THIS IS THE SUBMITTED VERSION OF Buerki, Andreas (2025) **Corpus Analysis of Phraseology** in Chapelle, C. A. (Ed.) *The Encyclopedia of AppliedLinguistics, 2nd ed.* Wiley.
THE PUBLISHED VERSION IS AVAILABLE AT
HTTPS://DOI.ORG/10.1002/9781405198431.WBEAL20509

Corpus Analysis of Phraseology

Andreas Buerki
Cardiff University
buerkiA@cardiff.ac.uk

Abstract

Phraseology designates both the study of usual turns of phrase in language and also the turns of phrase themselves (i.e. phraseological expressions). Traditionally, phraseology focussed on highly idiomatic, non-compositional expressions like idioms (e.g. *pull someone's leg*), but in recent decades and under the influence of the availability of corpora and increasingly sophisticated tools to interrogate them, the phraseological domain has expanded vastly; less clearly idiomatic expressions like collocations (e.g. *brush one's teeth*) and other usual sequences (e.g. *in other words*, *to be honest*), formulae (e.g. *Good morning*) and similar patterns are now considered part of phraseology. The corpus linguistic analysis of phraseological expressions has proven exceptionally fruitful in discovering expressions, understanding their prevalence, the range of permissible variation and links between phraseology and linguistic proficiency and genre among other aspects. Conversely, a phraseological corpus analysis can make important contributions to the understanding of text, discourse and practically any area of language and life.

Keywords

Corpus Linguistics, Phraseology, Formulaic Language

[A]Phraseology

The study of phraseological expressions has a long and rich history, with some of the earliest evidence of the scholarly lexicographical study of phrases in Europe dating back to the early middle ages. The modern study of phraseology in the European tradition, and the term *phraseology* to describe it, is most often traced back to Charles Bally, Ferdinand de Saussure's successor at the University of Geneva, and his volume on stylistics (Bally, 1909). Bally's work was most widely received and further developed by Soviet and Eastern Bloc scholars, whose work in turn became foundational to the study of phraseology across Europe and beyond. Present work in phraseology is

influenced by a diverse range of traditions of enquiry, forming a rich tapestry of perspectives, approaches, preoccupations, understandings and applications.

Accordingly, there exists today a plurality of understandings with regard to the linguistic phenomena that should or could be included under the umbrella of *phraseology*, as well as a range of views on the essential nature of such expressions and the terminological machinery used to describe different types of phraseological and related expressions. In general, theorists most often define the phraseological by recourse to elements shown in table 1. The element relating to form is invariably included, but typically either an aspect of use or an aspect of processing is then chosen to complement it; occasionally all three elements are found in definitions.

Table 1. Definitional Elements

Elements	**Typical descriptions**
Form	Combination of two or more words; sequences of words or other elements; phrases
Use	Conventional, usual; (typically) non-compositional or irregular; increased fixedness
Processing	Pre-fabricated; (appearing to be) stored and retrieved whole; reproduced

The element of form usually stipulates polylexicality (i.e. a phraseological expression must consist of more than one word), and hence an expression like *greenhouse* would be excluded, but *bus stop* or *as is* may be included, depending on how they fare with additional criteria. However, a number of theorists have argued that the concept of a word may not be stable enough to bear such heavy definitional weight and combinations of morphemes should also be admitted to the phraseological domain in certain cases. Such cases may include expressions in polysynthetic languages where a word may translate to a whole phrase in a more isolating language such as English.

Definitional elements describing use typically stipulate that phraseological expressions manifest usual turns of phrase, conventional ways of putting things in a speech community or phrasings preferred over available alternative ways of putting things. Thus, *making a mistake* is a conventional and usual turn of phrase, but *committing* or even *doing a mistake* less so. Particularly in corpus-

linguistic analysis, conventionality is often measured via relative frequency of occurrence, where frequency might be considered relative not only to corpus size, but also to alternative ways of expressing the same or similar sentiments (e.g. message to message-expression ratio; Wray, 2002, p. 31), or relative to pairings, for example of the word *mistake* with verbs other than *make*. Conventionality is also observed to bring a certain fixedness; in the example, the usual verbs to use with *mistake* are very limited, though the form of the verb, whether it be the gerund, present or past tense, for example, remains flexible. Fixedness also prevents the use of synonyms in idioms like *to jump ship* ('to join a rival group') which would fail to yield the same semantics if either *jump* or *ship* were substituted with synonyms like *skip* or *boat*. For this reason, an amount of fixedness is a further criterion often employed in definitions of phraseology that incorporate the element of language use. A final aspect of 'use' is a distinction that is often made between phraseolgical expressions that, like *to jump ship*, are idiomatic (also: non-compositional) and those that are non-idiomatic (compositional), like *make a mistake*. Although in these two cases, the distinction is clear, in many cases it is less clear-cut and may depend on questionable conceptions of the literal or de-contextualised meanings of single words, as Fleischer (1982, p. 38) pointed out. Nevertheless, it is a distinction often made and traditionally, the focus of phraseological research has been on idiomatic expressions, while non-idiomatic expressions were sometimes only considered phraseological 'in a broad sense' (Burger, 2007, p. 11). This sharp focus on idiomatic expressions has receded in more recent research. Non-idiomatic phraseological expressions are sometimes summarily referred to as *collocations*, and idiomatic ones as *idioms*, though both these terms are used with a range of different meanings in the literature.

The final definitional element relates to the manner of mental processing: an expression is seen as phraseological if its mental processing (in production and reception) is largely holistic. There is currently no consensus on what holistic processing might mean in detail, and mental processing is notoriously difficult to measure. For this reason, theorists using holistic processing as a definitional criterion instead employ arguments that plausibly suggest holistic processing for expressions in question rather than attempting to prove a particular manner of processing directly. For example,

non-compositional meaning (e.g. *spill the beans* 'reveal a secret') or form (e.g. *by and large* 'generally') suggest a unitary representation at some level, and evidence from psycholinguistic experiments on a small number of tested expressions suggest faster or different processing of usual word combinations vis à vis less usual combinations, regardless of compositionality. Judgements on whether expressions are 'or appear to be' (Wray, 2002, p. 9) holistically processed are then made in analogy to such cases.

While all definitional elements of table 1 are compatible with each other, the typical emphasis on either 'use' or 'processing' in a definition, alongside a 'form' element, is because use and processing place the focus on different aspects: matters of use, emphasising conventionality, usualness and/or shared meanings, are anchored at the level of a speech community, whereas mental processing happens in the heads individuals (and may differ between individuals processing the same expressions) and so is primarily anchored at the level of an individual. While it is possible to extrapolate and generalise across individuals' processing habits, and conversely investigate expressions that are usual or typical in an individual's language use, the emphases are clearly different. This interfaces with corpus analysis in that corpora (unless they are specialised corpora of the works of individuals) represent not the linguistic habits of individuals, but of speech communities and therefore interface more naturally with definitions that incorporate a 'use' element. Expressions falling under the definitional elements given above are routinely referred to as *phraseology* or *phraseological expressions*, but a range of other designations are also used including *Formulaic Language, Multi-Word Expressions* and other terms, as well as labels like *idiom, collocation, routine formula,* etc. for more specific sub-types.

[A]Corpus linguistic analysis and the study of phraseology

Corpus linguistics, the study of language through collections of machine-readable texts, and the field of phraseology have been significant influences on each other since the widening availability of sizeable corpora from the late 1990s onwards. Access to such corpus materials and increasingly sophisticated computational tools to interrogate them influenced three significant developments in

the study phraseological patterns in language: first, although the vastness of compositional, non-idiomatic phraseological expressions was previously known, the realisation took hold that those expressions could now be studied much more easily, that they were much more interesting than had perhaps been thought and compared to expressions with clear idiomaticity, they were extremely frequent in language use. This realisation led to a significant widening of phraseological research interests to include those less clearly idiomatic, but still highly usual patterns on par with the highly idiomatic phrases that had traditionally been the focus of research. Connected to this, it became possible to demonstrate that language as actually used and reflected in linguistic corpora is to an astonishingly large degree phraseological in nature, where previously it was still possible (if imprudent) to claim that 'virtually every sentence that a person utters or understands is a brand new combination of words, appearing for the first time in the history of the universe' (Pinker, 1994, p. 22). One of the earliest studies to show the inaccuracy of such statements was Altenberg's (1998) analysis of recurrent word combinations in the London-Lund spoken corpus. The third development connected to corpus linguistics was that phraseological patterns being noticeable in corpus analyses led to a remarkable increase in phraseological research being conducted and thus to research into phraseological questions becoming part of mainstream linguistic analysis, where it had previously been a more niche field of research globally. Conversely, the influence of phraseological concepts on corpus linguistic analysis in general is evident in some of the most central theoretical and methodological tenets of corpus linguistics: staples of corpus linguistic analysis like finding collocations of head words and deriving n-gram lists are informed and underpinned by phraseological theory.

The profound mutual influence of corpus linguistics and phraseology on each other has made possible advances in linguistic theory but has also enabled progress in key domains of applied linguistics, including lexicography (making possible corpus-informed dictionaries of idioms, phrasal verbs and collocations) and language pedagogy (e.g. via phraseologically aware, corpus-informed language learning materials), among many other domains.

[A]Current approaches to corpus analysis of phraseology

The main ways in which corpora are currently used to investigate phraseological patterns reach from methods to establish collocates of seed words and the strengths of their association, to the data-led discovery of phraseological expressions in corpus materials, to gaining a deeper understanding of the use of known expressions, including their prevalence, the variations and limits on variation that they show in language use and the functional, stylistic and other uses to which phraseological expressions and patterns are put.

[B]Discovery

When it comes to corpus linguistic methods employed in the data-led discovery of phraseological expressions, there are two basic types of approaches — seed-word based and n-gram based. An approach based on one or more seed words proceeds with the aim of discovering phraseological patterns associated with the seed word(s) in corpus materials by analysing words in proximity of where the seed word (in this context referred to as *node*) occurs. Typically, this involves automatic searches of left and right contexts in a span of 'a few words' (Sinclair, 1991, p. 170) either side of the node. The results of such searches are lists of candidate *collocates* that are subsequently filtered to obtain a list of bona fide collocates of the node word. Typical filters applied are the frequency of occurrence of a candidate collocate together with the node, or one of a range of statistical association measures, including MI-scores, t-scores, log likelihood and others that seek to measure the degree to which the co-occurrence of the node and collocate exceeds the level of chance co-occurrence, given the frequency of the individual words involved. Candidate collocates are then ordered in decreasing order of association and a cut-off score defined, above which co-occurrences are deemed of interest. This type of approach consequently produces expressions that are two-word combinations. For example, among collocates within a +/-3 word context of the word *ill* in the enTenTen21 corpus (Jakubíček, Kilgarriff, Kovář, Rychlý, & Suchomel, 2013), we find the words listed in (1).

(1) a. chronically d. patient(s) g. equipped j. effects

b. critically	e. cancer	h. bode(s)	k. repute
c. fall	f. cure	i. afford	l. societal

Depending on how the term *collocation* is defined, all of these can be considered as forming collocations with *ill*, but a closer phraseological analysis reveals a range of distinct types of expressions. These include collocations in the sense of Hausmann (1991), that is, structured co-selection tendencies of a base (in this case *ill*) with syntagmatically related collocates forming structural patters like ADV+ADJ (1a,b) or V+ADJ (1c) that are semantically transparent. Conversely, (1)d–e show conceptual associations that would not usually be classed phraseological at all, whereas (1)g–i show expressions (or fragments thereof) that are idiomatic in character (*can ill afford to do something* or *to bode ill* are not straightforwardly compositional expressions). (1)j–l are of the nature of multi-word terms denoting a single concept in a less than fully compositional manner (e.g. *a societal ill*).

Traditional measures of association, such as those mentioned above, neither take into account that words are not in fact expected to occur in random distribution even in the absence of notable associations between two words, nor is directionality taken into account (i.e. whether the association of the two words is one-sided or bidirectional). According to how the association between words is operationalised mathematically, different measures also lead to different expressions scoring high (the MI-score, for example, tends to award high scores to rare events and exclusive associations, whereas other measures require a lot of evidence of co-occurrence to award higher scores). Unidirectional measures have more recently been suggested; simple conditional probabilities, for example, look at how likely one word occurs, given the occurrence of the other close by (i.e. the proportion of the occurrences of one word that are in the vicinity of the other), showing in the case of (1)a, that 21% of occurrences of *chronically* are with *ill* in the enTenTen21, while only just over 1% of occurrences of *ill* are with *chronically* revealing a much stronger association of *chronically* with *ill* than vice versa. Typical corpus linguistic software applications and web services offer a range of tools and association measures to facilitate seed word based discovery methods. More

comprehensive extractions of phraseological patterns can be achieved by using large quantities of seed words (e.g. a large proportion of the content words of a language), but for comprehensive extractions of phraseological expressions, n-gram based approaches tend to be favoured.

An n-gram based approach to the discovery of expressions proceeds by splitting a text into partially overlapping word sequences of a chosen length, or a range of different lengths from two words (bigrams) to sequences as long as eight or more words (8-grams, etc.). As with candidate collocates, lists of such word sequences are subsequently filtered using various methods to identify phraseological expressions. Among the most common of those are frequency (e.g. used in the identification of *Lexical Bundles*, Biber, Johansson, Leech, Conrad, & Finegan, 1999, chapter 13) and stop lists that eliminate sequences composed of certain words, but association measures (as above) can also be used where they have been extended to become applicable to sequences longer than bigrams. More recent approaches have moved away from assumptions of the traditional association measures, however. Colson's corpus proximity ratio, for example, measures the fixedness of patterns by employing the ratio of the frequency of occurrence of an n-gram in its contiguous form, divided by the total number of times its constituents occur (including in at least partially interrupted form) within a window of each other (Colson, 2017).

N-gram based approaches are especially well-suited to the identification of usual sequences and turns of phrase that are fixed at the level of word form, such as those shown in (2)a–b.

(2)

a. in large part	e. with the [further/occasional/sole/notable] exception of
b. never mind	f. draw [heavily] on
c. for the first time in N years	g. [it was] raining cats and dogs
d. MAKE a name for RPRON	h. cats and dogs raining [from the sky]

There are techniques for accommodating morphological variation through hybrid n-grams that incorporate lemmas, parts of speech or other schematic elements as would be useful in the

identification of expressions like (2)c–d where *N* indicates a number, *MAKE* a lemma and *RPRON* (reflexive pronoun) a part of speech. In languages with rich inflectional morphology, this can be key to a successful identification of expressions. N-grams that skip positions allowing for intervening material (shown in square brackets) in (2)e–f are also employed. Positional variation as in (2)g–h remains a challenge that requires further methodological development and often represents a particular challenge in languages with a less rigid word order. Many corpus linguistic tools now offer n-gram based functionality and specialist software is also available. N-gram based discovery of phraseological expressions in short texts is possible using tools that access databases abstracted from large corpora to enable identification.

Alternative approaches based on transitional probabilities (points in a text where there is a high level of uncertainty regarding what word(s) might follow, cf. Gries & Mukherjee, 2010) are also used and increasingly, machine learning and AI-based approaches will play a role in the discovery and identification of phraseological expressions. Corpus linguistic discovery techniques have contributed very significantly to the theoretical understanding of phraseology and have served particularly to highlight the vast number of usual turns of phrase in language that may be only weakly idiomatic or fully compositional.

[B]Understanding use and prevalence

Apart from the discovery of new phraseological expressions and the comprehensive identification, extraction and comparison of expressions contained in different corpus materials, very significant amounts of corpus linguistic research have also been expended on building a better understanding of the use and prevalence of known phraseological expressions in authentic texts. Such known expressions might be sourced from a preceding discovery procedure or from lexicographical or other sources. The principal tools of investigation have been concordance lines (also keywords, or here key phrases, in context; KWICs). However, variation and flexibility in known phraseological expressions (as exemplified in the attested forms in (2)f–g and well as (3)a–c) have meant that

carefully crafted queries, using specialist query languages appropriate to the relevant corpora and their user interfaces, often need to be employed.

(3) a. It was raining like cats and dogs.

b. It came down like proverbial cats and dogs.

c. Despite the cats and dogs rain, ...

A major focus of research in this area has included investigations into idiom variation (e.g. Langlotz, 2006), including the exploration of the boundary between systematic variation and creative exploitation of idiomatic expressions as well as limits of and underlying principles behind such variation. Another focus has been on how idioms are typically used and apart from an appreciation for the flexibility within overall fixedness, main findings have included that idioms and proverbs are often not so much used but alluded to, or conversely used with various types of distancing devices such as the insertion of the adjective *proverbial* to modify nominal elements of idiomatic expressions (as in (3)b) and the addition of phrases like *as they say* or *as it were* (e.g. Moon, 1998). The textual and stylistic functions of phraseological expressions have similarly been studied, revealing, for example, concentrations of idiomatic expressions in headlines, titles and advertising or the differences in register associated with the use of phrasal verbs vs. non-phrasal verb equivalents. Lastly, corpus linguistic studies of usual sequences are able to document in detail not only the semantics of common turns of phrase and their status as conventional units, but pragmatic and other additional situational and cultural information that attaches to them.

Applications of corpus linguistic analyses of phraseology employing combinations of the above approaches have further contributed to knowledge in key areas outside their own domains. Examples of such areas include the developmental trajectories of L2 learning using learner corpora (e.g. Siyanova-Chanturia and Spina, 2020), the exploration of social and cultural change reflected in linguistic change (e.g. Buerki, 2020), evaluative language (Hunston, 2011) as well as linguistic

typology, discourse analysis and text linguistics, the role of phraseology in genre conventions and more, all illustrating the spectrum of applications to which corpus analysis of phraseology is put.

[A]Conclusions

The corpus linguistic analysis of phraseology has had a profound effect on corpus linguistics and the field of phraseology, leading in tandem to major advances in the understanding of language. While methodological challenges remain and theoretical stances are diverse and contested in this intriguing area of linguistics, the number of works that keep appearing shows that it is an accessible area of study that is welcoming of new contributions. The range of applications to which corpus analyses of phraseology have been put remains wide-ranging and diverse — any area of language and life will have phraseological aspects that can be corpus linguistically explored.

Cross-References

Formulaic Language and Collocation; Formulaic Sequences; Corpus Linguistics; (wbeal0235;wbeal0433;wbeal1402;wbeal20240;wbeal0033;wbeal0434.pub2)

References

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: Oxford University Press. DOI: 10.1093/oso/9780198294252.003.0005

Bally, C. (1909). *Traité de stylistique française*. Paris: Librairie C. Klincksieck.

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

Buerki, A. (2020). *Formulaic Language and Linguistic Change*. Cambridge: Cambridge University Press. DOI: 10.1017/9781108769976

Burger, H., Dobrovol'skij, D., Kühn, P. & Norrick, N. R. (2007). 'Phraseology: Subject area, terminology and research topics'. In H. Burger, D. Dobrovol'skij, P. Kühn & N. R. Norrick, (Eds.), *Phraseology: An international handbook of contemporary research* (pp. 11–19). Berlin: de Gruyter.

Colson, J. (2017). The IdiomSearch Experiment: Extracting phraseology from a probabilistic network of constructions. In R. Mitkov (Ed.) *Computational and corpus-based phraseology*. Cham: Springer. 16-28. DOI: 10.1007/978-3-319-69805-2_2

Fleischer, W. (1982). Phraseologie der deutschen Gegenwartssprache. Leipzig: Bibliographisches Institut.

Gries, S. & Mukherjee, J. (2010). Lexical gravity across varieties of English. *International Journal of Corpus Linguistics, 15*(4), 520–548. DOI: 10.1075/ijcl.15.4.04gri

Hausmann, F. J. (1991). 'Collocations in monolingual and bilingual English dictionaries'. In V. Ivir and D. Kalogjera (Eds.), *Languages in contact and contrast: Essays in contact linguistics* (pp. 225–36). Berlin: de Gruyter. DOI: 10.1515/9783110869118.225

Hunston, S. (2011). *Corpus approaches to evaluation: Phraseology and evaluative language*. London: Routledge.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In 7th International Corpus Linguistics Conference CL. 125–127. Available online at http://www.sketchengine.co.uk/wp-content/uploads/The_TenTen_Corpus_2013.pdf

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198236146.001.0001

Langlotz, A. (2006). *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. Amsterdam: John Benjamins. DOI: 10.1075/hcp.17

Pinker, S. (1994). *The language instinct*. London: Penguin.

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

Siyanova-Chanturia, A. & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning, 70*(2), 420–463. DOI: 10.1111/lang.12383

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511519772

Further Reading

Buerki, A. (2021). Reading discourses through their phraseology: The case of Brexit. In A. Trklja & Ł. Grabowski (Eds.), *Formulaic language: Theories and methods* (pp. 141–170). Berlin: Language Science Press. DOI: 10.5281/zenodo.4727671

Durrant, P. & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes, 30*(1), 58–72. DOI: 10.1016/j.esp.2010.05.002

Gisle, A. (2022). Phraseology in a cross-linguistic perspective: A diachronic and corpus-based account. *Corpus Linguistics and Linguistic Theory, 18*(2), 365–389. DOI: 10.1515/cllt-2019-0057

Mitkov, R. (Ed.) (2017). *Computational corpus-based phraseology*. Cham: Springer. DOI: 10.1007/978-3-319-69805-2

Contributor Bio:

Andreas Buerki is a Senior Lecturer at the Centre for Language and Communication Research at Cardiff Universtiy and vice-president of the European Society of Phraseology (EUROPHRAS). He is a founding member of the Cardiff Corpus Network of corpus linguists and author of a number of specialist corpus linguistic research software applications. He has published widely on phraseology and corpus linguistics, including the 2020 monograph *Formulaic Language and Linguistic Change* (CUP).