Words in Combination: Idioms and Composition

Francis **Bond**

Department of Asian Studies,
Palacký University, Olomouc, Czechia

<bond@ieee.org>

DAS {4|5}UJ2 2026

Faculty of Arts

Roadmap

Faculty of Arts

Defining Meaning

- When we use a word, we don't have to know everything about the referent
  - A *dog-cart* is a kind of **`cart`**
  - $\Rightarrow$ you can ride it
  - $\Rightarrow$ it has wheels
- We infer that it has many of the same properties as its **hypernym**, even though it may not have all
  - A *hover-car* is a kind of **`car`**
  - $\Rightarrow$ you can ride it
  - $\not\Rightarrow$ it has wheels
- Many of the properties may be irrelevant to the story at hand, and irrelevant to the syntax of the language

Faculty of Arts

How do we learn?

You shall know a word by the company it keeps

(Firth, 1957, p11)

- You see a new word **in context**
  buttoning up his pea-jacket,
- And you deduce information from the context
  - ? it is a kind of jacket (*yellow jacket*?)
  - ? with buttons
  - ? it is thick material (they are going to a stake out)
  - ? it has something to do with peas
    × not true (from the West Frisian word *pijjakker*, in which *pij* referred to the type of cloth used, a coarse kind of twilled blue cloth)
- We are getting better at doing this with computers
  - but people don't just use words
  - they have eyes and noses and other senses, and they have brains that link things

Faculty of Arts

How else do we learn?

- From word internal cues
  - *Television* “far vision”
  - *iphone* “internet phone” (also *individual, instruct, inform, inspire* from the ***iMac***)
  - 鯖 *saba* “mackerel” = 魚 fish; 青 blue
- From the sound
  - *bouba/kiki* ⋆ or ♣
  - *banged, beaten, battered, bruised, blistered, bashed*
  - mouth shape for *teeny weeny* vs *large*
- From images: -

Magnifying Glass

Faculty of Arts

Words are related in many other ways

- Domains: *ball, racket, net, love, ace*
- Origin: *chew, eat, drink* vs *masticate, consume, imbibe*
- ? come up with some words with different origins ?
  English or another language!
- Dialect: *ripper, bonza, sickie, no worries*
- Part-of-speech: *die, live* vs *death, life*
- When you learned them!
- and many more

All of these relations affect how you use and understand language.

Faculty of Arts

Roadmap

Faculty of Arts

Meaning is built up

- **Compositional Semantics**: the meaning of the whole depends (only) on the meanings of the parts and the method of combination.
- The hearer/reader's **interpretation** brings in much more
  - we bring in our existing knowledge
  - we make inferences
- These inferences are based on (or constrained by) the semantics
- two central ideas (formalized by: Katz and Fodor, 1963)
  - Semantic rules must be recursive to deal with infinite meaning
  - Semantic rules interact with syntactic rules to build up meaning
- Two major components:
  - A dictionary pairing lexical items with semantic representations
  - A set of **projection rules** that show how meaning is built up

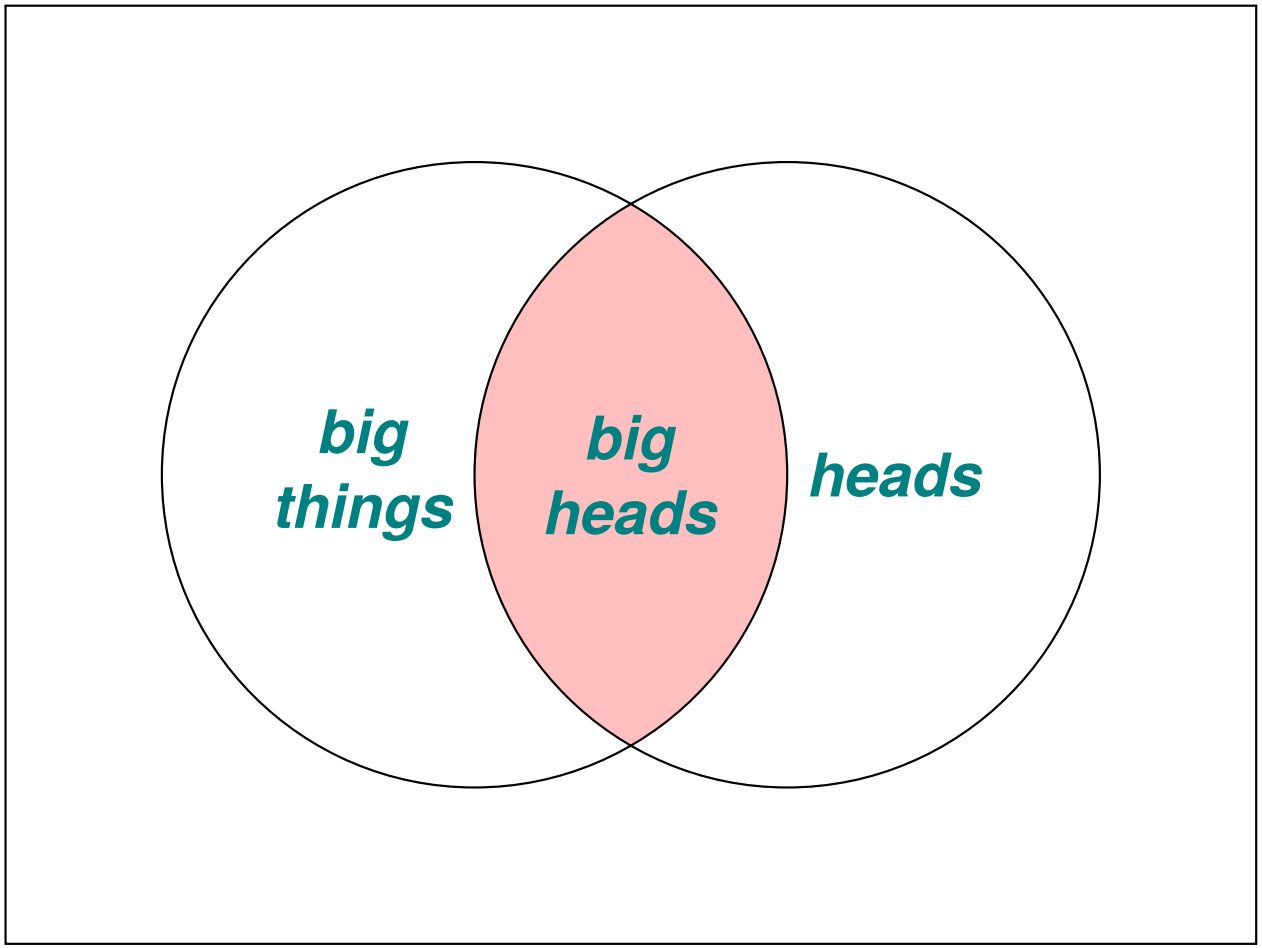
Faculty of Arts

Intersective Modification

- Consider the simplest case of a noun and an adjective
  - ***big*** “above average in size or number or quantity or magnitude or extent”
  - ***head*** “the upper part of the human body or the front part of the body in animals; contains the face and brains”
- Each constrains the world, one picks out things that are “big” and the other things that “are heads”.
- Together ***big head*** picks out things that have both properties: they are “big” and they “have heads”.

Faculty of Arts

This is like intersection for sets



- This is the simplest form of composition
  - although the meaning of big is not independent
  - *a big mouse* is smaller than *a small elephant*

Faculty of Arts

Other kinds of intersective modification

- **Manner**: *We live very quietly, sir* REDH
- **Restriction**: *That trick of staining the fishes' scales of a delicate pink is quite peculiar to China* REDH
- **Location:** *I would rather have my bracelets on him than on any criminal in London*
- **Time:** *one day in the autumn of last year*
- **State:** *And sit in the dark*

The syntactic dependency (the fact that one word/phrase is associated with another) helps us build the semantic model.

Faculty of Arts

Some exceptions

- Not all modification is intersective
  - *fake gun* is a thing like a gun: not a gun
  - *toy horse* is not a horse
  - ? come up with another example of non-intersective modification ?

  This requires different projection rules
- Word combinations (**multi-word expressions**) can pick up new meanings
  - *They have a big head* "They are vain"
  - *They are a red head* "They have red hair"

  This requires a richer lexicon
- There are many other ways of composing words (not just modification)
  - Semantic roles: *The dog barked*
  - Intensification: *They have a very big head*
  - Embedding: *I think they have a big head*
  - Quantification: *They have two heads/no head*

Faculty of Arts

Projection Rules

1. Projection rules combine with syntactic rules to produce the meaning of a sentence
   these can be grouped together in **signs** or **constructions**
   - Information is built up as we parse a sentence
     - Information is only added, never deleted
     - It must come from words or rules (or constructions)
2. Different languages show these combinations in different ways
   - English primarily uses word order
   - Japanese uses case-marking

   ...

? Consider *a very stout, florid-faced, elderly gentleman, with fiery red hair*
   - How many examples of intersective modification are there here? ?
   - Can you describe the other relations involved? ?

Faculty of Arts

Roadmap

Faculty of Arts

Idioms

- Some expressions clearly involve more than one orthographic word
  - compound noun
    - *grass snake*; *grass and tree snakes*
  - verb-particle
    - *I looked it up* vs *I looked up the very long word*
  - idiom
    - *going great guns, give the Devil his due*
    - *jog someone's memory*
    - *blow one's top*
  - And more
    San Francisco, ad hoc, by and large, Where Eagles Dare, kick the bucket, part of speech, in step, the Oakland Raiders, trip the light fantastic, telephone box, take a walk , do a number on (someone), take (unfair) advantage (of), pull strings, kindle excitement, fresh air, …
- Knowing the individual words is not enough to know the meaning (or usage)

Multiword Expressions (MWE)

There are many different kinds of irregularity.

MWE	Weirdness				
	Lexical	Syntactic	Semantic	Pragmatic	Statistical
ad hominem	+	?	?	?	+
at first		+			?
first aid			+		?
salt and pepper					+
good morning				+	+
cat's cradle	+		+		?

- Most of the time, we don't even notice
- Unless it is your second language
- Much more in Sag et al. (2002)

We analyze these differently

- **Fixed Expressions** are like words with spaces
  ad hominem “at the person”
- **Semi-Fixed Expressions** allow some inflection
  part(s)-of-speech “syntactic category”
- **Flexible Expressions** (Nunberg et al., 1994)
  - **Decomposable Expressions** we can treat the parts as having different meanings in context
    spill the beans “reveal the secrets”
  - **Non-decomposable Expressions**
    kick the bucket “die”
- **Institutionalized Phrases** are statistically marked
  black and white vs ?***white and black*** c.f. **白黒** *shirokuro* “white black”
  machine translation vs ?***computer translation***

Faculty of Arts

Semantic Opacity: A Gradient

(1) **Opaque:** *kick the bucket* ⇒ meaning unrelated to kicking/buckets

(2) **Partially transparent:** *spill the beans* ⇒ metaphorical extension

(3) **Highly transparent:** *strong tea* ⇒ collocational preference

Dimensions of semantic idiosyncrasy:

- Non-compositional meaning
- Restricted argument structure
- Fixed lexical choices
- Conventional metaphor/metonymy

Faculty of Arts

Verb-Particle Constructions

(4) *She looked up the word*

(5) *She looked the word up*

Particles may:

- Encode direction (*walk out*)
- Encode aspect (*eat up*)
- Encode idiomatic meaning (*give up*)

Faculty of Arts

Collocations and Selectional Preferences

(6) *strong tea* vs. **powerful tea*

(7) *heavy rain* vs. **strong rain*

Not fully idiomatic, but statistically entrenched.

For computational semantics:

- Collocations affect vector representations.
- Lexical choice is conventionalized.

Faculty of Arts

Formulaic Sequences and Lexical Bundles

In corpus linguistics (Biber et al., 1998):

(8) *on the other hand*

(9) *as a result of*

(10) *in the case of*

Properties:

- High frequency.
- Often discourse-organizing.
- Not necessarily semantically opaque.

Bundles show how usage frequency shapes grammar.

Faculty of Arts

How common are MWEs?

- They are very common in the lexicon
  - In wordnet, 41% of the entries are multiword (mainly compound nouns)
- But less common in the actual text (SPEC 4.5%: 296/6,641)
  24 are new (not in Wordnet 3.0); 55 are named entities
  - *take into one's confidence*
  - *take in*
  - *Sherlock Holmes*
  - *practical joke(r)*
  - *in love*
  - *get the better of*
  - *Panama hat*
  - *as good as one's word*
- It still seems as though wordnet (and dictionaries in general) are missing many MWEs

Faculty of Arts

Why are they important?

- If you think you know the individual words, then you might be confused
- Which is a problem if you are a translator:
  whoever crossed his path “whoever he met” SPEC

(11)

私道	を	渡ろう	と	する	人
shidou	wo	watarou	to	suru	hito
private-road	ACC	cross.let’s	QUOT	do	person

”whoever tried to cross his private road”

- Knowledge of MWEs is one of the things that separates a good speaker from a poor one
- From a linguist’s point of view, they also reveal something about how language is organized in our brains

Faculty of Arts

Idioms are hard to handle

(12) *Kim blew her top*

“Kim got angry”

No **blowing**, no **top**, no **her**

(13) キムは　頭に　　来た

kimu-wa atama-ni kita

Kim-TOP head-DAT came

“Kim got angry (lit: Kim came to her head)”

no **head**, no **coming**

- They are hard to identify
- They require work to represent

Faculty of Arts

The state of the art (translation)

(14) *Kim blew her top*

キムは、 彼女の 上を 吹きました
Kimu wa, kanojo no ue o fukimashita
Kim-TOP her-'s top-ACC blew
Kim exhaled on the upper part of her (Google Translate 2015)

キムは、 彼女の トップを 吹いた
Kimu wa, kanojo no toppu o fuita
Kim-TOP her-'s top-ACC blew
Kim exhaled on the shirt of her (Bing Translate 2015)

キムは、 トップを 吹いた
Kimu wa, toppu o fuita
Kim-TOP top-ACC blew
Kim exhaled on the shirt (Google Translate 2019)

(15) キムは頭に来た
Kim came to the head (Google+Bing Translate 2015)

Faculty of Arts

The state of the art (wordnet)

- Hyponym of *rage* with many lemmas
  - *flip one's lid*
  - *blow one's stack*
  - *flip one's wig*
  - *lose one's temper*
  - *blow up*
  - *throw a fit*
  - *hit the roof*
  - *hit the ceiling*
  - *have kittens*
  - *have a fit*
  - *combust*
  - *fly off the handle*
  - *blow a fuse*
  - *go ballistic*
- But by no means exhaustive
  - *lose one's cool*
  - *lose one's rag*
  - *lose one's grip*
  - *blow one's top*
  - *fly off one's handle*
  - *spit the dummy*

Faculty of Arts

The state of the art (large language models)

- FCB: Translate ”Kim blew her top” into Japanese
- ChatGPT: `The English idiom "Kim blew her top" translates to Japanese as "キムはキレた" (Kimu wa kireta). This phrase captures the sense that Kim lost her temper or became very angry.`
- FCB: What does “キムは頭に来た” mean
- ChatGPT: `The phrase "キムは頭に来た" (Kimu wa atama ni kita) translates to English as "Kim got angry" or more literally, "It came to Kim's head." It's another way to express that Kim became angry or upset, similar to saying "Kim blew her top" in English. The expression "頭に来た" (atama ni kita) is a common way in Japanese to describe someone getting angry or frustrated.`

Translated with openai’s ChatGPT 4.0 (2024-04-21)

Training on a lot of data, with a lot of parameters, makes it much better

Roadmap

1. Word Meaning Revisited
2. Compositionality
3. Multi-Word Expressions
4. Possessed Idioms
   - MWEs and Idioms in your Projects

Faculty of Arts

Possessive Idiomatic Constructions

- Some English idioms contain a variable
- ***wrack one's brains*** "think hard"

  (16) I wrack my brains

  (17) You wrack your brains

  (18) Kim wracks their brains

- The possessor of the 'brains' must be the subject
- The equivalent in Japanese does not: 知恵を絞る ***chie-wo shiboru*** "think hard: lit., squeeze knowledge"

Faculty of Arts

Possessed Idioms (i)

Structure	Example	Frequency
X_{NP} V_1 X's N_1	*lose one's mind*	137
X_{NP} V_1 [P_1 X's N_1]	*fly off one's handle*	40
X_{NP} V_1 X's N_1 [P_1 Y_{NP}]	*cast one's lot with someone/thing*	39
X_{NP} V_1 X's N_1 [P_1 D_1 N_2]	*have one's head in the clouds*	27
X_{NP} V_1 X's N_1 P_1	*cry one's eyes out*	22
X_{NP} V_1 X's own N_1	*blow one's own horn*	18
X_{NP} V_1+P_1 X's N_1	*pull up one's socks*	17
X_{NP} be [P_1 X's N_1]	*[be] off one's rocker*	13
X_{NP} V_1 X's N_1 [P_1 X's N_2]	*scratch one's ear with one's elbow*	13
X_{NP} V_1 D_1 N_1 [P_1 X's N_2]	*a dose of one's medicine*	10
X_{NP} V_1 X's N_1 A_1	*get one's hands dirty*	10
X_{NP} V_1 Y_{NP} [P_1 X's N_1]	*wind someone around one's finger*	10

- Extracted from lexicons
- Collected by me, my interns and students (Sara, Sheefa, Jia Qian)

Faculty of Arts

Possessed idioms (ii)

X_{NP} V_1 X's A_1(est)	*do one's best*	8
X_{NP} V_1 [P_1 X's N_1 [P_2 Y_{NP}]]	*pour out one's heart to someone*	7
X_{NP} aux+neg V_1 X's N_1	*not mince one's words*	5
X_{NP} V_1 Y_{NP} D_1 N_1 [P_1 X's N_2]	*give someone a piece of one's mind*	4
X_{NP} V_1 R_1 A_1 [P_1 X's N_1]	*too big for one's boots*	3
X_{NP} V_1 [P_1 D_1 N_1 P_2 X's N_2]	*by the skin of one's teeth*	2
X_{NP} V_1 N_1 [P_1 X's N_2]	*have egg on one's face*	2
X_{NP} V_1 X's N_1 [P_1 X]	*have one's wits about one*	2
X_{NP} V_1 X's N_1 and V_2 N_2	*have one's cake and eat it*	2
Remainder	*let grass grow under one's feet*	30
Total		**421**

- Surely we have not yet found them all
  - *blow one's load*, *get one's rocks off*, …
- Some of these have other properties (e.g., NPI):

(19) *Kim doesn't let the grass grow under his feet*

(20) ?*Kim lets the grass grow under his feet*

Faculty of Arts

The idiom lexicon “eat one’s words”, …

We can model these formally

(21) {}

Idiom entry	
INDEX FORM	***eat one’s words***
TEMPLATE	X_{NP} V_1 *X’s* N_1
EXAMPLE	*Kim eats her words*
EXAMPLE	*Kim is going to have to eat her words*
DEFINITION	*to retract one’s statement, especially with humility*
V_1	*S:* (V) *eat* (TAKE IN SOLID FOOD)
N_1	*S:* (N) *words* (THE WORDS THAT ARE SPOKEN)
$^*V_1$	*S:* (V) *swallow, take back, unsay, withdraw* (TAKE BACK WHAT ONE HAS SAID)
$^*N_1$	*S:* (N) *statement* (A MESSAGE THAT IS STATED OR DECLARED)
@TYPE	*decomposable …*

Faculty of Arts

The idiom lexicon …, “twiddle one’s thumbs”

(22) {}

Idiom entry	
INDEX FORM	***twiddle one’s thumbs***
TEMPLATE	X_{NP} V_1 *X’s* N_1
EXAMPLE	*Kim twiddles her thumbs*
DEFINITION	*to do nothing*
V_1	*S:* (V) *twiddle, fiddle with* (MANIPULATE)
N_1	*S:* (N) *thumb, pollex* (THE THICK SHORT FINGER)
@TYPE	*Nondecomposable*
PARAPHRASE	*X is idle*
@TEMPLATE	*X BE A*
@A	*S:* (ADJ) *idle* (NOT IN ACTION OR AT WORK)

Faculty of Arts

Corpus Examples I

- Examples of modification and distance in the BNC

(23) *cannot believe my own bloody eyes*

(24) *make one's unsteady way*

(25) *have one's humorous moment*.

(26) *The butcher had lined his pockets too thickly in the past at their expense, and Faith's will had been a warning, a pointer to their future.*

(27) *Now do thy speedy utmost, Meg,*

Faculty of Arts

Corpus Examples II

(28) *Maybe the parties foolishly signed a pact which then found its way into the hands of the DTI.*

(29) *Fitzgerald, who lost his rag with John McEnroe at Wimbledon, had already offered his view of the pair*

(30) *Nor is it the case that the Federal Republic is using the issue of democratic accountability to drag its feet on EMU.*

Faculty of Arts

Semantics of possessed idioms

- We can represent the semantics by transforming it.
  - If the structure can be kept (**decomposable**) then map the predicates to appropriate wordnet senses.

    (31) *Kim changes his tune* "Kim changes his opinion"
  - If it is more opaque (**non-decomposable**) then rewrite to a more suitable semantics.

    (32) *Kim blew her mind* "Kim became angry"
- Representing meaning is hard!

Faculty of Arts

Annotating MWEs

- Be aware of multiword expressions when you analyze the text
- There should be, on average, one in every two sentences
- We expect around 20% to be missing from Wordnet!
- If they are not already in the wordnet, add them in the comment you only have to comment on the first word

Faculty of Arts

References I

Dougla Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. CUP.

J. R. Firth. 1957. *Papers in Linguistics 1934-1951*. OUP.

Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. 1994. Idioms. *Language*, 70:491–538.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15. Springer-Verlag, Hiedelberg/Berlin.

Faculty of Arts