Formulaic sequence annotation guidelines

Julian Brooke

September 9, 2016

The annotation task is to identify whether a sequences of words is "formulaic" in your language. We are using the definition given by Wray (see attached pdf for more details). She defines a formulaic sequence (FS) as "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"; you can think of FS as being something weve learned and have stored in a mental lexicon, and can recognize and retrieve more easily than entirely novel constructions).

FSes cover a wide variety of linguistic phenomena, including some that are sometimes viewed as syntactic in nature (like verb/preposition combinations). The most obvious examples of FS include large, fixed expression such as "see the light at the end of the tunnel", but also include much shorter sequences such as semantically non-compositional noun phrases like "red tape". A sequence does not need be syntactically complete to be a formulaic sequence, in fact many involving verbs are not: "become conscious of" and "be one of the first to" are examples of FS which are complete with respect to FS, but have a syntactic gaps which can be filled with a variety of expressions which do not form part of the FS. These gaps can occur with a sequence, for instance "pick * up" or "worth * weight in gold". Different sequences will have different fillers for these gaps depending on their syntactic role; what distinguishes a gap from part of FS is that the content of a gap is restricted only to general syntactic or semantic categories, whereas parts of a sequence will be restricted to specific words. For example any small physical object can be picked up, justifying a gap in the middle of the phrase "pick * up", but the only verbs that regularly come before "conscious of" are "become" and "be", so the expression seems highly lexicalized (i.e. "prefabricated"),

and therefore formulaic.

Wray gives a number of criteria for identifying FS, a few of the most useful for this annotation task are:

1. Semantic non-compositionality. This is the easiest and most obvious indicator: if the sequence of words has a special meaning that is more than the meaning of its parts (like "red tape" and "kick the bucket").

2. Associated with a specific linguistic situation. An example of this is a phrase like "it would be great if * could", which is semantically compositional but also has a larger discourse function that has been formulized, namely as a expression of (unrealized) expectation.

3. The preferred way a native speaker (or a native sub-community) would express an idea. This can include specific, otherwise compositional words like "global warming" (as compared to, for instance, "earth warming"). If you speak another language, it may be useful to think whether the phrase translates directly into another language; if not, try to think of other ways to express the same idea, and see if they feel similarly idiomatic.

Generally speaking, as a native speaker you should have an intuition about which sequences have this special property (they are part of your mental lexicon in some way), or, in case when they aren't a formula you would use, at least some intuition about whether it would be formulaic to some community of native speakers (e.g. "ventral cord" is a term doctors would know, even if you dont). Many names are formulaic sequences, but not if they would not be recognized by some larger community of native speakers (i.e. the name of some random non-famous person), or are otherwise non-standard ways to refer to the entity with extra material that is derived in a rule-based manner (e..g "Barak Obama", "President Obama", and "President of the United States" are FSes, but "President of the United States Barak Obama" and "Mr. Barak Obama" are not). In English, another special case is phrases like "the crust" and "a stomach ache", which we would consider formulaic because (in common usage) the article is pretty much fixed ("the crust", "the filling" in the context of pies is similar to "the floor" "the ceiling when in a building, the library" or "the sun", "the universe", when living on earth. "A stomach ache" is like "a cold" vs. "the flu" vs "the runs" vs "cancer", getting an article in this case is pretty idiosyncratic, and therefore deserving

of formulaic status. Here are a larger list of kinds of linguistic phenomena that are often formulaic:

compound nouns
adjective/noun combinations
Names of well-known entities (people, places, locations)
verb/preposition combinations (both gappy, and non-gappy)
light verb constructions
multiword discourse connectives
verb/noun combinations
verb/adverb combinations
Proverbs, idioms, and clichs

However, almost any sequence could be formulaic, and so you should not use a fixed set of possible syntactic types to limit possible FS. For some FS, there is some variation in the form of some of the words ("become conscious of" could be "became conscious of"), and in other cases (like "red tape") the form is fixed ("red tapes" no longer has the idiomatic meaning); when annotating, you will be shown a particular common form but when there is variation in the possible form, you should view it as an annotation of all possible forms, not just the one you're seeing. Be careful with verb forms, don't exclude a form of "be" (or some other verb) from the canonical form just because it varies by inflection ("was/were/is up for" are all good forms). However if the grammatical differences that involve independent words (like perfect aspect in English, which is expressed with "have"), it should be included in a FS only if the grammatical aspect seems essential to the formula. Instead of 2 options (FS or not), you are being asked to annotate three options; a sequence can be totally non-formulaic, it can recall a formulaic sequence, or it can be a canonical formulaic sequence. You should annotate the middle option (recalls) when the phrase contains or otherwise immediately brings to mind something you would call an FS (it should recall on its own, without the examples), but there are important parts of the FS missing, or extra elements that are not essential to the FS. Generally speaking, canonical forms should not have extra, highly variable words included on either end, or in the middle. For example, if the n-gram is "he was up for it", both "he" and "it" are extra and shouldn't be included in the canonical form (consider that you can say "they were up for some annotation"). Note that there are

of course cases where a pronoun (often "it") is part of the canonical form even when it appears at a boundary, for example "it's a shame that", but in those cases you can't switch the pronoun for another one. Note that if a variable pronoun (or some other word) appears within a phrase then that is also not a canonical form. For example, "respect his opinion" only recalls a canonical form, the true canonical form is respect * opinion. More generally, (as alluded to above) canonical forms should not include words that can freely vary across some general semantic or syntactic class; we are only interested in sequences for which there seems to be some special lexical affiliation suggesting a fixed phrase, not a productive process. That said, there could be cases where there are a small, idiosyncratic set of lexical possibilities where we can still say, with some confidence, that each deserves to be considered a formulaic phrase in its own right even if the meaning is the same: for example "be available in a range of" and "comes in a range of" (size, shapes, colours, etc.) are two semantically equivalent phrases that should both be considered FS, they should not be generalized to "in a range of" just because there is a small amount of lexical variation. Similarity, "the point I'm trying to" is likely to be completed with "make"; the fact that "raise" is also possible (though less common) does not mean that it should be generalized; "the point I'm trying to" simply doesn't stand on its own without a verb, and there are only a handful of possible verbs. That is, if most of the time the phrase is used, a particular element is included, then it should be considered canonical (and the other form non-canonical) even if it can sometimes be elided under certain usage circumstances (e.g. "as painful as it is" can sometimes be written as just "painful as it is", or when "the Old World" loses its "the" in adjectival form). If there is a small amount of lexical variation which seems fairly idiosyncratic then prefer the larger versions with each rather than generalizing. If there is both quite a clear preference for a particular lexical choice, but also lots of potential substitutions, then both longer and shorter could be considered canonical: for instance "there is something wrong with" seems to merit its own formula, but "there is something * about" is also a good general formula. Otherwise, if the options don't seem to particularly distinguish themselves, or seem to be restricted only by a general (word class) rule (e.g. "on Monday"), then prefer a shorter version that doesn't include the element.

You are being shown a small set of examples, which are random instances selected from the corpus. You can use this as evidence to help you decide if a formulaic sequence is canonical, though keep in mind that you are seeing

only positive examples; if you think the FS might be smaller than what you see, you will have to rely on your own intuition about whether it occurs on its own. You should also be sensitive to the data in that you should only tag something as formulaic if you think the examples actually show evidence of that formulaticity, even if you think it is possibly FS.