“Keepwords”, the Limits of Creativity, and the Notion of the Core of an Idiom

CAREY BENOM

Kyushu University

YOUNG-MIN OH

Ritsumeikan Asia Pacific University

1 Introduction

Idioms are conventional, non-compositional multi-word expressions, a definition which encompasses everything from e.g. phrasal verbs such as *get up* to proverbs such as *Don’t count your chickens before they hatch*. Speakers employ a broad range of creative variation with idiomatic expressions.

In our investigation of idiomatic creativity (Benom 2023, Benom and Oh 2020, Oh 2018, 2020a, 2020b, 2022, Oh and Benom 2020, 2021), we have studied tens of thousands of corpus-derived creative uses of idioms in Korean, Japanese, and English (other corpus-based approaches include Langlotz 2006, Moon 1998, Tsuchiya 2013, and Wulff 2008). The impressive breadth of variation we observed led us to wonder if all components of an idiom are

Japanese/Korean Linguistics 30
Edited by Sara Williamson, Adeola Aminat Babayode-Lawal, Laurens Bosman,
Nicole Chan, Sylvia Cho, Ivan Fong, and Kaye Holubowsky.

potentially up for grabs, or if there are certain privileged keywords, a 'core' of the idiom, that is always present. Therefore, this paper reports on our attempt to address these questions: Q1) Does an idiom have a core? and Q2) If so, what is the nature of this core? Based on our inductive approach, we use an operational definition that asks what the relevant corpus data show us. Of particular interest is whether a core is comprised of all the content words in the idiom, or some subset thereof, or if it includes both grammatical and content terms (see Talmy 2000 on the importance of the distinction between the grammatical and lexical subsystems).

We will also ask a related but larger question: Q3) What are the limits of idiomatic creativity? In other words, at their most creative, how far can speakers push it? How creative is too creative? (Q3) is especially difficult to address with clarity, completeness, and empiricism, and therefore, among the answers we can offer here, those for (Q3) will be the most preliminary.

The structure of the paper is as follows: after providing some background below, which will lead us to consider one further question and to make five predictions, we will describe our methodology in Section 2. Section 3 will present and discuss our results, and Section 4 gives our conclusions.

1.1 Background

Idioms hold a unique place in the history of linguistic theory. They don't fit neatly into "building block" or "words plus rules" approaches to language, since they are neither words nor grammar, but share aspects of both (Croft and Cruse 2004). For this reason, they played an essential role in the birth of construction grammar (e.g. Fillmore et al. 1988).

Idioms are famous for their fixedness or inflexibility (Nunberg et al. 1994, Langlotz 2006:5), which is sometimes taken as definitional. After all, some canonical form must be entrenched for speakers to treat these multiword expressions as single units, and to serve as the basis from which creativity can be employed.

Yet, speakers do have flexibility, and recently, a broad range of idiom uses displaying impressive creativity has been the subject of an increasing number of investigations, including those cited in the second paragraph of this paper. Given this robust variation, the question of whether an idiom has keywords or a core is well-founded, though it turns the traditional mistaken assumption that idioms are all strongly fixed on its head, asking if anything *at all* is fixed.

But the terms *keywords* and *core* need to be defined carefully. In fact, we can envision two contrasting definitions. In the first, which we adopt, keywords are elements in the idiom which are (essentially) always present, regardless of how far speakers go with their creativity. These keywords, and potentially the larger syntactic structure, form the idiom's core.

The second definition is that the core is comprised of keywords which are so salient that they needn't be mentioned. Their meaning is implied or the meaning of the idiom as a whole is unchanged when used without them.

We employed the first definition because we wanted to know if an idiom has a "backbone" or a "foundation" that is necessarily present to permit creative uses. Therefore, we refer to the keywords that we are studying as *keepwords*, which we define as lexical items in the canonical form of an idiom that are essentially always present, even in the most creative uses. Speakers can delete or replace other terms, but not (usually) keepwords.

This brings us to an important point. There are two distinct types of scenarios in which lexemes from the canonical form can be missing in a creative use. We refer to them as contraction (hereafter CON) and substitution (hereafter SUB). We will exemplify them with creative uses of the idiom *Don't count your chickens before they hatch*.

(1) CON: *Don't count your chickens*.
(2) SUB: *Don't count your chickens before the bond gets passed.*

In the first example, a contracted form of the idiom is used, and *before they hatch* is missing. In the second, a clause is substituted into the idiom, and *they hatch* is missing. We treat CON and SUB as two distinct types of idiomatic creativity.

In previous work (Benom and Oh 2020, Oh and Benom 2021), we argued that most idiomatic creativity is motivated by the complexity of idioms' multi-layered and figurative semantics, and by speakers' need to define reference, given this web of interconnected meaning. We refer to "referential specification", which we define as the creative use of language in order to ground the abstract, figurative meaning of an idiomatic expression in the relatively concrete context to which it is applied by specifying information about reference. We note as crucial the fact that speakers' creativity is constrained by the requirement that the hearer must recognize that the production is intended as an instantiation of the idiom. This is, with both CON and SUB, the biggest reason why there are limits on idiomatic creativity. So we now ask in Q3 *What do these limits look like in our data?*

However, we will limit the scope of the present study to examining the lexical level, because that is a sufficiently complex (and rewarding) area of study. By studying the variation of most grammatical terms (cf. Tsuchiya 2013 and other corpus-based approaches, which primarily focus only on content terms), we hope to shed light on the syntax involved. However, the scope of this study prevents us from considering the syntactic structures which defy variation, and so we won't claim to understand the core (Q2) at a syntactic level. Fully addressing (Q3) will also require investigation of syntax, and on the limits, if any, on what types of elements can be added to an idiom. We intend to leave these issues to future research.

At this point, we will present one final question, about which we will make five predictions: Q4) Is there a difference between the terms that are not easily deleted through CON and those that are not easily replaced through SUB? First, we predict (P1) that there should be, based on two differences between the contexts.

The first difference is that CON is frequently conventionalized variation (as in (1)), while SUB often involves true creativity (i.e. the making of new meaning, as in (2)). The second difference is that, with SUB, a substituted term is a clue to the identity of the idiom. For one thing, it is nearly always the same part of speech. Furthermore, Benom (2023), a study of SUB in heart idioms in English and Japanese, finds just five types of semantic relationships between the canonical item and its lexical substitute. Speakers should be able to take advantage of these affordances in both syntax and semantics. With more clues, based on the requirements for creative idiom use we described, speakers have more leeway to employ creativity with SUB, and so we predict that they should employ CON and SUB differently.

Based on this argument our second prediction (P2) is that, since having more clues to the identity of the idiom available means more creative potential, short idioms and long idioms should behave differently with respect to CON vs. SUB.

Our third prediction (P3) is specific to CON: based on the principle of economy, keepwords should include the fewest lexemes sufficient for hearers to recognize the idiom. This means that keepwords should not co-occur frequently outside the idiom. The juxtaposition of a small number of crucial elements is all that the speaker needs to signal the idiomatic use, but a minimalistic approach won't work if the items are often used together in other contexts. This means that we are expecting results roughly like those in (1), rather than e.g. *don’t count* or *chickens hatch* being the keepwords. In addition, we predict (P4) that content words should be the strongest keepwords under CON, as they have lower frequency than grammatical words, and therefore co-occur less often, and they also have more specific and richer semantic content, making it likely that idioms are generally identifiable through their content words.[1]

Our final prediction (P5) is specific to SUB: speakers should employ substitutes for the terms that can give them maximum benefits, which, as we have argued (Benom and Oh 2020), are most commonly based on assisting

[1] For idioms that have (essentially) no flexibility, all words are keepwords, and therefore this only applies to flexible idioms.

with referential clarity. Therefore, utility in assigning reference should motivate most substitutions. We will assess to what extent our predictions hold true in Section 3 below.

2 Methods

Our operational definition of idiomatic creativity is based on variation from the canonical form, which we defined by employing as many idiom dictionaries as possible for each language (following the principle of *majority rules* when conflicts arose). We included case markers, but ignored the type of omission which is frequent in Korean and Japanese conversation. We ignored inflectional morphology such as tense, aspect, and person marking on verbs.

Our data come from in-depth studies of eight idioms each in Korean and English, and nine in Japanese,[2] using the largest corpus of each language we could find. For English, we used the enTenTen15 (13 billion words), for Korean, the koTenTen18 (1.7 billion words), and for Japanese, the jaTenTen11 (8.4 billion words). We tried to extract all idiomatic variation from the corpus in each case by searching for all terms in an idiom two at a time (e.g. *count* and *hatch*), within a span of 10 words of one another (5 on each side), and manually filtering the results. For both Korean and Japanese, we were successful, but the English corpus was the largest, and when we got more than 1000 results, we limited ourselves to analyzing the first 500.

Using web data is less than ideal for several reasons. Maybe the most crucial is that language use on the web is certainly different than spoken language, and therefore any conclusions we make cannot be assumed to be true for spoken language in general. But the size of the corpora is what allowed us to collect so many relevant uses. In fact, even with our huge corpora, data sparsity was still a problem for some idioms.

Previously, we had coded all data based on whether any lexical element was added, or contracted (CON), or substituted (SUB), vis-à-vis the canonical form, and whether the use shows variation in syntax. For this paper, we collected all data with the 25 idioms that was coded as CON (total = 12,498 uses for 3 languages) and SUB (total = 4,284 uses for 3 languages). Note that some uses involve both CON and SUB, and thus are present in both data sets.

[2] We studied a range of idioms in all three languages, and made efforts to match idioms by number of content words and number of morphemes, but we had to make compromises due to data sparsity. Additionally, our inductive approach meant that we allowed the results to dictate that we should distinguish the short and long idioms based on their behavior. The clear cutoff line which emerged (see Section 3) meant that two idioms each from Korean and English were included among short idioms, but just one from Japanese, and that one returned few data, so we added another short Japanese idiom. This means that the group of longer idioms contains six each from Korean and English and seven from Japanese.

3 Results and Discussion

Here, we will first briefly describe the larger trends in the data, and then examine the results for short idioms before we discuss the longer idioms in more detail. To balance depth and breadth given limited space, we will detail the results for approximately half of the idioms we studied.

The big picture: 1) Most idioms had a clear core in most environments, and all had one or more strong keepwords (defined below). Even idioms returning copious data had strong keepwords. 2) Context made a huge difference for longer idioms. Of the 6 short idioms investigated all had strong keepwords, and only one clearly varied by context. As for the longer idioms (19 total; see fn. 2), all had a clear core consisting of one or more strong keepwords in CON or SUB or both. Most (12/19) had different (sometimes overlapping) keepwords in the two contexts, while others (5/19) had one or more keepwords in one context, but none in the other. This means that 17/19 had a core *for a particular context*, showing the striking difference between CON and SUB. Just two longer idioms had the same keepwords in both contexts (like most of the short idioms). Our predictions were upheld in nearly all cases, and we will mention the few exceptions.

As for short idioms: our data showed that idioms with at most two content words and four morphemes (we allowed the results to determine this cutoff point) have far less creative potential and stronger keepwords which don't usually vary by context. Let's begin with this Korean idiom:

(3) *olibal-ul* *naymil-ta*
duck.feet-ACC[3] stick.out-DEC
Lit. 'duck feet are sticking out.'
Fig. 'play innocent, feign innocence, pretend not to know'

Here, all numbers represent the raw frequency of uses in which the form is absent in our corpora of variation. With CON and SUB combined: **olibal**(0) [-ul(33) naymil-ta(20)] (149) Tot:202. To explain, *olibal* 'duck feet' is never absent, and hence a strong keepword. The accusative marker *-ul* and *naymil-ta* 'stick out' were both missing, together, in 149 uses, and each was absent by itself 33 and 20 times, respectively. The total of 202 represents the number of uses in which any element of the idiom was missing. In some creative uses, multiple terms are absent, and because we use brackets sparingly, to preserve clarity of presentation (and because non-contiguous elements are sometimes

[3] Abbreviations used: ACC = accusative, ATT = attributive, DEC = declarative, GEN = genitive, LOC = locative, NEG = negative, NOM = nominative, PFV = perfective aspect, POT = potential, PRS = present tense, TOP = topic.

absent), the total number of uses in our corpus is not necessarily equal to the sum of the absences of each lexeme in the representation.

Arbitrarily, we defined the strongest keepwords as those absent in less than 5% of relevant creative uses,[4] and represent them in bold and underlined, as we did with *olibal* 'duck feet'. Weaker (potential) keepwords, absent in 5-10% of relevant uses, will only be underlined. We found few of this second type,[5] suggesting that we are capturing a real phenomenon, based on speakers' behavior, rather than merely confirming our pre-theoretical notion.

As with most other short idioms, distinguishing CON and SUB doesn't reveal much; CON: **olibal**(0)-[ul naymil-ta](149) Tot: 149; SUB: **olibal**(0)-ul(33) naymil-ta(20) Tot: 53. One final point to make about this idiom core, is that, under CON, the entire meaning of the idiom has been adopted by the word *olibal* 'duck feet' in its lexical semantics, and thus this seems to be a case of idiom wordization (Moon 1996). This is also true of the other short Korean idiom we studied. For clarity, henceforth, we will incorporate our results at the bottom of the presentation of the idiom itself. Here we show the combined results for CON and SUB.

(4) *ojilap* *-i* *nelp -ta*
the.front.part.of.an.outer.or.upper.garment-NOM wide-DEC
Lit. 'the front part of an outer or upper garment is wide.'
Fig. 'to be interfering, to be nosy'
CON + SUB: **ojilap**(0) [-i(167) nelp-ta(153)](1908) Tot:2228.

Data was plentiful, and the keepword's vigor, noteworthy. Substitutions for the adjective *nelp-ta* 'be wide' included *pwuli-ta* 'act, behave', in which *ojilap* metonymically refers to the full meaning of the idiom, and the substituted term helps to refer to someone acting in a nosy way (P5).

One short English idiom showed limited variation; (CON + SUB): **bite**(0) the(67) **bullet**(0) Tot:67. The other returned just 2 contractions, so we present the results for SUB only: beat(5) around (1) the(10) **bush** (0) Tot:16.

Data sparsity was also a problem with the Japanese *tedama-ni toru* 'beanbag-LOC take', Lit. 'treat like a beanbag', Fig. 'have someone in the palm of your hand'; (CON + SUB): **tedama**(0) -ni(3) toru(8) Tot:11. This prevents us from drawing any strong conclusions based on these results, as with *beat around the bush.* No other Japanese idioms ended up in the shorter group, so we added the following idiom:

[4] By "relevant creative uses", we refer to the total number of contracted uses of that idiom when we are considering each potential keepword in the context of CON, total SUB when we are considering SUB, and the combined total when we are considering all absences combined.

[5] We found 67 strong keepwords and 15 weaker (potential) keepwords total for all 25 idioms.

(5) *tana kara bota-mochi*
shelf from peony-rice.cake
Lit. 'Peony rice cake from the shelf'
Fig. 'have a stroke of unexpected good luck', 'pennies from heaven'
CON: [**tana**(0) kara(2983)](16) **bota**(0)-mochi(2972) Tot: 2999
SUB: tana(7) kara(4) botamochi(5) Tot:13

Our purpose was to try to gain some insight from additional data. This turned out to be extremely fortunate, since this idiom doesn't behave like any of the other short ones. The remarkable gap in the amount of data between CON and SUB is a result of the conventionalized contraction *tana bota* ('shelf peony'). This is interesting for several reasons. To begin, *bota* is not a free morpheme. It's bound on the right, but in *tana bota*, it's attached on the left. What's more, the kanji can be used in isolation to refer to peony flowers, but they are pronounced *botan* in that case. Speakers have combined the reduced form with the first content word in the idiom to create a compound word which, in its lexical semantics, bears the entire idiomatic meaning, and thus this seems to be yet another case of idiom wordization (Moon 1996). It also means that we have two content elements serving as strong keepwords. This fits (P3), except that one is a bound morpheme, rather than a lexeme. With SUB, we have no keepwords, despite limited data, but those few substitutions we did find were useful in assigning reference (P5) to both the source (*tana* 'shelf'; e.g. *hako* 'box') and the specific realization of the good fortune (*botamochi* 'peony rice cake' had substitutes such as *baritou* 'Bali'), and even both at the same time, but only after giving proper context by explicitly mentioning the beginning of the idiom:

(6) *tana-kara nara-nu reizouko-kara matcha-aisu*
shelf-from become-NEG fridge-from green.tea-ice.cream
Lit. 'not becoming from the shelf, green tea ice cream from the
Fridge'
Fig. 'lucky green tea ice cream from the fridge'

This gives us some clues about (Q3) which fit and support our earlier claims: speakers can substitute for almost the entire idiom, as long as they make sure to cite the idiom first, so that it is activated.

In summary, the two short Korean idioms we investigated had just one extremely strong keepword each, regardless of context, and despite copious data, and we believe that they are cases of idiom wordization. We saw something similar with the Japanese idiom in (5), as a compound-noun was created from the idiom through CON, though in that case, SUB showed different behavior. We speculate that most 2-content word idioms will include both of their content words as keepwords, and that these idioms are outliers, but more data is needed, and this is beyond our present scope. Short idioms' behavior mostly did not match (P1), though the short vs. long contrast itself was predicted by (P2). All keepwords were content terms, including those for CON (P3, P4). (P5) was supported when data was sufficient.

Now we will look at the results for longer idioms (3 or more content words, or 2 content words and 5 morphemes). We will begin with the Japanese idiom in (7) below.

(7) *me-kuso hana-kuso o warau*
eye-shit nose-shit ACC laugh
Lit. 'eye boogers laugh at nose boogers.'
Fig. 'the pot calling the kettle black'
CON: **me-kuso**(2) **hana-kuso**(0) [o warau(1)](2258) Tot:2261
SUB: me-kuso(11) hana-kuso(10) **o**(0) warau(44) Tot:58

With CON, the first two content words, which don't often co-occur elsewhere, are strong keepwords (P3, P4). With SUB, the only keepword is the accusative marker *o*, and the first two content terms are replaced with e.g. *ma-guso* 'horse shit' and *mimi-kuso* 'ear wax', which involve wordplay, but also *ningen* 'human' and *koumuin* 'civil servant', which helped speakers with referential clarity, and therefore are cases of referential specification (P5).

Keepwords under CON were nearly all content words (P4), whereas with SUB, grammatical morphemes were often among the strongest keepwords, as in the next two English examples.

(8) CON: **beggars**(0) can(31) not(33) be(8) **choosers**(0) Tot:72
SUB: beggars(29) can(11) not(5) **be**(1) choosers(21) Tot:67
(9) CON: Put(188) the(133) **cart**(0) **before**(6) the(70) **horse**(4) Tot: 230
SUB: Put(212) the(24) cart(119) before(30) **the**(9) horse(105) Tot: 368

In (8 CON), the only two content words, which infrequently co-occur outside the idiom, are strong keepwords (P3, P4). In (8 SUB), these same words are most frequently replaced, and *be* is the only strong keepword (*not* is a weaker one). Lexical replacements proved versatile, with e.g. an invitation to a lecture on investment practices called *Borrowers can be Choosers*, which uses SUB as an aid in reference (P5). Despite only containing two content words, the behavior of the idiom matches that of longer idioms, such as (9), where, again, content words are strong keepwords under CON (P4), but this time *before* expresses the relationship between the two. The fact that *cart* and *horse* can co-occur in non-idiomatic contexts motivates the presence of this grammatical lexeme in the core; it is necessary to evoke the idiom. Again, it is not the frequently co-occurring *put the cart* that are keepwords here, just as we predicted (P3). As for SUB: *the* is the only strong keepword, but there are two weaker keepwords, which, just as we saw in (9 SUB), suggest that speakers are preserving a skeleton of the larger structure, so they can be creative with the other elements. Substituting for *cart* by using wordplay such as *cartel* or *Descartes* was popular, but so were examples like *putting the renewable energy bandwagon before the cart,* and in most cases, the replacement helped with reference (P5), apart from a few examples of a dialectal difference which we coded as SUB (*carriage* for *cart*). Similar behavior was observed for the Japanese idiom in (10).

(10) *sendou ooku-shite fune yama-ni noboru*
captain many-do boat mountain-LOC climb
Lit. ‘(If you) make many captains, the ship will climb the mountains.’
Fig. ‘Too many cooks spoil the broth.’
CON: [**sendou**(3)**ooku-shite](**5)
[fune(13) [[yama(1) ni](26) noboru(2)](2)](167) Tot: 219
SUB: **sendou**(1) ooku-shite(138)
[**fune**(6) [yama(6) ni(13) noboru(34)](21)](3) Tot: 222

With CON, speakers preserve the first clause, which consists of two content words (P4), which, considering that these words seem unlikely to co-occur outside the idiom, is enough to trigger the idiom (P3), but with SUB, they instead keep the first noun of each clause, ‘captain’ and ‘ship’, elements which frequently co-occur outside the idiom, allowing them to achieve two objectives: 1) to preserve the structure as a whole, as a kind of skeleton, and 2) to permit reference to both cause (*ooku-shite* ‘make many’) and effect (*yama-ni noboru* ‘climb the mountains’) (P5).[6]

The larger grammatical structure is also implied by the keepword with SUB in this next Japanese idiom:

(11) *kare ki mo yama no nigiwai*
dry tree also mountain GEN bustle
Lit. ‘Dead trees are also (part of) the bustle of the mountain.’
Fig. ‘It’s better to have something boring than nothing at all.’
CON: [kare ki mo](7) yama(11) no(15) **nigiwai**(1) Tot: 34
SUB: kare(11) ki(18) mo(8) yama (11) **no**(1) nigiwai(8) Tot: 46

[6] Lack of space prevents us from showing examples of these, but in translation, e.g. ‘the ship will go to the mountains’ is replaced with ‘disagreements may drag on until the end’.

With CON, the only keepword is a content word (P4), but such extreme minimalism wasn't necessarily predicted by (P3). With SUB, a creative speaker can replace any of the content words with a substitute, but tends to preserve the grammatical structure of the genitive NP by keeping the genitive marker *no*. We see replacements that help in assigning reference (P5), including, for *kare ki* 'dead trees', *gareki* 'debris' and *gomi* 'trash'. Again, however, the lack of data makes us hesitant to come to any strong conclusions.

The final Japanese idiom we can discuss here is presented below.

(12) *se* *ni* *hara* *wa* *kaerarenai*
back LOC stomach TOP change.POT.NEG
Lit. 'You can't replace your back with your stomach.'
Fig. 'You can't make an omelet without breaking eggs.'
CON: **se**(0) **ni**(0) **hara**(1) wa(80) kaerarenai(50) Tot: 131
SUB: se(24) ni(27) hara(8) wa(97) kaerarenai(39) Tot:195

In (12), notice the relatively frequent substitution for elements that are strong keepwords under CON, showing the benefits of having a substituted lexeme as a clue to the idiom – just the type of results we predicted (P1). With CON, the first two content words, and the grammatical marker linking them, are keepwords. This is motivated by the fact that the two content words co-occur in other contexts, but with the locative marker *ni* linking them, they always instantiate the idiom (P3, P4). Substitutes for *se* ('back') included references to real-world sacrifices, *kishu* 'model' (e.g. of a phone) and *yume* 'dream' (P5).

Everything we have described to this point speaks to the limits of idiomatic creativity (Q3), but we would like to mention two common patterns in our data which reveal some more of those limits. The first is THIS OR THAT. One example can be found in the following results. CON: Penny(24) wise(25) and (292) pound (78) foolish (80) Tot: 297. In this case, there are no keepwords, since speakers either contract the idiom to *penny wise* or to *pound foolish*. The latter is used metonymically to express the full idiomatic meaning, but the former is used with a positive sense, to express only that part of the idiomatic meaning with which the lexemes correspond (see the typology of idioms in Nunberg et al. 1994, and see Oh and Benom 2021 on types of idiomatic contraction).

The second pattern we find is THIS AND THIS OR THAT, such as in the SUB results in (11) above, in which the genitive marker and any of the other lexemes (and a lexical substitute) were sufficient to activate the idiom. The content words were replaced one at a time, showing the great flexibility speakers had. THIS AND THIS OR THAT is also found in the results for SUB shown for the Korean idiom below.

(13) *paltung -ey pwul -i tteleci-ta*
top.of.the.foot-on fire -NOM fall-DEC
Lit. 'fire falls on the top of the foot'
Fig. 'be pressed for time, be in urgent need'
SUB: **paltung**(7)-ey(353)[pwul(105)-i(67)tteleci-ta(113)](1) Tot:522

Observe the number 1 just before the total. It means that speakers either substitute for *pwul-i* or *ttelecii-ta* without a problem, but not both. If you replace all of *pwul-i tteleci-ta* there won't be enough clues left to understand that it is a use of the idiom – even with the strong keepword. So, the strategy we see is that speakers keep the first content word (*paltung*) plus either of the other two content words/ phrases. The number 1 before the total represents an either/or meaning that is even more essential to the core of the idiom than the strong keepword, given their respective totals. With respect to (P5), the substitutes employed helped speakers with reference, including, for *pwul* 'fire' (the urgent issue), *kumyungwiki* 'financial crisis', and for *tteleci-ta* 'fall' (which refers to the problem "arriving" or becoming reality), *heykyelhata* 'solve' (referring to the resolution of the urgent issue).

The final Korean idiom space permits us to present is in (14) (cf. (7)).

(14) *ttong mwut-un kay-ka kye*
dung be.smeared.with-ATT.PFV dog-NOM chaff
mwut-un kay namwula-n-ta
be.smeared.with-ATT.PFV dog speak.ill.of-PRS-DEC
Lit. 'the dog smeared with shit scolds the dog smeared with chaff'
Fig. 'the pot calling the kettle black'
CON: ttong(5) mwut-un(5) kay(14)-ka(57) kye(30)
mwut-un(30) kay(35) namwula-n-ta(58) Tot: 65
SUB: ttong(111) **mwut-un**(3) kay(61)-**ka**(4) kye(104)
mwut-un(3) kay(53) namwula-n-ta(69) Tot: 220

With CON, the first two content words are weak keepwords. This is generally consistent with (P3, P4), given that they don't frequently co-occur elsewhere, but regarding their lack of strength as keepwords, we suggest that this flexibility is attributable to the idiom's length (seven content words), which simply provides speakers with more resources. With SUB, the structure of the whole is preserved by the repetition of *mwut-un* with the nominative marker in between the two uses, allowing speakers to creatively substitute for other elements. The first word (*ttong*) is a taboo term, and therefore most substitutes were based on the principles of taboo avoidance, the most common replacement being *mwe* 'something'. This was not in our predictions

(P5), but it is well-motivated. Speakers also primarily used *mwe* ‘something’ to replace *kye* ‘chaff’; we would speculate that it is a way of preserving the larger, parallel, structure, but we admit that this is a post hoc explanation. Most other SUB was motived by referential specification, however, as speakers replaced either use of *kay* ‘dog’ (e.g. with *Naver*, a Korean IT company), or the specific action of the first dog (the verb *namwula-n-ta*, replaced by e.g. *sengnayta* ‘be angry at’).

As our final English idiom, we will show the results for the idiom discussed above in (1–2).

(15) CON: do(329) not(263) **count**(9) your(71) **chickens**(7) before(198) they(207) hatch(195) Tot: 348[7]
SUB: do(114) not(35) **count**(1) your(291) chickens(39) before(28) they(44) hatch(36) Tot: 369

With CON, speakers use the first two content items (P4) as keepwords, and these are sufficient to activate the idiomatic meaning (P3). With SUB, the verb is the anchor, allowing speakers to use substitutes to refer to a wide range of roles (P5), including the one who is possibly assuming too much (*your*, replaced with e.g. *its, their*), what that they are relying on (*chickens*, replaced with *conspiracies, rate hikes, teachers*), and the specifics of their potential fruition (*hatch*, replaced in e.g. … *before he is in custody, before the bond gets passed*).

At this point, we will summarize how our predictions fared. (P1) stated that CON and SUB would show different results, and it was strongly validated. (P2) said that short and long idioms would behave differently, and they did, though there were 3 idioms of the 25 that did not fit the behavioral tendencies of their length. (P3), for CON, was that keepwords should not co-occur frequently outside the idiom. It turned out that this principle is precisely what is needed to motivate the patterning with longer idioms, in addition to the tactic of “use the first content words”.

For instance, consider *put the cart before the horse*, in (9 CON) above. If speakers were to keep only the first few words in a contraction, it would be insufficient to predict the idiom, since *put* and *the* and *cart* are frequently used together in various contexts. So, speakers use the 2nd and 3rd content words, but even that is not sufficient without the preposition. *Cart*, *before*, and *horse* used together (and in that order) are strong predictors of the idiom. In this and other cases, if the keepwords are the first content words, they don’t

[7] To re-iterate, this total refers to the total number of *uses* with CON. Absences are represented here without brackets for clarity, and since a single use can include multiple contracted elements, the total of the individual items does not equal the total number of uses.

co-occur frequently outside the idiom, and if keepwords aren't the first few content words of an idiom, we can motivate them in this way, showing that (P3) was also strongly affirmed. (P4) was another great success, since nearly all keepwords under CON were content words, and exceptions seemed well-motivated based on (P3). As for (P5), we showed that the majority of substitutions were motivated by referential specification, apart from a small number of cases, some of which were motivated by other, clear factors (such as taboo avoidance). Our five predictions were all upheld.

4 Conclusions and Looking Ahead

Here, we began by asking (Q1) if an idiom has a "core", and (Q2) what that core looks like in our data. All 25 idioms we studied showed a clear, strong core, but not in all contexts (CON vs. SUB). All 25 had one or more strong keepwords (as defined in Section 3) in at least one context. Most short idioms had strong keepwords that did not change based on context. Of the 19 longer idioms, just two showed the same type of pattern, with the same, strong keywords in both CON and SUB, while 12 had different (sometimes overlapping) strong keepwords for the different contexts, and 5 had one or more strong keepwords in one context only. As for (Q2), everything in the paragraph below on (Q3) below, and indeed, all of Section 3, is also part of our answer, but we cannot yet claim to understand the syntax of the core.

We looked at (Q3) the limits on idiomatic creativity, and we gained insight into many facets of the answer, including a) that keepwords are a real phenomenon, and that idioms, overall, have a core of the type we defined, b) the key role played by the requirement to maintain sufficient clues to invoke the idiom, which we showed works differently with CON (with keepwords being key content words which infrequently co-occur elsewhere) and SUB (with keepwords that preserve the syntactic skeleton of the whole, including many grammatical terms), including the strategy of citing the idiom in order to substitute for all content words, and c) the relevance of referential specification with SUB, as a motivation to shape the limits of what speakers can do, and need to do. Finally, d) we observed patterns such as THIS OR THAT and THIS AND THIS OR THAT. A fuller answer for (Q3) awaits future research.

Our predictions, repeated here, were almost entirely accurate, although there were some exceptions: (P1) there was a difference between CON and SUB, but mostly for longer idioms; (P2) short and long idioms did behave differently, with just 3 exceptions among the 25 idioms; (P3) with CON, the fewest lexemes sufficient to recognize the idiom were the keepwords; (P4) with CON, keepwords were content words; (P5) with SUB, non-keepwords were those most helpful with referential specification. We found that these predictions, along with the principle of using the first content words of an idiom, motivated nearly all the patterns in our data.

References

Benom, C. 2023. Argentina, Eat your Cows Out! Lexical Substitution in English and Japanese Heart Idioms. *Embodiment in Cross-Linguistic Studies*, eds. J. B. Kóczy and K. Sipocz, 369–91. Leiden/Boston: Brill.

Benom, C. and Y. M. Oh. 2020 (Nov 26.) Don't Count your Commodities Investments before they Hatch: Idiomatic Creativity, Concretization, and the Comprehensibility of "Impossible" Collocations. Presented at *The 56th Linguistics Colloquium*, online.

Croft, W. and D. A. Cruse. 2004. *Cognitive Linguistics.* Cambridge: Cambridge University Press.

Fillmore, C. J., P. Kay and M. C. O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64:501–38.

Langlotz, A. 2006. *Idiomatic Creativity*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Moon, K. H. 1996. The Generation and Extinction of Idiomatic Expressions. *Journal of Korean Linguistics* 28:301–33.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. New York: Oxford University Press.

Nunberg, G., I. A. Sag and T. Wasow. 1994. Idioms. *Language* 70(3):491–538.

Oh, Y. M. 2018. Adding Insult to Spilled Coke: A Corpus-based Study of Idiomatic Creativity. *Proceedings of the Japanese Cognitive Linguistics Society* 17:94–106.

Oh, Y. M. 2020a. A Systematic, Corpus-Based Methodology for Studying Idiomatic Creativity in Korean and Japanese. *Japanese/Korean Linguistics* 27:87–103.

Oh, Y. M. 2020b. A Corpus-based Study of Idiomatic Creativity: A Case Study of "me kara uroko ga ochiru" and "oya no sune o kaziru". *KLS Selected Papers* 2:103–18.

Oh, Y. M. 2022. A Web Corpus Driven Study of Idioms and their Variations in English, Japanese, and Korean from a Cognitive Linguistic Perspective. Doctoral dissertation, Kansai University.

Oh, Y. M. and C. Benom. 2020. A Systematic, Corpus-Based Methodology for Studying Idiomatic Creativity in Korean and Japanese. *Japanese/Korean Linguistics* 27:87–103.

Oh, Y. M. and C. Benom. 2021. Contraction as Idiomatic Variation. *Japanese/Korean Linguistics* 28:361–75.

Talmy, L. 2000. *Toward a Cognitive Semantics*. 2 vol. Cambridge, MA: MIT Press.

Tsuchiya, T. 2013. Teikeihyōgen o Kiban to Shita Gengo no Sōzōsei: Kanyōhyōgen to Kotowaza no Kakuchōyōhō ni Kansuru Shakai/ Ninchiteki Kōsatsu (Language Creativity Based on Formulaic Language: Extensional Usage of Idioms and Proverbs from a Social Cognitive View). Doctoral dissertation, Kyoto University.

Wulff, S. 2008. *Rethinking Idioms: A Usage-based Approach*. London/New York: Continuum.