• One morning I shot an elephant in my pajamas
• I didn’t shoot an elephant
• Imma let you finish but Beyonce had one of the best videos of all time
• I do uh main- mainly business data processing
• 一天早上我穿着睡衣射了一只大象
• The White House pledged to cut down the red tape for access to public documents.
Multiword expressions

- The White House pledged to cut down the red tape for access to public documents
Multiword expressions

The White_House pledged to cut_down the red_tape for access to public documents
## Multiword expressions

<table>
<thead>
<tr>
<th>type</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MW compounds</td>
<td>red tape, motion picture, daddy longlegs, hot air balloon, trash talk</td>
</tr>
<tr>
<td>verb-particle</td>
<td>pick up, dry out, take over, cut short, hold hostage, take seriously</td>
</tr>
<tr>
<td>verb-noun</td>
<td>pay attention (to), go bananas, lose it, break a leg, make the most of</td>
</tr>
<tr>
<td>support verbs</td>
<td>make decisions, take breaks, take pictures, have fun, perform surgery</td>
</tr>
<tr>
<td>coordination</td>
<td>cut and dried/dry, more or less, up and leave</td>
</tr>
<tr>
<td>connective</td>
<td>as well as, let alone, in spite of, on the face of it/on its face</td>
</tr>
<tr>
<td>fixed phrase</td>
<td>easy as pie, scared to death, go to hell, bring home the bacon</td>
</tr>
<tr>
<td>proverbs</td>
<td>Beggars can't be choosers. The early bird gets the worm.</td>
</tr>
</tbody>
</table>

Schneider 2014
Multiword expressions

- Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity

Baldwin and Kim (2010); Sag (2002)
Predictability

• The meaning and behavior of multiword expressions is typically not predictable from the individual words that comprise it.

• “dog”
• “top”
• “days”

• “dog days”
• “top dog”

Baldwin and Kim (2010)
Syntactic idiomaticity

• The syntax of the MWE isn’t predictable from its components

Baldwin and Kim 2014

by and large
Semantic idiomaticity

- The meaning of a MWE is not predictable from its components

kick  →  bucket  →  kick the bucket

Baldwin and Kim 2014
Pragmatic idiomaticity

• An MWE is associated “with a fixed set of situations or particular context”

<table>
<thead>
<tr>
<th>good morning!</th>
<th>Fixed greeting used at same time of day</th>
</tr>
</thead>
<tbody>
<tr>
<td>all aboard!</td>
<td>used in specific situation of boarding a train/ship</td>
</tr>
<tr>
<td>shock and awe</td>
<td>fixed phrased associated with specific moment in Iraq War</td>
</tr>
</tbody>
</table>

Baldwin and Kim 2014
Lexical idiomaticity

• At least one component of the MWE doesn’t appear in the vocabulary on its own.

| ad hoc | “created or done for a particular purpose as necessary” |

• Neither “ad” nor “hoc” are English words on their own.
Statistical idiomaticity

- The words in a MWE occur unusually frequently together compared to their individual frequency.

<table>
<thead>
<tr>
<th></th>
<th>flawless</th>
<th>immaculate</th>
<th>impeccable</th>
<th>spotless</th>
</tr>
</thead>
<tbody>
<tr>
<td>condition</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>credentials</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>hair</td>
<td>-</td>
<td>+</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>house</td>
<td>?</td>
<td>+</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>logic</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>timing</td>
<td>?</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note:* “+” = strong lexical affinity, “?” = neutral lexical affinity, “−” = negative lexical affinity.

Baldwin and Kim 2014; Cruse 1986
Compositionality

We can build up the meaning of a sequence by the combination of its parts

He ate the cheeseburger
Compositional differences resist compositionality.
MWE dictionaries

Random sample of WordNet MWEs:

- arctic willow
- blade apple
- cardiac valve
- de bakey
- glycerol tristearate
- line of descent
- madagascar cat
- vaginal discharge
- western red cedar
- works program

- WordNet contains multiword entries
red tape

Contents [hide]

1 English
  1.1 Etymology
  1.2 Noun
    1.2.1 Synonyms
    1.2.2 Translations
  1.3 See also
    1.3.1 Usage notes
  1.4 Anagrams

English [edit]

Etymology [edit]

- Thought to allude to the former practice of binding government documents in red-coloured tape

Noun [edit]

red tape (uncountable)

1. The binding tape once used for holding important documents together. [quotations ▼]
2. (metonymically, idiomatic) Time-consuming regulations or bureaucratic procedures.
His administration filed briefs that urged the Supreme Court of the United States to strike down bans as unconstitutional ('United States v. Windsor' and 'Obergefell v. Hodges');
Welcome to Phrases.com

Phrases.com is a large collection of common phrases, casual expressions and idioms – collaboratively assembled by our contributing editors. Entries can be easily browsed, searched for, rated, heard and translated to several common and not so common languages.

Improve your English and writing skills by navigating our comprehensive phrases dictionary alphabetically, or simply search by keywords.

stand for
To tolerate.
MWE Extraction

• In many cases, existing MWE lexical don’t cover the specific MWE present in a new domain.

• Several methods for extracting MWE from a corpus.
Collocations

“An arbitrary and recurrent word combination”
[Benson 1990; Baldwin and Kim 2010]
\( \chi^2 \)

- \( \chi^2 \) (chi-square) is a statistical test of dependence—here, dependence between the two variables of word 1 identity and word2 identity.

- For assessing the difference in two datasets, this test assumes a 2x2 contingency table:

```
   word 1  ¬word 1
  
word 2    7     104023
¬word 2   104    251093
```
To test whether “white house” is a meaningful collocation, we can ask: does the word *house* occur significantly more frequently after *white*?

\[
\chi^2
\]

<table>
<thead>
<tr>
<th>(w_1=\text{white})</th>
<th>(w_1=\neg\text{white})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_2=\text{house})</td>
<td>104</td>
</tr>
<tr>
<td>(w_2=\neg\text{house})</td>
<td>2</td>
</tr>
</tbody>
</table>

“red house”, “my house”

“red car”, “my dog”

“white dog”, “white truck”
For each cell in contingency table, sum the squared difference between observed value in cell and the expected value assuming independence.

\[ \chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]
<table>
<thead>
<tr>
<th></th>
<th>$w_1=\text{white}$</th>
<th>$w_1=\neg\text{white}$</th>
<th>sum</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_2=\text{house}$</td>
<td>104</td>
<td>1004</td>
<td>1108</td>
<td>0.076</td>
</tr>
<tr>
<td>$w_2=\neg\text{house}$</td>
<td>2</td>
<td>13402</td>
<td>13404</td>
<td>0.924</td>
</tr>
<tr>
<td>sum</td>
<td>106</td>
<td>14406</td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequency</td>
<td>0.007</td>
<td>0.993</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Assuming independence:

\[ P(w_1 = \text{white}, w_2 = \text{house}) = P(w_1 = \text{white}) \times P(w_2 = \text{house}) \]

\[ = 0.007 \times 0.076 = 0.00053 \]

Among 14512 words, we would expect to see 7.69 occurrences of *white house*.

<table>
<thead>
<tr>
<th></th>
<th>( w_1 = \text{white} )</th>
<th>( w_1 = \neg \text{white} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_2 = \text{house} )</td>
<td>7.69</td>
<td>1095.2</td>
</tr>
<tr>
<td>( w_2 = \neg \text{house} )</td>
<td>93.9</td>
<td>13315.2</td>
</tr>
</tbody>
</table>

\[ P(w_1 = \text{white}) \quad P(w_1 = \neg \text{white}) \]

\[ 0.007 \quad 0.993 \]

\[ P(w_2 = \text{house}) \quad P(w_2 = \neg \text{house}) \]

\[ 0.076 \quad 0.924 \]
What $\chi^2$ is asking is: how different are the observed counts different from the counts we would expect given complete independence?

$$\chi^2$$

<table>
<thead>
<tr>
<th></th>
<th>$w_1=\text{white}$</th>
<th>$w_1=\neg\text{white}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_2=\text{house}$</td>
<td>104</td>
<td>1004</td>
</tr>
<tr>
<td>$w_2=\neg\text{house}$</td>
<td>2</td>
<td>13402</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$w_1=\text{white}$</th>
<th>$w_1=\neg\text{white}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_2=\text{house}$</td>
<td>7.69</td>
<td>1095.2</td>
</tr>
<tr>
<td>$w_2=\neg\text{house}$</td>
<td>93.9</td>
<td>13315.2</td>
</tr>
</tbody>
</table>
\( \chi^2 \)

- With algebraic manipulation, simpler form for 2x2 table \( O \) (cf. Manning and Schütze 1999)

\[
\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}
\]
The $\chi^2$ value is a statistic of dependence with a probability governed by a $\chi^2$ distribution; if this value has low enough probability in that measure, we can reject the null hypothesis of the independence between the two variables.
$\chi^2$

5% probability mass from 3.84 forward; if $\chi^2$ is in this region, then we reject independence as being too unlikely (at $\alpha = 0.05$)
Chi-square is ubiquitous in corpus linguistics (and in NLP as a measure of collocations).

A few caveats for its use:

- Each cell should have an *expected* count of at least 5
- Each observation is independent
Why is part of speech tagging useful?
POS indicative of MWE

$((A \mid N)^+ \mid ((A \mid N)^*(NP))(A \mid N)^*)N$

<table>
<thead>
<tr>
<th>AN</th>
<th>linear function; lexical ambiguity; mobile phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>regression coefficients; word sense; surface area</td>
</tr>
<tr>
<td>AAN</td>
<td>Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase</td>
</tr>
<tr>
<td>ANN</td>
<td>cumulative distribution function; lexical ambiguity resolution; accessible surface area</td>
</tr>
<tr>
<td>NAN</td>
<td>mean squared error; domain independent set; silica based packing</td>
</tr>
<tr>
<td>NNN</td>
<td>class probability function; text analysis system; gradient elution chromatography</td>
</tr>
<tr>
<td>NPN</td>
<td>degrees of freedom; [no example]; energy of adsorption</td>
</tr>
</tbody>
</table>

Justeson and Katz 1995
MWE prediction

• Many phrases are ambiguous about whether they are a MWE in context.
  • the white house pledged to reduce red tape
  • he lives in the white house on the corner
  • Kim made a face at the policeman.
  • Kim made a face in pottery class.

Baldwin and Kim (2010)
MWE prediction

- Data: 55,000 tokens of web reviews annotated for MWE in context.

I googled restaurants in the area and Fuji_Sushi came_up and reviews were great so I made_ a carry_out _order

https://github.com/nert-nlp/streusle/
BIO notation

**Standard BIO entity notation**

No gaps, he was willing to budge a little on the price which means a lot to me. \((0|BI^+)^+\)

1-level: 0 0 0 0 0 B I O 0 0 0 0 B I I I I I 0

**Expanded BIO to accommodate one layer of nesting**

Gappy, he was willing to budge a little on the price which means a lot to me. \((0|B(o|bI^+|I)^+I^+)\)

1-level: 0 0 0 0 B b i I O 0 0 0 B I I I I I 0
MWE prediction

<table>
<thead>
<tr>
<th>preexisting lexicons</th>
<th>entries</th>
<th>max gap length</th>
<th>$\bar{P}$</th>
<th>$\bar{R}$</th>
<th>$\bar{F}_1$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0</td>
<td></td>
<td>46.15</td>
<td>28.41</td>
<td>35.10</td>
<td>2.44</td>
</tr>
<tr>
<td>WordNet + SemCor</td>
<td>71k</td>
<td>0</td>
<td>35.05</td>
<td>46.76</td>
<td>40.00</td>
<td>2.88</td>
</tr>
<tr>
<td>6 lexicons</td>
<td>420k</td>
<td>0</td>
<td>33.98</td>
<td>47.29</td>
<td>39.48</td>
<td>2.88</td>
</tr>
<tr>
<td>10 lexicons</td>
<td>437k</td>
<td>0</td>
<td>46.66</td>
<td>47.90</td>
<td>47.18</td>
<td>2.31</td>
</tr>
</tbody>
</table>

Best configuration with in-domain lexicon

2 lexicons + $MWtypes$(train)$_{\geq 1}$

<table>
<thead>
<tr>
<th>Supervised Model</th>
<th>$\bar{P}$</th>
<th>$\bar{R}$</th>
<th>$\bar{F}_1$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>74.39</td>
<td>44.43</td>
<td>55.57</td>
<td>2.19</td>
</tr>
<tr>
<td>6 lexicons</td>
<td>74.51</td>
<td>45.79</td>
<td>56.64</td>
<td>1.90</td>
</tr>
<tr>
<td>10 lexicons</td>
<td>76.08</td>
<td>52.39</td>
<td>61.95</td>
<td>1.67</td>
</tr>
<tr>
<td>best configuration with in-domain lexicon</td>
<td>76.64</td>
<td>51.91</td>
<td>61.84</td>
<td>1.65</td>
</tr>
</tbody>
</table>

2 lexicons + $MWtypes$(train)$_{\geq 2}$

---

Schneider et al. (2014), “Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut” (TACL)
Activity

• 13.mwe/JustesonKatz95.ipynb