Technology and Phrases

GILL PHILIP

Like many linguistics terms, "phrase" also has a nontechnical use in the general language. We notice and comment on (especially unusual) "turns of phrase"; most of us at some time in our lives have asked for help in "how best to phrase" a request or a letter; and, when travelling, we often make use of foreign phrase books. Rarely does it occur to us to define what we mean by "a phrase," or to wonder why phrases are sometimes more useful to us than single words are. This entry explains what phrases are and provides an overview of the most common phrasal types. It also discusses some of the ways in which phrases can be identified and retrieved automatically and semi-automatically from computer language corpora.

What Is a Phrase?

According to Gries (2008), a "phrase" consists of two or more word forms which together express a semantically complete meaning. The words may undergo morphological variation (in most phrases), and usually occur in an established sequence. This definition is maximally inclusive. Phrases encompass a heterogeneous range of linguistic structures including idioms (*throw caution to the wind*), proverbs (*too many cooks spoil the broth*), conventionalized similes (*as dead as a dodo*), quotations (*to be or not to be*), sayings (*if you can't beat 'em, join 'em*), catch-phrases (*here's one I prepared earlier*), formulaic greetings (*how do you do*), phrasal verbs (*set up*, *look into*, *get away with*), and complex prepositions (*in order to*, *on behalf of*). Some, such as *by and large* or *for the most part*, are completely frozen, meaning that the word forms and their sequencing are always the same: No inflected forms or additional elements are allowed, and no inversion of the words is possible.

Frozen phrases are very much the exception to the rule. The *canonical form* of a phrase, normally used as its citation form in dictionaries and other reference works, is in fact only one of its possible realizations, albeit typically the most commonly occurring one. Phrases bend to the will of the textual environment in which they are used: verbs inflect for number and person, tense and aspect (though change to verb mood is much rarer); personal pronouns have to agree with their grammatical subjects and objects; nouns are pluralized and attract modifiers; prepositions are altered so that they reflect grammatical relations more appropriately. Phrasal verbs belong to this middle cut of phrasal types: Their verbal component has full inflectional capabilities (*give up*, *gives up*, *gave up*, *given up*, *giving up*), and the verb's direct object can interrupt the phrasal sequence (*take off*; *take it off*; *take your coat off*). Phrasal verbs aside, many phrases contain a variable slot, often for a direct object, as in *bear something in mind*; or a possessive pronoun, as in *shake in one's shoes*.

To complicate matters further, some phrases appear not to have a canonical form at all, and indeed may not even require any single lexical item to be present. Instead they consist of a lexicogrammatical configuration which combines aspects of syntax and semantics in a more abstract manner. One such phrasal type is the *semi-prepackaged phrase*, which can be illustrated with *the faintest idea*, and its variants (Francis, 1993, p. 144). *Faintest* can be replaced by any superlative adjective denoting indefiniteness (including *least*, *slightest*,

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.

DOI: 10.1002/9781405198431.wbeal1189

remotest, *foggiest*), while *idea* can be replaced by any noun belonging to the same semantic set (e.g., *conception*, *notion*); yet not one of these single words is essential for the meaning to be conveyed successfully. Another kind of phrase which has no canonical form is the *idiom schema* (Moon, 1996). This differs from the semi-prepackaged phrase in that it features a central metaphorical conceit which provides the motivation for the meaning conveyed. Idiom schemas can exploit semantic sets in the realization of their variant forms, allowing content words to be replaced by others with a similar semantic content (e.g., *quake/tremble/shiver* in place of *shake* in *shake in one's shoes*), but ultimately the conceit takes precedence. The schema [numeral + hyponym + *short of a(n)* + superordinate] with its underlying conceit of incompleteness can be realized as *one sandwich short of a picnic*, *several cards short of a full deck*, *a few gallons shy of a full tank*, all of which are interpreted to mean "slightly crazy" (Moon, 1996, p. 252).

Identifying and Extracting Phrases Using Technology

Fully Automated Methods of Identifying and Extracting Phrases

The inherent plasticity of phrases poses problems for computational approaches to phraseology, with the result that at present, the fully automated identification and retrieval of phrases from computer language corpora have still not been achieved. This does not mean that phrases and other recurrent word sequences cannot be retrieved automatically, only that retrieval is either over- or under-selective. If over-selective, it excludes the many noncanonical instantiations of the phrase; if under-selective, manual intervention is essential to discard irrelevant examples.

It is in seeking to extract phrases from computer corpus data that the difference between phrases and lexical bundles becomes clear. *Lexical bundles* (Biber, Johansson, Leech, Conrad, & Finegan, 1999) are uninterrupted strings of three or more words which recur with significant frequency in a data set. Since they are defined in terms of statistical frequency, they can also be identified and retrieved statistically: any three-word sequence in a data set can be said to be a lexical bundle provided that its frequency matches or exceeds the established cut-off point.

Phrases are on the whole much "messier" than lexical bundles, and automatic identification is only possible for completely frozen phrases and for the canonical forms of fixed and semi-fixed phrases. In other words, only phrases which are also lexical bundles can be retrieved automatically. Unfortunately, a great many phrases have a tendency to be variable. Noncanonical forms, as well as instantiations of semi-prepackaged phrases and idiom schemas, lack the regularity of lexical bundles, so much so that instantiations of phrases are often unique. As a consequence, they cannot be located automatically on the basis of recurrence or absolute frequency. The vast majority of lexical bundles, on the other hand, while identifiable automatically, are missing the "semantic completeness" that is necessary if they are to qualify as phrases. To give an example, the lexical bundle *a matter of* is a part of many established phrases (e.g., *it's a matter of life and death*, *it's only a matter of time*, *as a matter of fact*), but it must combine with other words before it can be said to express a complete meaning. It can be defined as a phrasal fragment, but it is not a proper phrase, while *as a matter of fact* is both a (frozen) phrase and a lexical bundle.

Since the majority of phrases are semi-fixed rather than completely frozen, there is a risk that they will not be picked up by algorithms designed to extract lexical bundles (or *n-grams*). These algorithms search for uninterrupted ("contiguous") sequences of particular word forms, not interrupted sequences of lemmas. But where traditional n-gram location fails, more sophisticated search procedures come in. Recent advances have led to the development of skip-grams and concgrams, both of which are designed to overcome the

limitations of n-grams while remaining fully automatic (no human intervention required) and therefore empirically sound.

As the name suggests, a *skip-gram* (Guthrie, Allison, Liu, Guthrie, & Wilks, 2006) allows for interruptions to occur within the word string because the algorithm can ignore ("skip") up to four extraneous words. As a result, the recurrent words are located even if they do not always occur in an uninterrupted sequence. For example, it would allow the string *bear in mind* to be identified even when it occurs as *bear something in mind*, that is, with the direct object pronoun in its grammatically appropriate place, without classing *bear it in mind* or *bear what you said in mind* as distinct n-grams. The main limitation of skip-grams is that they are unable to process strings longer than three words in length (tri-grams), and the four-word skips, while apparently generous, may be insufficient even to allow for a modified noun in direct object position.

Concgrams (Cheng, Greaves, & Warren, 2006; Greaves & Warren, 2007; Greaves 2009) offer better coverage in both of these areas. Concgrams can process 2- 3- 4- and 5-grams, and the maximum skip between words is 50 characters, which is approximately 12 words (Greaves, 2009, p. 2). The n-grams can be extracted automatically (though this takes a very long time), or can be user-defined instead; for example, stipulating *bear* and *mind* to uncover all co-occurrences of those words, contiguous and noncontiguous. Additionally, the algorithm searches for the words in the n-gram irrespective of their sequential order. This makes it possible to extract a range of permutations which can be found in natural language, especially fronting (e.g., **in mind I'll bear it*) or reversal (e.g., *a wolf in sheep's clothing*, *every silver lining has a cloud*). These advances have greatly facilitated the extraction of phrases from corpora, but some problems remain. The most important of these is probably that inflected forms, being distinct character strings, belong to separate n-grams (a search for *bear in mind* will not retrieve *borne in mind*). There is also the perennial issue of meaningfulness: Algorithms calculate frequency and recurrence but cannot differentiate between a string that is meaningful and one that is not. As a result, the data retrieved using skip-grams or concgrams require manual intervention to discard irrelevant hits before analysis can take place.

Semi-Automatic Methods of Identifying and Extracting Phrases

The study of phrases has a long history in the pre-corpus era, and many phraseologists limit their study to decontextualized canonical forms. Corpus phraseology is largely a branch of corpus lexicography, and the most prevalent reason for studying phrases in corpus data is to examine how they are actually used in text, to improve descriptions for dictionaries and language learners. In order to do this effectively, it is essential to go beyond the canonical form alone and examine all the uses of a phrase in the data.

Before the appearance of concgrams, corpus phraseologists had no option but to make judicious and sometimes inventive use of the available query software. Locating phrases in all their permutations is not straightforward, because corpus query software is designed to locate character strings (words), not meanings. Different scholars have tackled the matter in different ways, but in essence all of them have resorted to multiple searches which are then collated and verified manually. The searches performed make use of well-established software tools which are included as a matter of routine in corpus query packages, namely the wild-card character and the intervening span. The wild card can be attached to the end of a word to represent any further characters, and therefore to retrieve inflectional variants of regular verbs, noun derivations, and so forth; it can also be used in place of any single word. The intervening span option allows skips to be inserted, typically up to a maximum of five words. Additionally, most commercially-available corpora have been lemmatized. Lemmatizing software indexes all inflected forms under their lemma, so if the data set has been lemmatized, it is possible to extract all inflected forms

in a single search. Applied to *bear in mind*, this would draw together *bear* and all its inflected forms: *bears*, *bearing*, *bore*, and *borne*. The advantage of a lemmatizer over a simple wild card is that it includes irregular forms alongside the regular ones: In other words, *bore* and *borne* are located with the help of a lemmatizer but not with a wild card (it would only retrieve *bear*, *bears*, and *bearing*).

Philip (2008, 2011) explains in some detail the kind of procedure that has to be followed in order to extract phrases semi-automatically from corpus data. The procedure is admittedly laborious, but at present the most exhaustive available. A first search extracts the canonical form. Subsequent searches are designed to extract inflectional variants, truncated forms (by shortening the string), inversion (by reversing the string), and substitution (using wild cards to replace actual words). In all of these searches, the number of possible intervening words is set at the maximum permitted by the query software (typically five). The combined queries are then collated and duplicates eliminated. Following this comes manual selection of the examples, in which instances of the phrase, including apparent exploitations and permutations, are set to one side and non-candidates discarded. Finally, the relevant concordances are subjected to analysis, which may be computer-aided if the selection is saved as a new file and run through a concordance package.

Working With Phrases

Extracting phrases from corpora is a means to an end, not the end itself. Only once the instantiations of a phrase—be it frozen or variable, in its canonical form or a variant—have been gathered together is it possible to proceed with linguistic analysis.

Phrases are meaningful by definition, but meaningfulness comes in several guises. In addition to the more familiar semantic meaning, phrases are also associated with pragmatic, functional meanings. Corpus phraseology views phrases as lexical items which, like single words, are part of "units of meaning" (Sinclair, 1996), meaning that they attract collocations, colligations, semantic preferences, and semantic prosodies in "concentric rings of phraseology" (Philip, 2011, p. 38). The semantic prosody of a phrase is its function, so it should come as no surprise that phrases can perform any one of a range of functions: They can be evaluative, situational, modalizing, text-organizational, or all of these (Moon, 1992, p. 495) in addition to, or indeed instead of, conveying information. Some phrases (e.g., *on the other hand*, *as it were*) are primarily pragmatic and have minimal semantic value; others—famously idioms—combine semantics with a range of expressive functions. These functions can be ascertained only by studying phrases in context. Corpus linguistic tools, automatic and semi-automatic, have allowed phraseology to move beyond semantics and syntax toward a fuller understanding of what phrases are used for, and why.

SEE ALSO: Formulaic Language and Collocation; Formulaic Sequences; Lexical Bundles and Technology; Lexical Priming; Semantic Prosody

References

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

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*(4), 411–33.

Francis, G. (1993). A corpus-driven approach to grammar: Principles, methods and examples. In M. Baker, G. Francis, & E. Tognini Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 137–56). Philadelphia, PA: John Benjamins.

Greaves, C. (2009). *ConcGram 1.0: A phraseological search engine*. Amsterdam, Netherlands: John Benjamins.

Greaves, C., & Warren, M. (2007). Concgramming: A computer-driven approach to learning the phraseology of English. *ReCALL, 19*(3), 287–306.

Gries, S. T. (2008). Phraseology and linguistic theory. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). Amsterdam, Netherlands: John Benjamins.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). A closer look at skip-gram modelling. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 1222–5). Genoa, Italy.

Moon, R. (1992). There is reason in the roasting of eggs: A consideration of fixed expressions in native-speaker dictionaries. In H. Tommola, K. Varantola, T. Salmi-Tolonen, & J. Schopp (Eds.), *EURALEX '92: Proceedings* (pp. 493–502). Tampere, Finland: University of Tampere.

Moon, R. (1996). Data, description, and idioms in corpus lexicography. In M. Gellerstam, J. Järborg, S-G. Malmgren, K. Norén, L. Rogström, & C. R. Papmehl (Eds.), *EURALEX '96: Proceedings* (pp. 245–56). Gothenburg, Sweden: Gothenburg University Press.

Philip, G. (2008). Reassessing the canon: "Fixed" phrases in general reference corpora. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 95–108). Amsterdam, Netherlands: John Benjamins.

Philip, G. (2011). *Colouring meaning: Collocation and connotation in figurative language*. Amsterdam, Netherlands: John Benjamins.

Sinclair, J. M. (1996) The search for units of meaning. *Textus, 9*, 75–106.

Suggested Readings

Cowie, A. (Ed.). (1998). *Phraseology: Theory, analysis, and applications*. Oxford, England: Oxford University Press.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam, Netherlands: John Benjamins.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London, England: Routledge.

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. Amsterdam, Netherlands: John Benjamins.

Moon, R. E. (1998). *Fixed expressions and idioms in English.* Oxford, England: Clarendon.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, England: Cambridge University Press.