Is there a grammar of idioms?

Christiane Fellbaum

Princeton University

Corpus/computational perspective

One definition of “idiom”

Statistically significant collocation of two or more words

E.g., *bucket* and *kick* and *spill* and *beans* co-occur in corpora with frequency far greater than chance

One definition of “idiom”

Collocational strength between words can be measured, falls along a sliding scale

Allows automatic identification of idiom candidates in corpora

Can identify the members of an idiom

and their idiosyncrasies in the context

“Idiom”

Statistical analysis also identifies frequent, idiosyncratic **compositional** collocations like *strong coffee, brush one’s teeth* and *take a walk.*

Focus here is on wholly or partly non-compositional collocations

Scope of this talk

VP idioms only

Syntactically well-formed: V (NP)* (PP)

hit a nerve

give s.b. the glad eye

look a gifthorse in the mouth

cut to the quick

(I'll ignore the DP-or-no-DP debate here and simply refer to NPs)

Not considered here

Idioms with irregular phrase structure

nothing doing, by and large

These are resistant to grammatical operations, lexical variation

Arguably have no grammar, but are fixed multi-word units

Implausible idioms

pay through the nose
give s.o. the cold shoulder
chew s.o. out
give an arm and a leg
lose one's head
make s.o.'s blood boil
melt s.o.'s heart
fry one's brains
give up the ghost
mop the floor with s.o. (non-accompanying reading)

Inconsistent with our world knowledge
violate selectional restrictions

Plausible idioms

pinch pennies
cry wolf
buy the farm
hit the books
smell a rat
face the music
shoot oneself in the foot
point the finger at s.o.
fall off the wagon
flog a dead horse
bark up the wrong tree
twist s.o.'s arm
stab s.o. in the back

“Plausible” idioms

Plausible idioms have alternative literal readings

Manual classification of German idiom candidates retrieved from a 1 bio-word corpus showed that the literal meaning is intended about half the time on average

(but great variation across idioms)

Questions

How to account for morphosyntactic and lexical variation in idioms?

How to account for apparent constraints?

Is there a grammar of idioms?

A grammar for sub-classes of idioms?

Data problem?

Many proposals for the grammar of English idioms are based on

--small number of high-frequency idioms (*kick the bucket, let the cat out of the bag*)

--constructed, rather than attested data

Constructed data problem

No agreement in the literature as to which syntactic operations are “permitted” for a given idiom

Positing a subset of idiom-specific operations can result in mini-grammars

(These vary across linguists)

Grammars based on classes of idioms can leak; members do not show uniform class-specific behavior

Data problem

We cannot prove a negative: just because we do not encounter an idiom in a particular form doesn't mean it cannot be used in this way

Idioms are not very frequent. We encounter the "canonical form" most often—this may bias our intuition, esp. in the absence of a context

Constructed data problem

Should we rule out variations of idioms because they “seem strange”?

Large corpora, Web provide contexts

Data

Data from corpora/Web show great range of syntactic and lexical variation

Not always consistent with constructed data in the literature nor the proposed rules

Rules, constraints exist, but do not point towards a distinct grammar of idioms

Even infrequent variations show adherence to grammar of literal language

Hallmark of idioms is non-compositionality

Have meaning only as a lexically specified construction ("conventionality")

Believed to be (more or less) "frozen"

Depending on the degree of compositionality, grammatical operations are believed to be blocked (other than verbal inflection)

Some propose a scale from completely non-compositional and frozen to fully compositional and flexible

kick the bucket is often said to be at the "frozen" end

let the cat out of the bag is at the other end

take the bull by the horns is somewhere in between

Corpus/Web data seem to bear out compositionality-flexibility correlation

“Encoding”, “internally regular” idioms

Some idioms are (partially) compositional
*Take **the bull** by the horns*
(the bull = challenge)

bull shows flexibility:
The bull was taken by the horns
I won’t take this bull by the horns
etc.
(Nunberg, Sag, Wasow 1994)

“beat a dead horse”

dead horse = dead issue

This dead horse was beaten into oblivion/weeks ago/senselessly (passive)

CBers love to beat dead horses (plural*)*

Not trying to beat a half-dead horse any more dead (lexical modification*)*

But it's not so easy

Corpus data show that semantic transparency and metaphoric status are not straightforwardly related to variability

Non-compositional idioms show the full range of behavior of free language

Large-scale corpus-based analysis of German idioms

(Fellbaum 2006, 2007)

See also Moon's corpus-based study of English idioms

Corpus of 1 billion words

Targeted study of pre-selected VP idioms with high-frequency heads

Automatic systems should identify and process (semi-) fixed idioms easily

Searched with regular expressions, allowing for morphosyntactic and lexical variation

Iteratively fine-tuned searches to accommodate unexpectedly high degree of variation

Example
Non-compositional

Kein Blatt vor den Mund nehmen

No leaf/sheet before the mouth take

Not cover one's mouth with a sheet

"speak out freely, without inhibition"

Negative polarity item (like many idioms)

Blatt does not refer

Attested variations

(my translations)

passivization (*no sheet was taken…*)

topicalization (*what the speaker did not take to cover his mouth was a sheet*)

pronominalization (…*took it to cover his mouth*),

relativization (*the leaf that he covered his mouth with…*)

Attested variations

(my translations)

Polarity reversal (*ein Blatt*… “a leaf”)

Determiner, number variation

dieses (“this”) Blatt, tausend (“a thousand”) Blätter

Lexical variation: compounding

Notenblatt (*sheet of music*)

Idiom-specific morphemes

Ins Fettnaepfchen treten

Lit. ‘step into the little grease pot’

“commit a faux pas”

Fettnaepfchen doesn’t seem to occur outside this idiom

Does not refer in the idiom

Variations

Das F., in das [er] getreten ist, ist riesig

(relativization)

Ins F. trete ich bestimmt mal

'I'll definitely commit a faux pas at some point'

(topicalization)

Berlusconi: ein Mann, viele F.

(quantification)

Immer trat [er] ins bereitstehende F.

(adjectival modification)

Fettnaepfchen can occur without the verb,
perhaps because it evokes the entire idiom

Metaphoric component can stand alone

Idiom-independent polysemy?

Or has the frequency of the idiom led to *spill* alone evoking the idioms

No tellin what he spilled behind closed doors

What he spilled to the young prosecutor

Rhett is understandably excited about his upcoming record, and from *what he's spilled* so far, it will be quite the album

Kick the bucket

Perhaps the prototypical idiom

Relatively frequent

Much discussed in the literature

Frequently encountered claim:

kick the bucket is opaque and frozen, hence
#the bucket was kicked last night
can only receive a literal interpretation

Web data

There is a certain comfort in that. ***The bucket will be kicked.*** *Then you can go about discovering what happens to a guy after he buys the farm. Heaven? Hell?*

Live life to the fullest, you never know ***when the bucket will be kicked.***

Nominalization

*No**, no kicking of the bucke**t...not anytime soon.*

*the paper in question looks at the economic inequalities that result **from one person's untimely kicking of the bucket** and another one's living*

More nominalizations

*I am young but have experienced **more bucket kicking** within my immediate family and circle of family friends than I can shake a fist at*

*here's a short list of things I hope to continue to avoid from now until **bucket kicking time**.*

To die is not necessarily to be frozen

Passivization in other idioms meaning “to die”

…and *the dust was bitten* by many a poor creature, sand-flies principally

…advise the coroner of the district where *the clogs were popped*

…Amy Winehouse sang it before *her clogs were popped*

Impersonal passive and nominalization seem consistent with these idioms' meaning

Cf. *there was a lot of dying/his dying*

Impersonal passive (German*): Es wurde gestorben*

But not personal passives

Lexical variation

Denied in the literature. But we find examples like these:

Our little brother Willie has *kicked the **pail***

I ain't yet *kicked the **pail***

Objections to the data

“Such data must be discounted”

Dismissed as “playfulness”

“Playfulness” is context-dependent

Distinguish “principled” from “occasional” variation in idioms. The former only must be accounted for by the grammar (A. Sabban)

Objections to these objections

Aren't most syntactic operations and lexical variations conditioned by context, information structure?

Why should idioms be different?

Where does "playfulness" begin and "principled" use of language end?

Is playfulness (if we can identify it) not interesting, as it is presumably subject to the speaker's grammar?

Alternative account

Idioms are subject to all grammatical processes provided their meaning is preserved and they can be recovered by the hearer

How does this play out in light of the data?

Constructions

Some syntactic constructions carry meaning

(Fillmore)

Part of an idiom's meaning may be conveyed by the construction, which must be preserved to convey the idiomatic meaning

Meaningful construction

E.g., many idioms are malefactives

mop the floor #(with someone)

cook s.o.'s goose/#a goose

the bell tolls #(for s.o.)

Maleficient argument is required for idiomatic reading

Semantics of construction
Aspect (perfective)

in den Farbtopf gefallen sein

Lit. have fallen into the paint pot

'wear too much make-up'

X ist eine Laus ueber die Leber gelaufen

Lit. a louse ran across X's liver

'be in a bad mood'

Refer to states; stative aspect must be preserved for the idiomatic reading

This does not preclude lexical variation:

Na, Ihnen sind aber ein paar fette Laeuse ueber die Leber gelaufen

'Well, it seems you had several fat lice running across your liver'

(i.e., you are in a really bad mood)

Context

Clefting, Topicalization

He spilled the beans

Not found with idiomatic reading:

What was spilled were the beans

It was the beans that were spilled

The beans, he spilled them

These data are entirely consistent with free language

Out of context:

#It was a family secret that he told Mary

#What he told Mary was a family secret

#The family secret, he told it

Need appropriate information structure

The event (VP) can be focused:

What happened was that he spilled the beans

It was the spilling of the political beans that upset the senator

Spill the beans is what he did!

Constraints on lexical variation

Recall definition of idiom as a statistically frequent, salient co-occurrence of two or more lexemes

Their co-occurrence evokes the meaning of the idiom

Constituents do not necessarily have to appear in any syntactic configuration

(unless the construction carries part of the meaning)

*Judt valued those who, like him, sought to save the **baby** even as they **chucked out** bucketloads of **bathwater***

Configuration theory (Cacciari & Tabossi 1988)

Literal meanings of idiom components are activated

Idiom is recognized when a "key" in encountered

Key may not always be a function word

E.g., definite article in the absence of an antecedent (*the bucket, the bull*,..) (Fellbaum 1993)

Psycholinguistic view

Idiomatic meaning of frequent plausible English idioms is processed faster than literal meaning

(Glucksberg & Cacciari 1994)

Evidence for their storage as "long words"?

Or: effect of frequency of constellation of lexemes

(i.e., we encounter *kick the bucket, spill the beans* most frequently in their idiomatic uses)

To process the idiom, the hearer/reader must identify the “canonical” constituents and recognize that their co-occurrence is not random

Variant must be sufficiently similar in sound and/or meaning

Semantic and phonological properties of an idiom constituent can be varied in a context-appropriate way

Recovering idioms from constituent(s) is easy for lexemes like *gifthorse, Fettnaepfchen*, which can evoke the idiom without the verb being present

Harder for common lexemes like *beans, strings, bucket* (do they fit in the context?)

Still harder when "canonical" lexeme is substituted

Lexical substitution is constrained, systematic

Semantically similar words: synonyms, super- and subordinates

Compounds

ASL idiom variant

Your ears should have been burning =>

Your eyes should have been burning

Produced spontaneously by ASL interpreter for Deaf speaker (fluent in English)

Real Housewives of OC' Recap: The *Pussycat's Out of the Bag* ..

Pride organisers have *let* the *Pussycat out of the bag* about a multi-million record selling global superstar lined up to perform at this year's…

The Tiger is out of the bag (headline referring to Tiger Wood's recently discovered secret life)

Lexical variations

Lexical substitutions for idiom components with no referents

Hier tanzt der Bär (lit. 'here dances the bear,' "this is where the action is"

Hier rappt/steppt der Hummer (lit. 'here raps/ tapdances the lobster')

This is a productive idiom

Substituted lexemes are members of the same category (animals, performance verbs)

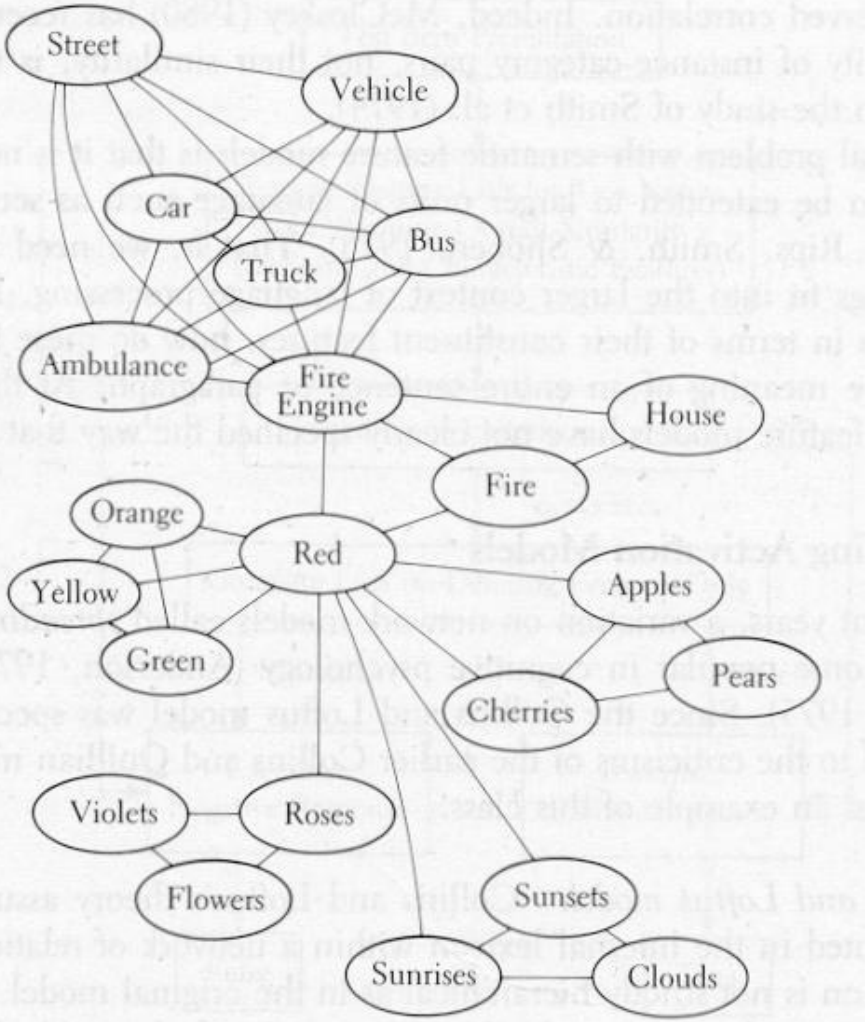
Production and processing of lexical variations is consistent with one theory of human semantic representation and memory

A theory of the mental lexicon

The mental lexicon is a semantic network

Similar words are stored in the same vicinity

Processing involves spreading activation


Street
Vehicle
Car
Bus
Truck
Ambulance
Fire Engine
House
Fire
Orange
Red
Yellow
Apples
Green
Pears
Cherries
Violets
Roses
Flowers
Sunsets
Sunrises
Clouds

Lexical variation is limited to near neighbors

Lexical substitution

Corpus shows some substitution of phonologically similar words (~puns)

Breezy does it! (ad for sandals)

Sleazy does it (report about dishonest lawyer)

Word association, speech error and Tip of the Tongue data suggest that phonologically similar words are "stored" in close proximity

This organization is orthogonal to, but interacting with a semantic one

Data are consistent with psycholinguistic theories

Hybrid theory (Cutting & Bock 1997)

Idioms are represented as separate entries **and** processed like free language

Superlemma theory (Sprenger et al. 2006):

Idioms are conceptual units whose lexemes are bound to both their idiomatic and their free use

This allows for production and processing of variations

Conclusion

Research on the grammar of idioms should be based on authentic data

No/less of an idiom-specific grammar may be needed

Data may not need to be shoehorned into sub-grammars for (classes of) idioms

Not supportive of implicational hierarchy of syntactic transformation (Fraser)

Nor radial theory (Dobrovolskij & Piirainen)

Conclusion

Idioms are idiosyncratic tuples of lexemes

Syntactic arrangement is limited by the idiom's meaning

--meaning-bearing construction

--event-specific (e.g., no personal passive for unaccusative verbs like *die*; impersonal OK)

Lexical variation is systematic

Can be modeled within independent theory of the lexicon (semantic network)