formulaic sequences* (e.g., Simpson-Vlach & Ellis, 2010).

Logically, this statistical approach is used mostly with large corpora of hundreds of thousands, if not millions, of words. These corpora are created to be representative of specific registers of language. A purely frequency-based approach has some serious limitations if used with smaller data sets, particularly because minimum frequency cutoffs may be very difficult to set. For example, in a corpus of a million words, a frequency cutoff of 40 occurrences per million words would mean a given sequence would need to occur 40 times in order to be considered formulaic or a lexical bundle. But if a corpus has, say, 100,000 words, the sequence would only need to occur four times. This makes it challenging to identify sequences as formulaic using frequency-based methods alone when working with small corpora. A further limitation of using frequency alone as a criterion for formulaicity is that additional steps are also required to eliminate meaningless combinations of words – some content-specific word combinations can easily pass the frequency test and yet not be formulaic, for example, proper names such as *Ford Motor Company* or verbal tics or nonlexical fillers such as so, so, so. . . . Furthermore, it would be difficult to rely wholly on frequency for identifying formulaicity transcribed conversations on a range of topics. Many sequences which are formulaic might appear only once or twice in such a diverse and small set of data.

A very important drawback of using frequency-based analysis is the fact that frequency gives us absolutely no real indication of the holistic processing, which is often an important concern in some types of research – there is no way to rely on frequency measures to identify formulaic sequences in individual idiolects unless extra means of analysis are used. A good example of this is evident in a study by Schmitt, Grandage, and Adolphs (2004) in which formulaic sequences were identified in a corpus by means of statistical measures. The sequences were then integrated into spoken dictation tasks in which dictated texts were designed to overload the participants’ short-term memory capacities. Evidence of holistic
storage of the sequences in the participants’ reconstructions of the dictations varied greatly from participant to participant (Schmitt, Grandage, and Adolphs, 2004).

Some word sequences which appear to be formulaic in terms of saliency and unitary meaning or function may not actually appear at particularly high frequency in a given corpus. For example, in spite of or how are you may not be frequent in any particular corpus or genre, but we would probably agree that they show formulaicity, because they contain words which very commonly occur in this order, and they have a particular unitary meaning or function. The probability that words will co-occur like this can be measured statistically using measures of association such as Mutual Information (MI). MI is a measure of how likely a given set of words are to occur together in a set sequence in comparison to chance. MI does not have a particular statistical significance cutoff, but many researchers have used an MI of 3.0 or higher as an indication of statistical strength of co-occurrence (e.g., Church and Hanks, 1990; Hunston, 2002). A higher MI represents a stronger probability of co-occurrence, and is quite a strong objective measure of formulaicity, if used in combination with other frequency measures. Other similar measures of strength of association of words are used in corpus linguistics. For example, Gries (2008, 2012) uses a measure called the Fisher-Yates exact probability test in examining the strength of the relationships between a given word and a construction in which it occurs. Some studies combine data of various types such as corpus measures of association, eye tracking, and response latency. These types of measures are generally called psycholinguistic measures.

When frequency measures are not feasible for analysis of a small or very specific corpus, it is possible to look at a large general corpus, such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA). These huge sets of language data can provide a wealth of information about particular word combinations in real-life language use. An example of this is in a study by Wood and Namba (2013) in which they identified useful formulaic sequences to help individual Japanese university students improve their oral presentations. The researchers first identified useful sequences by means of native speaker/proficient speaker intuition, and then turned to the COCA to check their frequencies and statistical strength of co-occurrence. The researchers looked for the sequences in the spoken language subcorpus of the COCA, using a frequency cutoff of at least ten occurrences per million words and with a Mutual Information score (MI) of at least 3.0. In this way, they could be certain that the identified sequences were frequent in speech and that they consisted of words with a high frequency of co-occurrence. A novel means of determining formulaicity is the use of online search engines such as Google. Shei (2008) pointed out that there are really no readily available corpora which are large enough to give full coverage of language use for many types of investigation. Shei presents a strong case for researchers and educators to use the internet as an enormous corpus, readily exploited by means of a search engine such as Google to identify and retrieve word sequences for research and language teaching and learning support. It is a simple matter of Googling a given sequence and examining the resulting hits, which may contain extremely valuable information about its frequency, form, variability, and functions.

**Psycholinguistic Measures**

Studies in which identification of formulaic language is a focus have used measures of processing speed. These measures may include reaction times (e.g., Conklin & Schmitt, 2012), eye movement (e.g., Underwood, Schmitt, & Galpin, 2004), and electrophysiological (ERP) measures (e.g., Tremblay & Baayen, 2010).
Classifying and Identifying Formulaic Language

Measures such as eye tracking or response latencies require that participants read, making them unsuitable for research involving children or nonliterate individuals. It is obvious that these psycholinguistic measures are useful in identifying sequences psycholinguistically stored by any one individual, but they are much less helpful in showing how commonly used a formulaic sequence may be in a broader speech community. These measures may reveal formulaic sequences which are rare, unusual, or one-off, used idiosyncratically by a speaker.

Acoustic Analysis

A common criterion for identification of formulaic sequences in speech is phonological coherence, a term coined by Peters (1983). Phonological coherence is a characteristic of formulaic production in which a word sequence is uttered fluently, with no hesitations and an unbroken intonation Peters (1983, p. 8). Formulaic sequences exhibit certain prosodic characteristics, such as alignment with pauses and intonation units, resistance to internal dysfluency, no internal hesitations, fast speech rhythm, and stress placement restrictions (see Lin, 2010, 2012). It is important to bear in mind that phonological coherence characterizes formulaic sequences in a given individual’s idiolect, and that analysis of this type is restricted by the quantity of data which can be processed by an individual and the technological tools used to record and analyze speech data.

Criteria Checklists and Native-Speaker Intuition

Researchers sometimes discover that frequency, psycholinguistic processing, or acoustic analysis measures are insufficient to identify formulaic sequences in various types of data, especially spoken data. This is often resolved by means of criteria checklists that blend specific features associated with formulaicity.

A proponent of the use of such checklists has been Wray (2002), who reviewed methods of detecting formulaic sequences in many data types. She notes that use of corpus analysis computer software is one possible method of identification, but points out some shortcomings of reliance on frequency in particular:

It seems, on the surface, entirely reasonable to use computer searches to identify common strings of words, and to establish a certain frequency threshold as the criterion for calling a string “formulaic” . . . (however) problems regarding the procedures of frequency counts can be identified. Firstly, corpora are probably unable to capture the true distribution of certain kinds of formulaic sequences. . . . The second serious problem is that the tools used in corpus analysis are no more able to help decide where the boundaries between formulaic sequences fall than native speaker judges are.

(pp. 25, 27, 28)

It is obvious that small data sets composed of spoken discourse present challenges for computer corpus analysis software. For one thing, the discourse or topic-specific speech in such data sets, combined with the small total word count, make it very difficult to rely on frequency alone, since some sequences might occur only once or be used very idiosyncratically. It is also often the case that formulaic sequences blend into surrounding language; many also have large fillable slots, presenting a great challenge for corpus analysis software. Research involving second language learners often produces data with large numbers
of nonstandard or idiosyncratic sequences. In the end, a researcher can turn to one quite daunting measure in identifying formulaicity in language, what Wray terms “the application of common sense” (p. 28).

Native-Speaker Judgment

Fortunately, it is readily possible to apply common sense to the task of identifying formulaic sequences, especially that of second language speakers, by examining language performance and comparing it to native-speaker use of formulaic sequences. This involves the use of native-speaker judgment and a checklist of criteria. Wray (2002, p. 23) points out five challenges inherent in this type of procedure:

1. It has to be restricted to smaller data sets.
2. Inconsistent judgment may occur due to fatigue or alterations in judgment thresholds over time.
3. There may be variation between judges.
4. There may not be a single answer as to what to search for.
5. Application of intuition in such a way may occur at the expense of knowledge we do not have at the surface level of awareness.

Recall how challenging it was at the beginning of this chapter to isolate formulaic sequences without any guiding criteria. This challenge can be at least partially overcome by use of a checklist of specific criteria. The standard procedure for this involves judges studying the criteria which inform a checklist, and examining a corpus to apply the criteria and identify sequences which appear to be formulaic. A high degree of interrater reliability among judgments is a good general measure of the strength of a given judgment.

A number of checklists have been used in such research; some checklists developed for specific populations, others more general. Following are descriptions of three such checklists which have been used in various studies: an early checklist elaborated by Coulmas (1979); a checklist applicable to a range of child and adult native or non-native speakers (Wray & Namba, 2003); a checklist used to identify formulaicity in second language acquisition of speech fluency (Wood, 2006, 2009, 2010).

Coulmas, 1979

Coulmas (1979, p. 32) lays out nine specific criteria for formulaicity:

1. At least two morphemes long (i.e., two words)
2. Coheres phonologically
3. Individual elements are not used concurrently in the same form separately or in other environments
4. Grammatically advanced compared to other language
5. Community-wide formula
6. Idiosyncratic chunk
7. Repeatedly used in the same form
8. Situational dependence
9. May be used inappropriately
Wray and Namba, 2003

Wray and Namba (2003) presented a very flexible and comprehensive checklist, originally used in a study of speech of bilingual children. The checklist is applicable to many types of data and consists of 11 criteria, rated on a Likert Scale of 1 to 5. This is quite refined in that it deals with the issue of gradience or ranges of formulaicity:

1. By my judgment, there is something grammatically unusual about this wordstring.
2. By my judgment, part or all of the wordstring lacks semantic transparency.
3. By my judgment, this wordstring is associated with a specific situation and/or register.
4. By my judgment, the wordstring as a whole performs a function in communication or discourse other than, or in addition to, conveying the meaning of the words themselves.
5. By my judgment, this precise formulation is the one most commonly used by this speaker/writer when conveying this idea.
6. By my judgment, the speaker/writer has accompanied this wordstring with an action, use of punctuation, or phonological pattern that gives it special status as a unit, and/or is repeating something s/he has just heard or read.
7. By my judgment, the speaker/writer, or someone else has marked this wordstring grammatically or lexically in a way that gives it special status as a unit.
8. By my judgment, based on direct evidence or my intuition, there is a greater than-chance-level probability that the speaker/writer will have encountered this precise formulation before, from other people.
9. By my judgment, although this wordstring is novel, it is a clear derivation, deliberate or otherwise, of something that can be demonstrated to be formulaic in its own right.
10. By my judgment, this wordstring is formulaic, but it has been unintentionally applied inappropriately.
11. By my judgment, this wordstring contains linguistic material that is too sophisticated, or not sophisticated enough, to match the speaker’s general grammatical and lexical competence.

Native-Speaker Judgment: Wood, 2010

Wood (2010) published a study examining the possible effect of use of formulaic language on speech fluency in second language learners of English. Identifying formulaic sequences was central to the methods used in the research.

Five criteria were integrated into the checklist. They were used as guides for native-speaker judges, and no one criterion or combination of criteria was required in order for a sequence to be labeled as formulaic.

1. **Phonological coherence and reduction.** Formulaic sequences may be produced with phonological coherence (Coulmas, 1979; Wray, 2002), lacking internal pausing and exhibiting a continuous intonation contour. Phonological reduction is also possible, involving phonological fusion, reduction of syllables, deletion of schwa, all of which are characteristic of the most frequent phrases in English (Bybee, 2002).

2. **The taxonomy used by Nattinger and DeCarrio (1992).** This taxonomy was described above, and is not required in every case, but as a potential guide to determining formulaicity. If a sequence matched a category in the taxonomy it might be flagged as formulaic.
3 **Greater length/complexity than other output.** This is a typical feature of formulaic language in second language speech. For example, a participant might say *I would like...* or *I don’t understand*, using these structures only in these particular sequences and never successfully using *would* or negatives using *do* in other contexts.

4 **Semantic irregularity, as in idioms and metaphors.** Sequences which meet this criterion are usually standard phrases and expressions.

5 **Syntactic irregularity.** Formulaic sequences can be syntactically irregular, especially idioms.

In Wood’s (2010) study, non-native-like sequences were accepted as formulaic. A sequence could have been misperceived and stored as an idiosyncratic string, as in *what’s happened* instead of *what happened*, or *thanks god* instead of *thank god*. The need to produce second language speech under the pressure of recording also might have caused utterances to contain irregular features, articulatory slips, or gaps and inaccuracies. Therefore, a sequence could match the checklist criteria and still be idiosyncratic, misperceived, stored with errors, or misarticulated.

The three checklists described here are organized so that none of the criteria are necessary; nor must all be met in order for a sequence to be labeled formulaic. Wray and Namba (2003) use a 5-point Likert scale for each criterion, from *strongly agree* to *don’t know* to *strongly disagree*. All three checklists are quite different from the distributional, psycholinguistic, and acoustic measures described earlier, since they place primary importance on informed judgments based on a range of criteria. The checklists show considerable agreement on the characteristics of formulaicity, all making reference to phonological characteristics and complexity.

**Future Directions**

Formulaic language is certainly not a unitary or monolithic construct. The categories overlap, display imprecision, and are subject to considerable interpretation. Judging whether a sequence is a collocation or an idiom is tricky, despite the fact that a number of researchers have attempted to assign specific characteristics to particular types. Many multiword items defy strict or ready classification; for example, sequences like *and then* or *sooner or later* are difficult to categorize. The advent of corpus analysis technology and methods have enabled the discovery of new types of formulaic sequences. It is difficult to see the importance of the distinction between, for example, identifying a sequence identified using frequency alone and identifying one using a combination of frequency and other statistical measures such as Mutual Information.

In general, it has been established that one can determine possible formulaicity in a number of ways. One way is by using frequency statistics, applying criteria to a specific corpus, or checking the frequency or mutual information of individual items in very large corpora such as the BNC, COCA, or even internet search engines. Formulaicity may be determined in whole or in part by psycholinguistic or acoustical features of a sequence and its processing. Expert or native-speaker judgment about formulaicity is a good way to gauge formulaicity with spoken data or smaller or quite specific data sets. In this type of procedure, a checklist of characteristics of formulaicity provides a guide for judges. Formulaic language can be challenging to identify, and it may be best to employ various means or a combination of measures, and even then absolute certainty is likely to be elusive. Whether one uses corpus frequency, statistical measures of co-occurrence acoustical features, or judges and
checklists, decisions are likely to be expressed with some hedging. Presumably future developments in research will yield more exact means of determining formulaicity.

**Further Reading**


This volume represents the only existing overview of formulaic language for researchers and students. It contains chapters which summarize current knowledge on a full range of topics, including categorization, identification, mental processing, corpus-based research, spoken and written language, and teaching.


This is a classic work on formulaic language in which Wray lays a strong theoretical and research foundation for studies of formulaic language. It includes essential perspectives on processing, acquisition, and research methods.


This volume represents one of the first compilations of original research on formulaic language from a range of perspectives. Many of the studies are excellent models of research in corpus and psycholinguistic explorations of formulaic language.

**Related Topics**

Phraseology, formulaic language in language teaching, lexical approach, lexical bundles, construction grammar, multiword units

**References**

Allerton, D. J. (1984). Three (or four) levels of word cooccurrence restriction. *Lingua, 63*(1), 17–40.


