Multiword Expressions and Lexicalism

Jamie Y. Findlay
University of Oxford

Proceedings of the LFG'17 Conference

University of Konstanz

Miriam Butt, Tracy Holloway King (Editors)

2017

CSLI Publications

pages 209–229

http://csli-publications.stanford.edu/LFG/2017

Keywords: multi-word expressions, lexicalism, idioms, Tree Adjoining Grammar

Findlay, Jamie Y. (2017). Multiword Expressions and Lexicalism. In Butt, Miriam, & King, Tracy Holloway (Eds.): *Proceedings of the LFG'17 Conference, University of Konstanz* (pp. 209–229). Stanford, CA: CSLI Publications.

Abstract

Multiword expressions (MWEs) such as idioms exhibit a tension between phrase-like and word-like properties. Much recent work has treated idioms as exclusively phrase-like, by positing special idiom versions of the words they contain. In this paper, I argue that such approaches are unappealing, and suggest that by following the ideas of Abeillé (1995), we can provide a more satisfying analysis that respects the special status of MWEs. This is implemented by replacing the context-free grammar standardly assumed for LFG c-structures with a Tree Adjoining Grammar (Joshi et al. 1975). This allows us to represent idioms in a single place, even when their parts can be individually modified and/or targetted by morphosyntactic operations.

1 Background

1.1 Multiword expressions

Multiword expressions (MWEs) are of interest to linguistic theory because of the tension they exhibit between a divided (phrase-like) and a unitary (word-like) nature. Consider the idiom in (1):

(1) *take the biscuit* 'be egregious/shocking'

This is clearly made up of multiple, independently recognisable English words, which inflect individually (for example, the past tense form is *took the biscuit* not *take the biscuit-ed*), i.e. it is like a common-or-garden phrase. At the same time, however, it has a unitary, and non-compositional, semantics, which only emerges when the words are used together. Notice that neither word can bear (some part of) the idiomatic meaning alone:[1]

(2) a. #What a dramatic biscuit! ($\neq$ What a dramatic shock/outrage!)
b. #That really takes it. ($\neq$ That's really egregious/shocking.)

Because of their idiosyncratic semantics, and the fact that the parts must co-occur, it seems necessary that these expressions be stored somehow.

The scope of the label 'MWE' is broad, and includes such phenomena as periphrases, nominal compounds, phrasal verbs, and idioms (Baldwin & Kim 2010). These each raise their own analytical problems, but the common challenge which they pose is how to resolve the tension between their word-like and phrase-like

[†] For helpful and insightful discussion on this topic, I would like to thank Doug Arnold, Ash Asudeh, Alex Biswas, Mary Dalrymple, Timm Lichte, John Lowe, Stephen Pulman, and Manfred Sailer. Thanks also to the two anonymous reviewers for their constructive feedback. This research was carried out while I was the recipient of a UK Arts and Humanities Research Council grant (grant no. AH/L503885/1), which I gratefully acknowledge.

[1] Here and throughout, the # marker of semantic oddity is used to indicate that the intended idiomatic reading is not available.

properties.[2] The focus of this paper is idioms (on which there is a considerable literature: see e.g. Katz & Postal 1963; Chomsky 1980; Nunberg et al. 1994), but much of what is said carries over to the analysis of other kinds of MWE as well.[3]

1.2 Idioms

Idioms are non-compositional in the sense that their meanings are not a function of the literal meaning of their parts and the way they are put together. Their meanings therefore have to be learned, and oftentimes seem to be just as arbitrary as any given lexical entry. For example, although *kick the bucket*, *look a gift horse in the mouth*, and *shoot the breeze* might all have originated in perfectly coherent metaphors, it is now the case that for many, if not most, speakers they are synchronically opaque. (All the same, as Nunberg et al. 1994: 492–493, fn. 2, point out, speakers do recognise that *some* figuration is at play, they may just have no idea what particular metaphor is being evoked.)

In spite of their non-compositional semantics, idioms nonetheless appear in the syntax as multiple, distinct word forms, and these can be separated, modified, and, as mentioned, inflected individually. It is this (morpho)syntactic flexibility which makes idioms challenging for linguistic analysis: ideally, they should be stored locally, as a unit, to account for their unitary properties, but their parts must also be individually accessible to the syntax, and may ultimately end up separated. For, although some idioms share the limited syntactic flexibility of periphrases and other kinds of MWEs like compounds (as in (3)), others show a considerable degree of freedom, whereby their parts can end up arbitrarily far apart (as in (4)).

(3) a. Old Man Mose kicked the bucket.

[2] Ackerman et al. (2011) discuss this tension in morphology under the rubric of the *principle of unary expression*, whereby each lexeme is preferably to be expressed in syntax as "a single morphophonologically integrated and syntactically atomic word form", and how this is challenged by the facts of periphrasis, where cells in a lexeme's paradigm appear to be filled by more than one word form.

[3] Although, of course, not all of what is said carries over to the analysis of all other kinds of MWE. For example, as a reviewer sensibly points out, simple nominal MWEs like 'New York' or 'Jack the Ripper' can plausibly be treated as 'words with spaces' (Sag et al. 2002), i.e. as atomic lexical items that just so happen to be written as multiple words. That said, however, it is clear that these expressions are not totally immune to linguistic analysis, as evidenced by word play – 'Newer York', 'Jack the former Ripper', etc. – and so we might prefer to represent them as full NPs, with the accompanying internal structure, but mark them in some way so as to 'close off' their internal structure in normal usage.

On the other hand, there are complex predicates and light verb constructions (LVCs), also often considered to be MWEs. These are (at least semi-)productive, and (to some extent) follow systematic combinatorial rules. They thus constitute a different class of MWEs, analytically speaking, from the semantically idiosyncratic idioms I examine in this paper. This is not just a self-serving distinction on my part: LVCs, at least, also exhibit markedly different psycholinguistic properties from idioms, being *harder* to process than literal expressions (Wittenberg & Piñango 2011), unlike idioms which are *easier* to process. This points to a more complex kind of semantic composition for LVCs, perhaps along the lines outlined by Lowe (2015) for complex predicates, and a less complex kind for idioms, along the lines outlined in this paper.

b. #The bucket was kicked (by Old Man Mose).
c. #Which bucket did Old Man Mose kick?
d. #The bucket that Old Man Mose kicked was {sudden/sad/…}.

(4) a. He pulled strings to get me assigned to his command.
b. Strings were pulled to get me assigned to his command.
c. Which strings did he pull to get you assigned to his command?
d. The strings that he pulled got me assigned to his command.

Similarly, although some kinds of idiom only allow so-called 'external' modification (Ernst 1981), whereby adjectives which appear inside the expression actually take scope over the whole idiom meaning (delimiting a domain, as in (5)), many allow extensive internal modification or quantification over sub-parts of their meaning (as in (6)).

(5) *External modification*:
a. Musicians keep composing songs 'til they **kick the proverbial bucket**.
(= … 'til, proverbially speaking, they kick the bucket.)
(GloWbE)[4]
b. Britney Spears […] **came apart at the mental seams**.
(= Mentally, Britney came apart at the seams.)
(http://bit.ly/2jZmYKP)
c. Let's say […] you want to **return the oral sex favour** he happily gives to you.
(= In the domain of oral sex, you want to return the favour.)
(http://bit.ly/2y4jeOx)

(6) *Internal modification*:
a. Delhi's politicians **pass the polluted buck**.
(The issue which is being avoided is polution.)
(http://on.ft.com/2y4fbBJ)
b. Maybe by writing this book I'll offend a few people or **touch a few nerves**.
(= I will upset a few people or annoy someone in a few ways.
≠ I will cause the same irritation multiple times.)
(http://bit.ly/2y56ibi)
c. Tom won't **pull family strings** to get himself out of debt.
(The connections which Tom won't exploit are family ones.)
(http://bit.ly/2y4tKFg)

This syntactic flexibility exacerbates the tension between the divided and unitary nature of idioms, since it sharpens the feeling that they are made up of words which enter the syntax individually, and yet they still retain their idiosyncratic, and collocationally restricted, semantics.

[4] Corpus of Global Web-based English (Davies 2013).

In this paper, I address one common theme in recent work on idioms, which seeks to resolve this tension by coming down on one side of it, treating idioms as truly phrasal, being made up of special versions of the words they contain, and having no unitary identity. I demonstrate that there are a number of problems with this approach, both theoretical and empirical, and argue that it cannot be sustained. Instead, I advocate a change to the LFG architecture, increasing the power of c-structure using a Tree Adjoining Grammar, which enables us to adopt a version of Abeillé's (1988, 1995) approach to idioms.

2 The lexical ambiguity approach

One common approach to idioms in lexicalist theories is what I propose to call the *lexical ambiguity* approach (LA). In such an approach, idioms are treated as made up of special versions of the words they contain, which combine to give the appropriate meaning for the whole expression. For example, words like *pull* and *strings* become ambiguous, meaning either **pull**′ and **strings**′ in the literal phrase *pull strings*, or **exploit**′ and **connections**′ in the idiom. This kind of approach resolves the tension in favour of treating idioms as phrase-like: they are no longer seen as single lexical items, but rather collections of separate lexical items which conspire to create the overall meaning.

Examples abound in the literature: Sailer (2000) in HPSG, Kay et al. (2015) in SBCG, Lichte & Kallmeyer (2016) in LTAG, and Arnold (2015) in LFG, for instance. Not all of what I discuss in this section is relevant to all of these approaches, and so it should not be read as a direct rebuttal of the explicit claims they make, but rather as an objection to the overall philosophy which they share.

2.1 Strengths of LA

Before my objections, however, let us consider the strengths of such an approach. LA is particularly well suited to explaining so-called decomposable idioms (what Nunberg et al. 1994 call *idiomatically combining expressions*), where the meaning of the whole can be distributed across the parts. Examples of this include *pull strings*, as mentioned, where *pull* ≈ **exploit**′ and *strings* ≈ **connections**′, as well as *spill the beans*, where *spill* might be identified with **divulge**′ and *beans* with **secrets**′. Other examples are given in (7) and (8).

Since the idiom meaning is assigned to the individual words in LA, this immediately explains the fact that parts of these idioms can be separated by syntactic operations, as in (7), or that they are open to internal modification and/or quantification, as in (8), because they are simply ordinary words, and can undergo all the processes ordinary words can.

(7) a. Cantor duly ran to teacher and **the beans got spilled**.
(http://bit.ly/2k6741B)

b. Who's at the centre of the **strings that were quietly pulled**?
(http://imdb.to/2y87Ilf)

c. Wait until next month, and we'll see **which bandwagon he jumps on**.
(http://bit.ly/2k25tcR)

(8) a. Yet from Carnap's point of view, Quine's argument in §5 is **beside the main point**, which is whether the notion of a semantical rule is a purely logical one.
(http://bit.ly/2k3EL3N)

b. Sorting out that little mess required **pulling several strings**.
(http://bit.ly/2k1aQZQ)

c. Brace yourselves as Claudine **spills some untold beans**.
(http://bit.ly/2k1spZY)

2.2 Problems with LA

Despite this obvious advantage, by essentially ignoring the tension which MWEs pose, and coming down entirely on one side of it, LA leaves a number of questions to be answered, some of which, I suggest, cannot be answered satisfactorily.

2.2.1 Selectional restrictions

If *pull* can mean **exploit**′ and *strings* can mean **connections**′, we clearly have to prevent them occurring apart from one another:

(9) a. #You shouldn't pull his good nature.
($\neq$ ... exploit his good nature.)

b. #Peter was impressed by Claudia's many strings.
($\neq$ ... Claudia's many connections.)

The most straightforward way to do this is to treat idiom formation as a kind of limit case of selectional restriction, and make those restrictions mutual:[5]

(10) *pull* V $(\uparrow \text{PRED}) = \text{`pull}_{id}\text{'}$
$(\uparrow \text{OBJ PRED FN}) =_c \text{strings}_{id}$

(11) *strings* N $(\uparrow \text{PRED}) = \text{`strings}_{id}\text{'}$
$((\text{OBJ} \uparrow) \text{ PRED FN}) =_c \text{pull}_{id}$

All lexical theories will have some way of identifying individual lexemes; in this case, we use the PRED feature, but other frameworks have similar options (the *lex-id* or LID features in HPSG/SBCG, for example).

[5] Gazdar et al. (1985) propose to instead enforce these restrictions in the semantics, by making use of partial functions (so that idiomatic functions are undefined unless they are passed their idiomatic complements as arguments). Unfortunately this elegant solution runs into insoluble problems when it comes to relative clauses, and will necessarily over- or undergenerate. See Pulman (1993: 50f.) for details.

As written, however, the restrictions in (10) and (11) are too strong, since this idiom can passivise, and so it is not true that *strings* must be the *object* of *pull*:

(12) Strings were pulled for you, my dear. Did you really think the Philharmonic would take on a beginner like you?
(http://bit.ly/2y8gIqF)

One way to loosen the restriction is by moving the constraint from f-structure to s-structure (or, equivalently, to a-structure, if one prefers a different architecture):

(13) $$\textit{pull} \quad \text{V} \quad \begin{array}{l} (\uparrow \text{PRED}) = \text{‘pull}_{id}\text{’} \\ ((\uparrow_\sigma \text{ARG}_2)_{\sigma^{-1}} \text{ PRED FN}) =_c \text{strings}_{id} \end{array}$$

(14) $$\textit{strings} \quad \text{N} \quad \begin{array}{l} (\uparrow \text{PRED}) = \text{‘strings}_{id}\text{’} \\ ((\text{ARG}_2 \uparrow_\sigma)_{\sigma^{-1}} \text{ PRED FN}) =_c \text{pull}_{id} \end{array}$$

Instead of requiring that idiomatic *strings* be the object of idiomatic *pull*, we require that it be its second/internal argument.

But this doesn't help with relative clauses:

(15) *The strings (that) he pulled...*

$$\begin{bmatrix} \text{PRED} & \text{'STRINGS}_{id}\text{'} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'PULL}_{id}\text{'} \\ \text{TOPIC} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{"HE"} \end{bmatrix} \\ \text{OBJ} & \end{bmatrix} \right\} \end{bmatrix}$$

In the standard analysis of restrictive relative clauses such as this, the word *strings* bears no direct relation to *pull*: it is not its object nor its ARG2 – the anaphoric element is instead – so (15) shouldn't be licensed.[6] Falk (2010) sees this as evidence for an 'unmediated' analysis of relative clauses, where we abandon the anaphoric element which mediates between the clauses in favour of a direct relationship between the predicate of the relative clause and the head noun. If we stick with the 'mediated' version, however, we cannot explain the distribution of some idioms, at least not without introducing *ad hoc* disjunctive specifications of the relationship between their parts.

[6] Miriam Butt (p.c.) suggests that a properly articulated argument structure would solve this problem, by allowing a single argument (say, the internal argument of *pull*) to be multiply realised (as both the head noun *strings* and the relative pronoun, either overtly or merely functionally), but I do not see how this is intended to be cashed out formally. Firstly, all versions of a-structure appeal to some variety of (Lexical) Mapping Theory to link arguments to GFs, and (L)MT assigns only a single GF to each argument. Secondly, if a-structure is positioned in the correspondence architecture before c-structure, as it is by Butt et al. (1997), it cannot permit a single argument to map to two separate realisations, because the relation α from a-structure to c-structure is a function.

2.2.2 Non-decomposable idioms

Although LA looks strong when it comes to decomposable idioms, it is not so clear how such an approach should handle non-decomposable ones, like *kick the bucket*, *blow off steam*, *shoot the breeze*, etc. (what Nunberg et al. 1994 call *idiomatic phrases*), where there is no obvious way of breaking down the meaning of the idiom such that its parts correspond to the words that make up the expression.

Assuming a resource-sensitive semantics, as is common practice in LFG (e.g. Asudeh 2012), we are forced to say that only one of the words in the expression bears the meaning, and the rest are semantically inert. For example, perhaps there is a $kick_{id}$ which means **die**′, and selects for special semantically inert forms the_{id} and $bucket_{id}$.

But the choice of where to locate the meaning is ultimately arbitrary. While it might intuitively seem to make sense to assign it to the verb, since it is the head of the VP which makes up the expression, formally it makes no difference: we may as well have $bucket_{id}$ meaning **die**′, or even the_{id}, provided they select for the other inert forms and then pass their meaning up to the whole VP.[7]

In addition, we also now face a huge proliferation of semantically inert forms throughout the lexicon.[8] What is worse, each of these must be restricted so that it does not appear outside of the idiomatic context. For example, say that we want a semantically vacuous *the* to use in *kick the bucket*. To prevent it appearing spuriously elsewhere (e.g. **The Kim sneezed*), it must, as discussed in Section 2.2.1, impose restrictions on what it can occur with. But if it says that it must be the specifier of idiomatic *bucket* (or the specifier of the object DP of idiomatic *kick*), then it cannot appear in other idioms which involve the word *the*, such as *shoot the breeze*. The *the* in *shoot the breeze* must be different from the one which appears in *kick the bucket*, since it imposes different selectional restrictions. But this means that we need as many *the*s as there are expressions which include it. Instead of

[7] One possible argument for the head-based analysis is that VP idioms systematically retain the aspect of the literal use of the verb (McGinnis 2002):

(i) a. Hermione was dying for weeks.
b. #Hermione was kicking the bucket for weeks.
[*Kick* is punctual: the only idiomatic reading of (ib) would be that Hermione died repeatedly, whereas (ia) can describe a single, protracted dying event.]

(ii) a. Harry ate his vitamins {in two seconds flat/*for five minutes}.
b. Harry ate his words {in two seconds flat/*for five minutes}.

However, I think this is part of the much larger issue of how much the literal meaning of an idiom persists in its figurative use. Cf. also Ernst (1981) and his discussion of examples like *pulling [Malvolio's] cross-gartered leg*, where a modifier appropriate to the literal but not figurative meaning is used.

[8] Arnold (2015) suggests using manager resources to eliminate the need for semantically inert forms, for example by having a special idiomatic *kick* which simply throws away the meaning of *the bucket*. Arnold himself notes a number of shortcomings of this approach, since it makes the wrong predictions about modification and cannot easily explain variation in syntactic flexibility. See the Appendix for more details.

having to expand the lexicon by as many entries as there are idioms, we have to expand it by as many entries as there are *words in idioms*. This seems suspect from an analytical point of view, and undoes much of the elegance of LA.

2.2.3 Processing

Swinney & Cutler (1979) showed that idioms are processed in the same way as regular compositional expressions; i.e. there is no special 'idiom mode' of comprehension which our minds switch into when confronted with idiomatic material. At the same time, these authors and others have found that idiomatic meanings are processed faster and in preference to literal ones (Estill & Kemper 1982; Gibbs 1986; Cronk 1992; i.a.). These findings are challenging for LA, for, in the LA approach, semantic composition of idioms is exactly the same as of literal expressions. There is no reason to think idioms should be processed any faster; if anything, we might expect them to be slower, since they involve ambiguity by definition.

3 Extending the power of c-structure

If we do not represent idioms as units, it is difficult to ensure that they always appear in the correct collocational environments. It is also difficult to handle instances where the semantics is itself seemingly unitary. Finally, it is a mystery why idioms should be processed faster than literal expressions, when formally they are identical. Rather, all of these findings plead for what I would imagine seems intuitively appealing anyway: that idioms are inserted *en bloc*, being stored in the lexicon as units, albeit with some internal structure.

The major obstacle to this in LFG is that the non-local character of idioms is ill-suited to the strict locality of context-free grammar rules. What I propose, therefore, is to add power to the c-structure component so that such non-local relations *are* statable. Tree Adjoining Grammar (TAG: Joshi et al. 1975; Abeillé 1988), with its 'extended domain of locality', offers such a possibility.

3.1 LTAG

In this subsection, I introduce very briefly the key features of TAG. For a fuller introduction, see Abeillé & Rambow (2000).

Whereas a context-free grammar is a string-rewriting system, a TAG is a tree-rewriting system. This means that, in a TAG, trees, not words, are the elementary components of the grammar. 'TAG' is a broad term for a mathematical formalism, just as 'context-free grammar' is. Lexicalised TAG (LTAG) is the linguistically relevant subtype, where each tree must be 'anchored' by at least one word form (Schabes et al. 1988).

A TAG consists of a set of *elementary trees* and the two operations of *substitution* and *adjunction* for combining them. In the next two parts, I discuss these two components in turn.

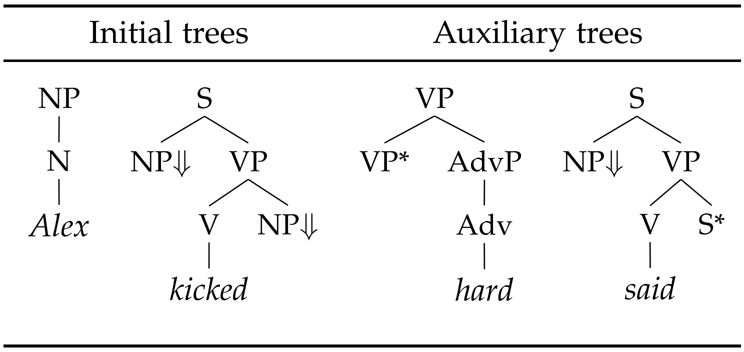


Table 1: Some elementary trees

3.1.1 Elementary trees

Elementary trees come in two types: *initial* and *auxiliary* (Table 1). An initial tree is a tree where all of the frontier nodes are either terminals or else non-terminals marked as *substitution sites* by a down arrow (⇓).[9] Substitution sites correspond to the arguments of a predicate.

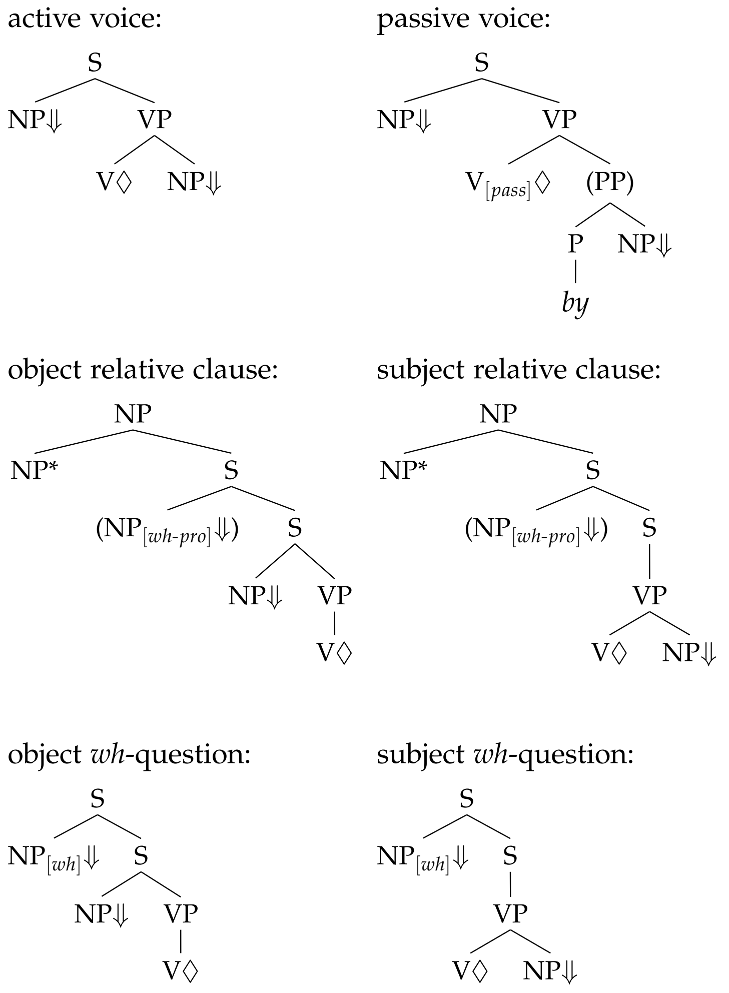
An auxiliary tree is an elementary tree in which one of the frontier nodes is specified as the *foot* node, and marked with an asterisk (*). This node must be labelled with the same symbol as the root node of the auxiliary tree.

Predicates are associated with *tree families*, sets of trees which represent their potential syntactic realisations. For example, Figure 1 shows part of the tree family for a transitive verb, including active and passive voice versions, relative clauses headed by the subject or object, and *wh*-questions where the subject or object is fronted. Such tree families are shared by all verbs of a particular class, and so we omit the specific head verb and mark the node where it appears with a lozenge (◊). Nodes marked with brackets are really abbreviations for pairs of trees, one where the subtree rooted in the bracketed node appears and one where it is absent.

One thing to note about TAG elementary trees is that because we are no longer restricted to the strict locality of context-free rules, *viz.* a node and its daughters, we obtain what is called an *extended domain of locality*: what counts as local, i.e. what can appear in a single object in the grammar, has expanded. Subject-verb agreement, for example, no longer needs to be mediated via features passed up to the VP (so that in reality we have subject-VP agreement), since the subject and the verb now both appear in the same elementary structure, and so dependencies between them can be directly encoded.

Abeillé (1988, 1995) has observed that such an extended domain of locality offers a particularly natural way of describing idioms. We simply allow elementary trees to be ‘multiply anchored’, so that more than one frontier node is filled by a terminal node, as in (16):

[9] I depart from standard TAG practice of using ↓ so as to avoid confusion with the LFG metavariable.


active voice:

passive voice:

object relative clause:

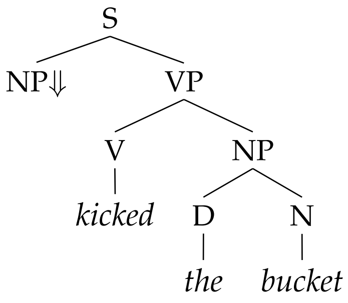
subject relative clause:

object *wh*-question:

subject *wh*-question:

Figure 1: (Partial) tree family for a transitive verb

(16)

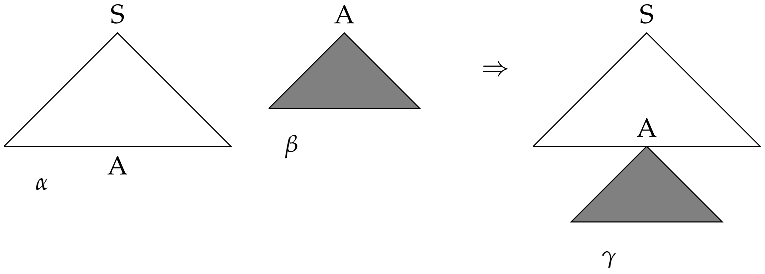


In this way, single lexical entries can contain more than one word form.

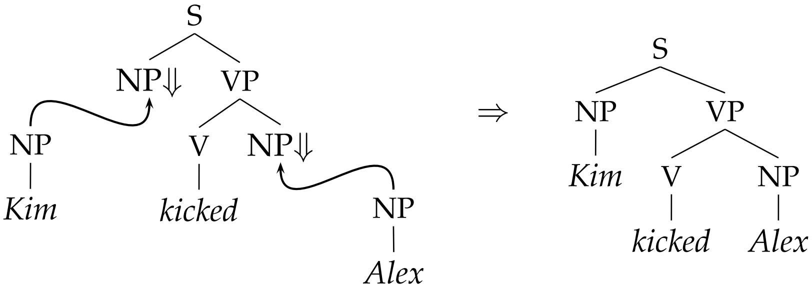
3.1.2 Substitution and adjunction

TAG provides two operations for manipulating elementary trees. Substitution is simply the replacement of an appropriate substitution site by an elementary or derived tree whose root node matches the symbol at the substitution site. This is illustrated schematically in (17), and with a linguistic example in (18):

(17)

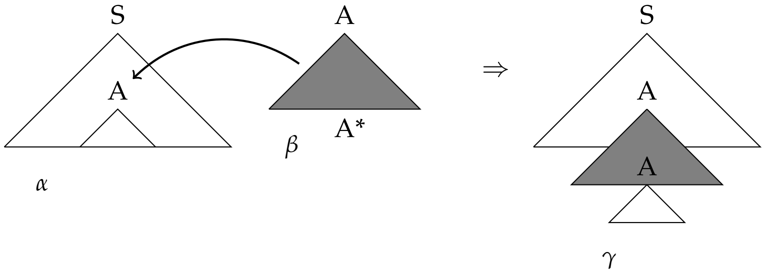


(18)



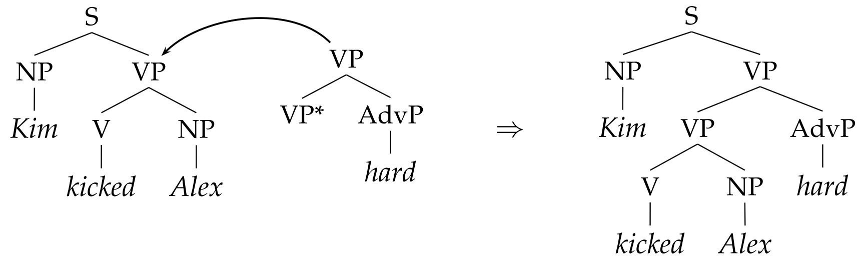
Adjunction is shown schematically in (19):

(19)



To adjoin β into α, we remove the subtree rooted in A from α, replace it with β, and then attach the subtree which we removed to the foot node of β. This produces a larger tree, γ. In effect, the auxiliary tree is inserted at the adjunction site and 'expands' the node around itself. This is commonly used to model the behaviour of modifiers, which adjoin to the node they modify:

(20)



In addition to modifiers, this is also how LTAG accounts for unbounded dependencies. As we saw in Figure 1, *wh*-dependencies are encoded locally in the elementary trees for a verb. Sentential embedding verbs are modelled as auxiliary trees in TAG, and this means that they can be adjoined to the interior nodes in such *wh*-extraction trees. The result of this is that such trees can grow from the inside out, meaning that the *wh*-element and the verb can end up arbitrarily far apart, even though they are represented locally in the lexicon. We will see an example of this in §3.3. This ability to represent relationships locally, even though the parts involved may ultimately appear separated, is one of the key advantages of the TAG approach to idioms.

3.2 TAG-LFG

Before we see exactly how this approach deals with the idiom facts identified above, let us see how a TAG can be incorporated into the LFG architecture.

In standard LFG, a lexical entry is a triple (W, C, F), where W is a word form, i.e. the terminal node in the phrase-structure tree, C is a c-structure category, i.e. the pre-terminal node, and F is a functional description, i.e. a set of expressions spelling out additional linguistic information via the correspondence architecture. In TAG-LFG, a lexical entry is instead a triple $(\langle W \rangle, T, F)$, consisting of a list of word forms, a tree, provided by some metagrammar, and a functional description.[10] An example is given in Figure 2.

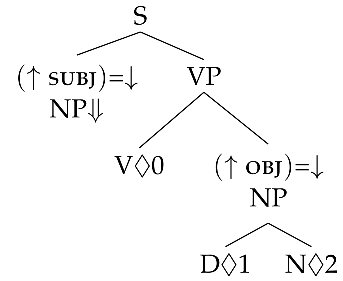
The word forms occur as a list because the trees for MWEs will be multiply anchored. For regular lexical entries, this list will be a singleton. The word form list is separated from the tree because the two elements of the entry come from different parts of the grammar: the word forms come from the morphology, and the trees from the 'syntactic lexicon' where tree schemata are stored. The lexical anchors, marked with $\Diamond$s, are numbered according to the list index of the word form that is to be inserted there.

The functional description remains the same, although it now allows reference to more remote nodes, and so instead of $\uparrow$ or $\downarrow$ I use node labels as a shorthand for

[10] A metagrammar (Candito 1996; Crabbé et al. 2013) is a formal system for describing generalisations both across and within grammars. For example, the fact that all transitive verbs will have tree families that contain many of the same trees can be captured by shared inheritance in a type hierarchy of the familiar kind.

$$\langle W \rangle = \langle \text{kicked}, \text{the}, \text{bucket} \rangle$$

$T =$



$$F = (\mathrm{S}_\phi \text{ TENSE}) = \text{PAST}$$

$$\vdots$$

$$\lambda x.\exists e[die(e,x)] : (\mathrm{S}_\phi \text{ SUBJ})_\sigma \multimap \mathrm{S}_{\phi\sigma}$$

Figure 2: TAG-LFG lexical entry for *kicked the bucket*

the nodes in question.[11,12]

Crucially, by complexifying c-structure in this way, we do not change the overall computational complexity of LFG. TAGs are mildly context sensitive, which makes them more powerful than context-free grammars, but it has been shown that LFGs in general are already more than mildly context sensitive, owing to the power of f-structure (Berwick 1982).[13]

3.3 Accounting for the idiom facts

In this subsection, we put to use the formalism just introduced, and demonstrate how the TAG-based approach to idioms is implemented.

Differences in syntactic flexibility can be represented in the different tree families which the idioms are related to. For instance, *kick the bucket* would not include any trees in its tree family beyond the simple active voice. If we think of tree families as types in a hierarchy, then *kick the bucket* only inherits from the active voice tree type. This level of granularity in types is needed because other words and idioms inherit different combinations of the basic tree schemata. Idioms like *spill the beans*, for example, are readily passivisable, but distinctly odd in questions or

[11] In reality, the node labels are not the nodes: they are the output of a node labelling function λ applied to each node (Kaplan 1995).

[12] In addition, since the functional descriptions must be resolved once all adjunctions and substitutions have taken place, we cannot see the trees as being manipulated derivationally by the operations of substitution and adjunction. Rather, we view the trees as *tree descriptions* (Vijay-Shanker 1992), which, together with the combining operations, license a set of derived trees which make up the grammatical sentences of the language in question. Cf. the notion of context-free grammar rules as 'node admissibility conditions' (McCawley 1968) already taken as standard in LFG.

[13] TAG-LFG is analogous to a feature-based TAG where recursive feature structures are permitted, which sets it apart from standard FTAG, e.g. that advocated by Vijay-Shanker & Joshi (1988), where such recursion is banned precisely in order to prevent FTAG from becoming intractable like LFG.

relative clauses:

(21) a. Jimmy Schementi spilled the beans back in August.
(http://bit.ly/2xKbtuh)
b. The beans were spilled back in August.
c. #The beans that Jimmy spilled back in August have caused problems for us.
d. #Which beans did Jimmy spill back in August?

Then there are verbs like *cost* which do not passivise, but can have their objects relativised on. (Object questions here are likewise dubious when in the form of a *which*-phrase.)

(22) a. The horses cost two thousand pounds.
b. *Two thousand pounds was/were cost (by the horses).
c. Emma [. . .] indignantly pledges to repay him the two thousand pounds that the horses cost.
(http://bit.ly/2xITrsb)
d. {What/#Which two thousand pounds} did the horses cost?

And of course there are regular transitive verbs and fully flexible idioms like *pull strings*, where all four possibilities are attested:

(23) a. We ate the rice and beans with delight.
b. The rice and beans were eaten with delight.
c. Our only reward then was rice and beans which we ate with delight.
(http://bit.ly/2yGktQ2)
d. What/Which rice and beans did you eat?

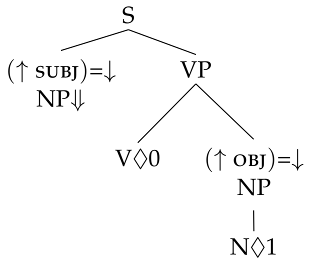
(24) a. We are pulling strings to find them jobs.
(http://bit.ly/2xIxSYO)
b. Strings were pulled in the US and Mexico to ensure this happened.
(http://bit.ly/2xJcORO)
c. Thanks to some strings we pulled with our partners, we're giving away 1000 gifts an hour.
(http://bit.ly/2xIarP7)
d. Which strings did he pull to visit Dreamworks?!
(http://bit.ly/2xIjKyt)

An articulated inheritance hierarchy of tree schemata can capture these different types of predicate, and so we can use the same tools to describe the different types of idioms. This is of course descriptive rather than explanatory, and it is possible there are semantic/conceptual motivations behind some of the restrictions, but I do not address this question here.

The internal modifiability of decomposable idioms can be achieved by simply associating more than one meaning constructor with their lexical entries, accounting for their internal modifiability. Figure 3 gives an entry for active voice *pulled strings*, including meaning constructors corresponding to the verb and its argument.

$\langle W \rangle \;=\; \langle \text{pulled}, \text{strings} \rangle$

$T \;=$



$$F \;=\; (\mathrm{S}_{\phi}\ \textsc{tense}) = \textsc{past}$$

$$\vdots$$

$$\lambda x.connections(x) : (\mathrm{N}_{\phi\sigma}\ \textsc{var}) \multimap (\mathrm{N}_{\phi\sigma}\ \textsc{restr})$$

$$\lambda x \lambda y.\exists e[exploit(e,x,y)] : (\mathrm{S}_{\phi}\ \textsc{subj})_{\sigma} \multimap (\mathrm{S}_{\phi}\ \textsc{obj})_{\sigma} \multimap \mathrm{S}_{\phi\sigma}$$

Figure 3: TAG-LFG lexical entry for *pulled strings*

The long-distance dependency facts fall out straightforwardly from the standard TAG approach. As noted, the presence of adjunction as a combining operation means that long-distance dependencies can be encoded locally in the lexicon. This is as true for sub-parts of idioms as it is for *wh*-dependencies and the like. Figure 4 gives an example for the relative clause-containing NP *strings Kim claimed Sandy pulled*. We start with the relative clause elementary tree for *strings. . . pulled*, and through adjunction of the embedding verb *claimed*, the parts of the idiom are separated. This could of course be repeated indefinitely.

Finally, the TAG-based approach also aligns with the psycholinguistic findings, as noted by Abeillé (1995). A parse involving an idiom will involve fewer elementary trees: in *Alex kicked the bucket*, for example, it will only involve the trees for *Alex* and for *kicked the bucket*, instead of the four trees *Alex*, *kicked*, *the*, and *bucket*. On the assumption that a simpler parse is faster, this makes sense of the increased processing speed found with idioms.

4 Conclusion

Idioms and other MWEs exhibit a tension between their phrase-like and word-like tendencies. Current work in lexicalist and other formal frameworks seems to be in favour of ignoring this tension and coming down entirely on one side of it, by treating idioms as phrases made up of special homophonous versions of the words they contain. I advocate an alternative, based largely on Abeillé's (1995) earlier work on idioms in French.

Part of the problem is that a context-free c-structure has too narrow a definition of locality to describe the relationship between the parts of idioms directly, and

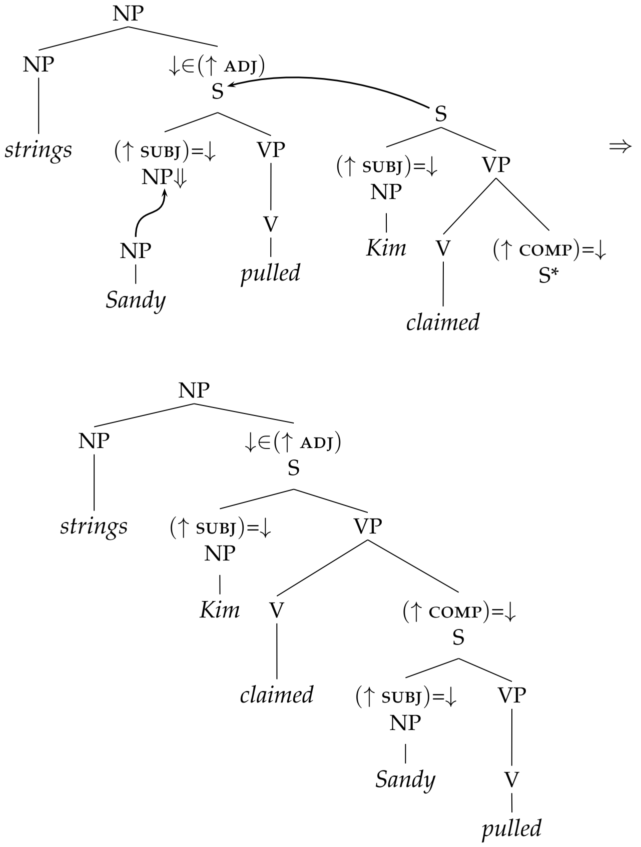


Figure 4: Derivation of *strings Kim claimed Sandy pulled*

such a relationship cannot easily be modelled at other levels of description either. By using a TAG instead, we can take advantage of the extended domain of locality this formalism offers, and also the operation of adjunction it provides: this makes it possible to describe the relationships between idiom parts locally, even if they are ultimately realised arbitrarily far apart. This allows us to describe each idiom in one place, in the lexicon, while still recognising its multiword status by associating it with more than one word form. Further work is needed to investigate the best way to develop a metagrammar which incorporates LFG annotations. It is possible that standard LFG c-structure rules might form the basis of such a metagrammar, thus offering a pleasing way to incorporate existing analyses into the new framework.

References

Abeillé, Anne. 1988. Parsing French with Tree Adjoining Grammar: some linguistic accounts. In *Proceedings of the 12th conference on computational linguistics*, 7–12. Budapest, HU.

Abeillé, Anne. 1995. The flexibility of French idioms: A representation with Lexicalized Tree Adjoining Grammar. In Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schreuder (eds.), *Idioms: Structural and psychological perspectives*, Hove, UK: Lawrence Erlbaum.

Abeillé, Anne & Owen Rambow (eds.). 2000. *Tree Adjoining Grammars: Formalisms, linguistic analysis and processing*. Stanford, CA: CSLI Publications.

Ackerman, Farrell, Gregory T. Stump & Gert Webelhuth. 2011. Lexicalism, periphrasis, and implicative morphology. In Robert D. Borsley & Kersti Börjars (eds.), *Non-transformational syntax: Formal and explicit models of grammar*, 325–358. Oxford, UK: Wiley-Blackwell.

Arnold, Doug. 2015. A Glue Semantics for structurally regular MWEs. Poster presented at the PARSEME 5th general meeting, 23–24th September 2015, Iaşi, Romania.

Asudeh, Ash. 2012. *The logic of pronominal resumption*. Oxford, UK: Oxford University Press.

Asudeh, Ash, Mary Dalrymple & Ida Toivonen. 2013. Constructions with lexical integrity. *Journal of Language Modelling* 1(1). 1–54. http://jlm.ipipan.waw.pl/index.php/JLM/article/view/56/49.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing* (2nd edn.), 267–292. Boca Raton, FL: CRC Press.

Berwick, Robert C. 1982. Computational complexity and Lexical-Functional Grammar. *American Journal of Computational Linguistics* 8. 97–109.

Butt, Miriam, Mary Dalrymple & Anette Frank. 1997. An architecture for linking theory in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG97 Conference*, Stanford, CA: CSLI Publications. http://web.stanford.edu/group/cslipublications/cslipublications/LFG/LFG2-1997/lfg97butt-dalrymple-frank.pdf.

Candito, Marie-Hélène. 1996. A principle-based hierarchical representation of LTAGs. In *Proceedings of the 16th conference on Computational Linguistics (COLING)*, 194–199. Association for Computational Linguistics. http://dx.doi.org/10.3115/992628.992664.

Chomsky, Noam. 1980. *Rules and representations*. New York, NY: Columbia University Press.

Crabbé, Benoît, Denys Duchier, Claire Gardent, Joseph Le Roux & Yannick Parmentier. 2013. XMG: eXtensible MetaGrammar. *Computational Linguistics* 39(3). 591–629.

Cronk, Brian C. 1992. The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics* 13. 131–146.

Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. http://corpus.byu.edu/glowbe.

Ernst, Thomas. 1981. Grist for the linguistic mill: Idioms and 'extra' adjectives. *Journal of Linguistic Research* 1(3). 51–68.

Estill, Robert B. & Susan Kemper. 1982. Interpreting idioms. *Journal of Psycholinguistic Research* 11(6). 559–568.

Falk, Yehuda N. 2010. An unmediated analysis of relative clauses. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG10 Conference*, 207–227. CSLI Publications. http://web.stanford.edu/group/cslipublications/cslipublications/LFG/15/papers/lfg10falk.pdf.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum & Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

Gibbs, Raymond W., Jr. 1986. Skating on thin ice: Literal meaning and understanding idioms in context. *Discourse Processes* 9. 17–30.

Joshi, Aravind K., Leon S. Levy & Masako Takahashi. 1975. Tree adjunct grammars. *Journal of Computer and System Sciences* 10(1). 136–163.

Kaplan, Ronald M. 1995. The formal architecture of Lexical-Functional Grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell, III, & Annie Zaenen (eds.), *Formal issues in Lexical-Functional Grammar*, 7–28. Stanford, CA: CSLI Publications.

Katz, Jerrold & Paul Postal. 1963. Semantic interpretation of idioms and sentences containing them. In *Quarterly progress report no. 70*, 275–282. Cambridge, MA: MIT Research Laboratory of Electronics.

Kay, Paul, Ivan A. Sag & Daniel P. Flickinger. 2015. A lexical theory of phrasal idioms. Unpublished ms., CSLI, Stanford. http://www1.icsi.berkeley.edu/~kay/idiom-pdflatex.11-13-15.pdf.

Lichte, Timm & Laura Kallmeyer. 2016. Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In Christopher Piñón (ed.), *Empirical issues in syntax and semantics 11*, Paris: Colloque de Syntaxe et Sémantique à Paris (CSSP).

Lowe, John J. 2015. Complex predicates: an LFG+glue analysis. *Journal of Language Modelling* 3(2). 413–462.

McCawley, James D. 1968. Concerning the base component of a transformational grammar. *Foundations of Language* 4(3). 243–269.

McGinnis, Martha. 2002. On the systematic aspect of idioms. *Linguistic Inquiry* 33(4). 665–672.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Potts, Chris. 2005. *The logic of conventional implicatures* (Oxford Studies in Theoretical Linguistics 7). Oxford, UK: Oxford University Press.

Pulman, Stephen G. 1993. The recognition and interpretation of idioms. In Cristina Cacciari & Patrizia Tabossi (eds.), *Idioms: Processing, structure, and interpretation*, 249–270. London, UK: Lawrence Erlbaum.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword Expressions: a pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Compuational Linguistics (CICLing-2002)*, 1–15. Mexico City, MX.

Sailer, Manfred. 2000. Combinatorial semantics and idiomatic expressions in Head-Driven Phrase Structure Grammar. Doctoral dissertation, Eberhard-Karls-Universität Tübingen.

Schabes, Yves, Anne Abeillé & Aravind K. Joshi. 1988. Parsing strategies with 'lexicalized' grammars: Application to Tree Adjoining Grammars. In *Proceedings of the 12th Conference on Computational Linguistics (COLING)*, 578–583. Association for Computational Linguistics. http://dx.doi.org/10.3115/991719.991757.

Swinney, David A. & Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior* 18. 523–534.

Vijay-Shanker, K. 1992. Using descriptions of trees in a Tree Adjoining Grammar. *Computational Linguistics* 18(4). 481–517. http://dl.acm.org/citation.cfm?id=176313.176317.

Vijay-Shanker, K. & Aravind K. Joshi. 1988. Feature structure based Tree Adjoining Grammars. In *Proceedings of the 12th Conference on Computational Linguistics (COLING '88)*, 714–719. Association for Computational Linguistics. http://dx.doi.org/10.3115/991719.991783.

Wittenberg, Eva & Maria Mercedes Piñango. 2011. Processing light verb constructions. *The Mental Lexicon* 6(3). 393–413.

Appendix

Arnold's (2015) approach to idiom composition uses manager resources to eliminate the need for many semantically inert forms, although it still requires ambiguity of the head word. For instance, idiomatic *kick* has the meaning constructor in (25):

(25) $\lambda x \lambda Q.\exists e[die(e,x)] : (\uparrow \textsc{subj})_\sigma \multimap [[(\uparrow \textsc{obj})_\sigma \multimap \uparrow_\sigma] \multimap \uparrow_\sigma] \multimap \uparrow_\sigma$

This consumes the meaning constructor for literal *the bucket*, which has the form given in (26), and discards the meaning.

(26) $\lambda P.the(b, bucket(b), P(b)) : \forall H[\uparrow_\sigma \multimap H] \multimap H$

In fact, it is possible to implement this at the phrasal level and in this way avoid having any lexical ambiguity (cf. Asudeh et al.'s 2013 approach to constructions). We associate a disjunction of idiom templates with the VP rule, including, e.g. Kick-the-Bucket:

(27) $$\text{VP} \quad \rightarrow \quad \begin{array}{c} \text{V}' \\ (\{@\textsc{Kick-the-Bucket}| \ldots\}) \end{array}$$

(28) $\textsc{Kick-the-Bucket} :=$

$(\uparrow \textsc{pred fn}) =_c \text{kick}$
$(\uparrow \textsc{obj pred fn}) =_c \text{bucket}$
$(\uparrow \textsc{obj spec pred fn}) =_c \text{the}$

$\lambda P \lambda y.\exists e[die(e,y)] : [(\uparrow \textsc{subj})_\sigma \multimap \uparrow_\sigma] \multimap (\uparrow \textsc{subj})_\sigma \multimap \uparrow_\sigma$

This consumes the meaning constructor for *kick the bucket* once it has been composed, and then returns the idiom meaning as a dependency on the subject.

Such an approach also allows an extension to decomposable idioms:

(29) $\textsc{Spill-the-Beans} :=$

$$(\uparrow \text{PRED FN}) =_c \text{spill}$$
$$((\uparrow_\sigma \text{ARG}_2)_{\sigma^{-1}} \text{ PRED FN}) =_c \text{bean}$$

$$\lambda P \lambda x \lambda y . \exists e[divulge(e, x, y)] : [(\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{ARG}_2) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{ARG}_2) \multimap \uparrow_\sigma$$

$$\lambda Q \lambda v . secret(v) : [(\uparrow_\sigma \text{ARG}_2\ \text{VAR}) \multimap (\uparrow_\sigma \text{ARG}_2\ \text{RESTR})] \multimap (\uparrow_\sigma \text{ARG}_2\ \text{VAR}) \multimap (\uparrow_\sigma \text{ARG}_2\ \text{RESTR})$$

However, this approach ultimately seems untenable, since it makes entirely the wrong predictions about modification (a point which Arnold 2015 notes): since the manager throws away the object's meaning, it can do this just as well before or after that meaning is modified, as it will correspond to the same Glue expression in either case. This predicts two things: (a) that modification should be possible in cases like *kick the bucket*, but simply have no effect on the meaning, and (b) that modification should be ambiguous in cases like *spill the beans*, either affecting the meaning or not, depending on the order of composition. Neither of these predictions is borne out: internal modification of *bucket* is not innocuous, but results in a loss of idiomaticity, as in (30), and interpreting internal modification in cases like (6), above, is not optional.

(30) #Sandy kicked the red/painful/sudden/... bucket.

A technical get out is available at least in the *kick the bucket* cases. As Arnold (2015) suggests, we can include the following constraint in the idiomatic head (or, equally, the template):

(31) $\neg(\uparrow \text{OBJ ADJ})_{\sigma_{\langle et,et \rangle}}$

This prevents the object having normal $\langle et, et \rangle$ modifiers, but allows expressive/emotive modifiers, as in (32), which are presumed to have a different semantic type (Potts 2005):

(32) Alex kicked the proverbial/bloody bucket.

This is purely stipulative, however, and, what is more, it doesn't help in any way with the internally modifiable cases, where such modifiers explicitly *are* allowed. The issue there seems to be much more fundamental, since there is no straightforward way to enforce a particular ordering on a Glue derivation, which is ultimately what is required.