Multiword Expressions

- Ling 226: Sign-Based Construction Grammar
- **Instructor:** Ivan A. Sag (sag@stanford.edu)
- **URL:** http://lingo.stanford.edu/sag/L226
Purely Compositional Analysis Won’t Suffice for MWEs

- **The Overgeneration Problem:**
  
  *telephone booth/box*, but also *telephone closet*  
  *call/phone/ring up*, but also *telephone up*

- **The Idiomaticity Problem:**
  
  The meaning of *kick the bucket* is unrelated to the meanings of *kick*, *the*, or *bucket*.

- **Parsing Problems:**
  
  E.g. *in (*the) step*, on (*the) sale*, *by and large*, *No can do.*
Purely Non-Compositional Analysis (‘Words with Spaces’) Won’t Suffice for MWEs

- **The Flexibility Problem:**
  *look up the tower* is ambiguous ("glance up at" vs. “consult a reference about” (the tower))
  *look the tower up* is unambiguous ("consult a reference . . .”)

- **The Lexical Proliferation Problem:**
  Light verb constructions often come in families:
  *take a walk, take a hike, take a trip/flight...*
  Individual listing results in considerable loss of generality and lack of prediction.
A Taxonomy of MWEs [adapted from Bauer (1983)]

- **Lexicalized Phrases:**
  - Fixed Expressions
  - Semi-Fixed Expressions
  - Syntactically Flexible Expressions

- **Institutionalized Phrases**
  Compositional phrases cooccurring with markedly high frequency (in a given context).
Fixed Expressions

- *by and large, in short, kingdom come, every which way*

- *ad hoc (cf. ad nauseum, ad libitum, ad hominem,...), Palo Alto (cf. Los Altos, Alta Vista,...), etc.*
Fixed Expressions are Fully Lexicalized

- They undergo neither morphosyntactic variation (cf. *in shorter) nor internal modification (cf. *in very short).
- A simple words-with-spaces representation is sufficient.
Semi-Fixed Expressions

- *kicks/kick/kicked/kicking the bucket* ("die"), *part(s) of speech, perjure him*(self)/them*(selves)*
- adhere to strict constraints on word order and composition,
- but undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection.
- can be treated as a word complex which is lexically variable at particular positions.
Semi-Fixed Expressions: U.S. Sports Team Names

- the (Oakland) Raiders
- an/the [(Oakland)Raiders] player
- the [Raiders and 49ers].
- the league-leading (Oakland) Raiders.
- an [(Oakland) Raider] spokesman
- *the (Oakland) 49ers
Syntactically-Flexible Expressions: Semantically Decomposable Idioms

- *take advantage (of)*, *pull strings*, *keep tabs on*, *jump on (the) bandwagon*

- **Syntactic Flexibility:**
  Strings had been pulled to get Sandy the job.
  It was the close tabs they kept on our parents that upset us most.

- **Internal Quantifiability:**
  The FBI kept closer tabs on Kim than they kept on Sandy.
  They took more advantage of the situation than they should have.
Internal Modifiability:
Many Californians jumped on the bandwagon that Perot had started.
She left no legal stone unturned.

Nunberg, Sag and Wasow (1994):
Only semantically decomposable idioms are flexible (in English) and only semantically decomposable idioms allow internal quantification and modification.

Flexibility is highly variable.
The Transformational Myth about Flexible Idioms

- Parts of idioms (e.g. *pull* and *strings*) are uniformly inserted in underlying phrase markers.
- Transformational operations define the space in which parts of idioms may be separated, e.g.:
  
  Strings were **pulled** __ (to get the job).

  Strings seem __ to __ have __ been **pulled** __ (to get Chris the job).

  What *strings* did Pat **pull** __ to get Chris the job?
McCawley’s (1981) Paradox

If the parts of idioms are uniformly inserted in underlying phrase markers, then one of the following examples should be ill-formed:

- Pat *pulled* [[the strings] [that __ got Chris the job]].
- [[The strings] [that Pat *pulled* __ ]] got Chris the job.

Both are well-formed, therefore no uniform assumptions about lexical insertion and the transformational analysis of relative clauses predicts the observed data.
A Further Problem

- Even if some solution could be found to McCawley’s paradox, the fundamental finding of Nunberg et al. (1994) would still remain mysterious.

- Why should semantic decomposability correlate with the ability to undergo transformational movement? (Transformations are semantically blind)
Where to Draw the Line?

- No soap opera worth its bubbles would spill all the beans in one episode if it could dribble them out over many. (Riehemann 2001)
- Microsoft released more details on its Zune player and service, but the cat still has its back paws stuck in the bag. Microsoft’s official announcement of its much-hyped Zune music player came today, just … [ABC News - Sep 14, 2006]
Syntactic Variability 1
(after Riehemann 2001)

- **modification**: Diana spilled the royal beans.
- **open slots**: lose X’s way,
- **passive**: The beans were spilled.
- **raising (control)**: The hatchet appears to have been buried.
  (The piper wants to be paid.)
- **topicalization**: The other beans, she’ll probably spill later.
Syntactic Variability 2

- **distribution over several clauses:**
  (The McCawley Paradox)

- **pronominal reference and ellipsis:**
  I thought the hatchet had been buried, but *it* appears not to have been __. They thought the cat was out of the bag, but *it* wasn’t __.
Further Considerations 1

- *Properties shared between idiomatic and literal words:* kick, kicked, kicking,...

- *No literal interpretation:* close up shop, tend shop,...

- *Restricted flexibility:* caught in the middle, taken aback, fit to be tied, caught short, written in stone... (*passive only*)
Further Considerations 2

- **idiom families**: lose one’s mind (marbles, wits); get off one’s ass (tush(ie), rear (end), butt, duff, tuchus...); throw someone to the dogs (lions, wolves,...)...
Syntactically ‘Deep’ Dependencies

- **Adjectives and Specifiers**: bark up the wrong tree, give me some skin...
- **Adverbs and Adjuncts**: to put it mildly, skate on thin ice, ...
- **Headless Idioms**: get/set/start/keep/have the ball rolling, up the creek without a paddle,...
- **What’s X Doing Y?**: What’s this fly doing in my soup?
Decomposable Idioms

- The relationship between words in decomposable idioms can be captured using a partially semantic mechanism [Nunberg et al. (1994)].
- Flat semantic representations like the MRS representations proposed by Copestake et al. (1995, 2006) are especially well suited to this.
- *cat out of the bag* can be described in terms of the following semantic relationships, where $i_{\text{cat}}$ and $i_{\text{bag}}$ are the meanings corresponding to the idiomatic senses of *cat* “secret” and *bag* “hiding place”:

$$[i_{\text{cat}}(x) \land i_{\text{bag}}(y) \land \text{out}(x, y)]$$
Every dog chased some cat.

- top $h_0$
  $h_1: \text{every}(x, h_3, h_2), h_3: \text{dog}(x), h_7: \text{cat}(y), h_5: \text{some}(y, h_7, h_6), h_4: \text{chase}(e, x, y)$

- Let $h_0 = h_1$ and $h_2 = h_5$ and $h_6 = h_4$. (Every dog has wide scope.)

- Let $h_0 = h_5$ and $h_6 = h_1$ and $h_2 = h_4$. (Some cat has wide scope.)
Idiomatic Constructions

- **cat_out_of_bag :=**
  
  \[
  \left[ \text{SEM} \left[ \text{RELS} \left( h_1:i_{\text{cat\_rel}}(x), h_2:i_{\text{bag\_rel}}(y) \right) \right] \right] \cdot h_3:out\_rel(x,y)...
  \]

- **i_cat :=**
  
  \[
  \left[ \text{SEM} \left[ \text{RELS} \left( h_1:i_{\text{cat\_rel}}(x) \right) \right] \right] & / \text{cat\_n1}.
  \]

- **i_bag :=**
  
  \[
  \left[ \text{SEM} \left[ \text{RELS} \left( h_2:i_{\text{bag\_rel}} \right) \right] \right] & / \text{bag\_n1}.
  \]
Two Fundamental Problems:

- We need to ensure that all elements are present.
- We need to make sure that idiom chunks don’t appear elsewhere.
  
  *(i) Sandy was out of the bag.
  *(i) Kim objected to those strings.
  *(i) We liked the tabs.
  *(i) The strings were offensive.
Locality Problem 5:
Control in Serbo-Croatian (Zec 1987),
Halkomelem Salish, ...

NP\_i promise [COMP he\_i,\_j VP]
NP persuade NP\_i [COMP he\_i,\_j VP]
Locality Problem 6:
English Idioms with Pronominal Genitives

- He lost [his/*her marbles].
- They kept [their/*our cool].
- I/*Kim/*You lost [my way].
External Argument (XARG)

The feature **XARG** is used to specify a distinguished element (e.g. subject, possessor, or object) within a given phrase. The value of **XARG** is either a sign or else the distinguished element **none** (cf. other analyses in HPSG).
Kim took [unfair [advantage]] of the situation.
Kim spilled [the [political [beans]]].

**Solution:** LID is passed up from head-daughter to mother when a modifier is present.

Related to this are transparent nouns:
Kim took [the [kind [of [unfair [advantage of the situation]]]] that was typical of the bourgeoisie of that era].
Hilary would keep [that [kind [of [a [promise]]]]].
Problem 9: Semantically Decomposable Idioms

- Each idiomatic verb requires the presence of the appropriate selected idiom chunk.
  (Keep tabs on/*of...; *Keep advantage of; Pull strings/*twine...)

- We must ensure that the idiom chunks occur only in the presence of an appropriate selector.
  (*The tabs bothered me.; *We objected to their umbrage...)
Syntactically-Flexible Expressions: Semantically Decomposable Idioms

- *take advantage (of), pull strings, keep tabs on, jump on (the) bandwagon*

- **Syntactic Flexibility:**
  Strings had been pulled to get Sandy the job.
  *It was the close tabs they kept on our parents that upset us most.*

- **Internal Quantifiability:**
  The FBI kept closer tabs on Kim than they kept on Sandy.
  *They took more advantage of the situation than they should have.*
- spill the beans,
- keep tabs on,
- pull strings,...
strans-v-lxm

FORM  ⟨ pull ⟩

ARG-ST  ⟨[SYN NP[ ]] [SYN NP[LID i-strings-fr]]⟩

SEM  ⟨[SEM [IND i]′ [SEM [IND j]]]⟩

SYN  [CAT [LID X]]

SEM  FRAMES  ⟨X: [pulling^{strings-fr}]]⟩

FORM  ⟨ strings ⟩

ARG-ST  ⟨ ⟩

SYN  [CAT [LID X]]

SEM  ⟨IND s ⟩

SEM  FRAMES  ⟨X: [i-strings-fr]⟩

SEM  ⟨X: [ENTITY i]⟩
They’re pulling strings to get you the job.
We have pulled strings more than once.
We pulled strings to get invited.
He pulls strings whenever he can.
specify that frames are classified as idiomatic frames (of type \textit{i-frame}) or canonical frames (of type \textit{c-frame})

idiomatic predicators (e.g. $\text{pull}^{\text{strings-fr}}$, $\text{spill}^{\text{beans-fr}}$) are classified as c-frames.

The \textit{i-frame} analysis is motivated by the basic fact that an idiomatic argument (e.g. \emph{strings} in its idiomatic sense) can only appear in construction with the right governor (e.g. \emph{pull} in its appropriate idiomatic sense).

The reason why examples like these only allow a nonidiomatic interpretation, and are therefore hard to contextualize, is that the listemes for the verbs in these sentences select arguments whose \texttt{LID} value must be of type \textit{c-frame}:

- Leslie found the strings that got Pat the job.
- We resented their tabs.
- The beans impressed us.
The motivation for classifying idiomatic predicators as c-frames is that, in spite of their idiomatic meanings, they project phrases (typically VPs or Ss) that freely appear in nonidiomatic environments. That is, their distribution shows none of the restrictions that idiomatic arguments must obey, e.g.

1. I think [Kim spilled the beans].
2. They tried to [pull strings to get Lee the job].
3. [With [my kids [keeping tabs on the stock market]]], I can finally think of retiring.
4. [Taking care of homeless animals] is rewarding.
In this analysis, both idiomatic arguments and idiomatic predicators correspond to the idiomatic meanings of the parts of an MWE: *pullstrings-fr* might be glossed as ‘exert’ and *i-strings-fr* as ‘influence’.

Because the idiomatic meaning is distributed over the parts in this way, it is possible to modify or quantify these parts using the very same analysis that is responsible for the modification and quantification of nonidiomatic expressions.
Syntactic Flexibility:

- *i* The bucket had been kicked many times in that community.
- *i* It was the bucket(s) that they had kicked that upset us most.
- *i* Europeans will kick more buckets this year than last.
- *i* Many Californians were kicking the bucket that the Georgetown kool aid made available to them.
Kick the bucket: two listemes

\[
\begin{align*}
\textit{intrans-v-lxm} & \\
\text{FORM} & \langle \text{kick} \rangle \\
\text{ARG-ST} & \left\langle \begin{bmatrix} \text{SYN} & \text{NP} \end{bmatrix}, \begin{bmatrix} \text{SEM} & \text{IND} \end{bmatrix} \right\rangle, \left\langle \begin{bmatrix} \text{SYN} & \text{MKG} \end{bmatrix} \right\rangle \\
\text{SEM} & \left\langle \begin{bmatrix} \text{FRAMES} & \text{death-fr} \end{bmatrix} \right\rangle \\
\text{cn-lxm} & \\
\text{FORM} & \langle \text{bucket} \rangle \\
\text{SYN} & \left\langle \begin{bmatrix} \text{CAT} & \text{LID} \end{bmatrix} \right\rangle \\
\text{SEM} & \left\langle \begin{bmatrix} \text{IND} & \text{none} \end{bmatrix} \right\rangle \\
\text{FRAMES} & \langle \rangle
\end{align*}
\]
Although lexemes licensed by this bucket listeme have i-bucket-fr as their Lid value, they have no frame on their Frames list and an Ind value of none.

Hence the bucket contributes nothing to the semantic composition of the sentence and provides nothing for a modifier to modify or for a quantifier to restrict.

This can predict the absence of idiomatic readings for such modifications and quantifications.

The failure of idiomatic kick to passivize can be accounted for by its classification as an intransitive verb lexeme.

This could be replaced by a less stipulative account, should one be properly motivated.
Note that although the idiomatic *bucket* provides no semantic argument for an internal modifier, the idiomatic *bucket* may nonetheless be modified by metalinguistic elements, which do not make reference specifically to the common noun’s meaning or index.

Thus we find contrasts like the following:

- *’*Kim kicked the awful bucket.
- *’*They kicked the bucket that they knew was inevitable.
- They kicked the proverbial bucket.
A Persistent Default Analysis (Sag 2006)

\[
\begin{align*}
\text{PHON} & \quad \langle \text{pull} \rangle \\
\text{SYN} & \quad \text{VAL} \langle \text{SYN NP}_i \rangle, \quad \text{SYN NP[LID } \text{i\_strings\_rel}\rangle \\
\text{SEM} & \quad \text{RELS } \langle h_0: i\_\text{pull\_rel}(i, j) \rangle
\end{align*}
\]
\[
\begin{align*}
\text{PHON} & \quad \langle \text{pulled, some, strings} \rangle \\
\text{SYN} & \quad [\text{VAL} \langle \text{NP}_i \rangle ] \\
\text{SEM} & \quad [\text{RELS} \langle h_0: i \_\text{pull}_{\text{rel}}(i, j) \rangle \\
& \qquad \langle h_3: \text{some}_{\text{rel}}(j, h_1, h_2), h_1: i \_\text{strings}_{\text{rel}}(j) \rangle ]
\end{align*}
\]
Motivation for Persistent Defaults

- Kim baked. [bread, cake, etc.; not ham, etc.]
- Sandy drinks. [alcohol]
- They’ve eaten. [a meal]
- They climbed all day. [upward]
[PHON \(\langle \textit{bake} \rangle\)]

[SYN \[\text{VAL} \langle [\text{SYN NP}_i], [\text{SYN NP}[\text{LID} /_p i_\text{-flour\_bsd\_rel}]] \rangle\]]

[SEM \[\text{RELS} \langle h_0: \text{bake\_rel}(i, j) \rangle\]]
relation

\[ \ldots \]

strings_rel

\[ i_{\text{strings}}_{\text{rel}} \quad l_{\text{strings}}_{\text{rel}} \]
<table>
<thead>
<tr>
<th>PHON</th>
<th>⟨ strings ⟩</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYM</td>
<td>CAT [noun]</td>
</tr>
<tr>
<td></td>
<td>LID [strings_rel /ₚ l_strings_rel]</td>
</tr>
<tr>
<td></td>
<td>VAL ⟨ ⟩</td>
</tr>
<tr>
<td>SEM</td>
<td>INDEX i</td>
</tr>
<tr>
<td></td>
<td>RELS ⟨ h₀:₀(i) ⟩</td>
</tr>
</tbody>
</table>
form ⟨pulled, the, strings, that, got, Kim, the, job⟩

SYN VP

form ⟨pulled⟩

SYN V

form ⟨the, strings, that, got, Kim, the, job⟩

SYN NP[LID i_strings_rel]

form ⟨the, strings⟩

SYN NP[LID i_strings_rel]

form ⟨that, got, Kim, the, job⟩

SYN [CAT [SELECT [0]]]

that

form ⟨got, Kim, the, job⟩

SYN [GAP [0]]
form ⟨the, strings, that, Pat, pulled, got, Chris, the, job⟩

SYN S

FORM ⟨the, strings, that, Pat, pulled⟩
SYN NP[LID i_strings_rel]

FORM ⟨got, Chris, the, job⟩
SYN VP

FORM ⟨the, strings⟩
SYN NP[LID i_strings_rel]

FORM ⟨that, Pat, pulled⟩
SYN RC[GAP ⟨0⟩]

that

FORM ⟨Pat, pulled⟩
SYN S[GAP ⟨0⟩]

FORM ⟨Pat⟩
SYN NP

FORM ⟨pulled⟩
SYN VP[GAP ⟨0⟩]
Idiom Parts without a Literal Sense

\[
\begin{align*}
\text{FORM} & \langle \text{umbrage} \rangle \\
\text{SYN} & \begin{cases}
\text{CAT} & \langle \text{noun} \rangle \\
\text{LID} & \langle 0i_{-}\text{umbrage\_rel} / p \uparrow \rangle \\
\text{VAL} & \langle \rangle \\
\text{SEM} & \langle \text{INDEX} i \rangle \\
\text{RELS} & \langle h_0:0(i) \rangle 
\end{cases}
\end{align*}
\]
Unresolved Issues for the Localist Architecture

- Can Binding Theory be Purely Local?
- Are Local Analyses Adequate for All Kinds of Idiomaticity? (See Sailer and Richter 2010)
- Are There Other Non-Local Phenomena that Pose Difficulties?
Conclusions

- The Head Feature Principle, the Nonlocal Feature Principle and other principles of HPSG/SBCG provide a theory of the extension of local domains.
- Particular proposals about features (GAP, LID, XARG, etc.), together with those principles, constitute hypotheses about what information is systematically transmitted outside of local domains.
- HPSG/SBCG provides a comfortable home for such analyses.
- Other frameworks (to the best of my knowledge) have not as yet provided explicit hypotheses about how such theoretical issues are to be resolved.