Multiword Expressions: A pain in the neck of lexical semantics

Computational Lexical Semantics

Gemma Boleda
Universitat Politècnica de Catalunya, Barcelona, Spain
gboleda@lsi.upc.edu

Stefan Evert
University of Osnabrück, Germany
stefan.evert@uos.de
Conventional approach to semantics (still!)

- Propositional meaning = compositional semantics + word meaning
- There is a (fairly obvious) problem ...

(from http://www.museoffire.com/tutorials.html)
The phrase *kick the bucket* does not have a compositional interpretation: it is impossible to compute its meaning from the individual word meanings of *kick* and *bucket*. Even most native speakers are not aware of the origins of this phrase.

Such non-compositional phrases are generally called **multiword expressions** (MWE).

- many other terms: multiword units (MWU), lexicalised word combinations (LWC), dictionary headwords, collocations, …
- non-compositionality is just one property that can make word combinations “special”, but the most important one for semantics.
What are multiword expressions?

My working definition of multiword expressions (MWE)

A **multiword expression** is a combination of two or more words whose semantic, syntactic, … properties cannot fully be predicted from those of its components, and which therefore has to be listed in a lexicon.

Three characteristic aspects of MWE (Manning & Schütze)

- **non-compositionality**: semantically (semi-)opaque
- **non-modifiability**: syntactically rigid
- **non-substitutability**: lexically determined
A note on terminology

empirical collocations
significant cooccurrence
(Firth, Sinclair, …)

semi-compositional pairs
phraseology & lexicography
(e.g. Hausmann)

lexical collocations

lexicalised expressions
non-compositional or
otherwise idiosyncratic
(NLP, e.g. Choueka)

multiword expressions

“collocation”
is a confusing notion
at the heart of the
MWE debate
Subtypes of multiword expressions

- idioms
- figurative expressions
- lexical collocations
- light verbs (SVC, FVG)
- institutionalised phrases & clichés
- complex lexical items (MWU)
- English noun compounds
- named entities
- particle verbs (VPC)
- (multiword) terminology
Scales of MWE-ness

Compositionality
Semantic dimension

Flexibility
Syntactic dimension

Substitutability
Lexical dimension

Compositionality:
- Opaque idiom
- Semi-compositional
- Decomposable metaphor
- Limited variability
- LWC
- Morphosyntactic preferences
- Semi-fixed construction
- Productive MWE pattern
- Selectional restrictions

Flexibility:
- Rigidity
- MWU
- N-gram

Substitutability:
- Productive MWE pattern
- Completely determined (no substitution)
- Partly determined
- Limited variability
- LWC
- Morphosyntactic preferences
- Semi-fixed construction
- Productive MWE pattern
- Selectional restrictions
Subtypes of multiword expressions

- Idioms
- Figurative expressions
- Lexical collocations
- Light verbs (SVC, FVG)
- Institutionalised phrases & clichés
- English noun compounds
- Complex lexical items (MWU)
- Particle verbs (VPC)
- Named entities
- (Multiword) terminology
A case study on lexical combinatorics: the collocates of bucket (n.)
There are 2264 different lemmata in your collocation database for "[lemma = "(bucket)_SUBST" %c]". (Your query "{bucket/N}" returned 1363 matches in 527 different texts)

<table>
<thead>
<tr>
<th>No.</th>
<th>Lemma</th>
<th>Total No in the whole BNC</th>
<th>As collocate</th>
<th>In No. of texts</th>
<th>Log-likelihood value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a_ART</td>
<td>2193551</td>
<td>469</td>
<td>284</td>
<td>1986.2363972664</td>
</tr>
<tr>
<td>2</td>
<td>water_SUBST</td>
<td>36441</td>
<td>159</td>
<td>122</td>
<td>1571.131370213</td>
</tr>
<tr>
<td>3</td>
<td>the_ART</td>
<td>6053575</td>
<td>548</td>
<td>218</td>
<td>1444.9331784248</td>
</tr>
<tr>
<td>4</td>
<td>of_PREP</td>
<td>3047415</td>
<td>367</td>
<td>250</td>
<td>1107.9487512</td>
</tr>
<tr>
<td>5</td>
<td>and_CONJ</td>
<td>2621850</td>
<td>315</td>
<td>189</td>
<td>937.47283124924</td>
</tr>
<tr>
<td>6</td>
<td>in_PREP</td>
<td>1874068</td>
<td>223</td>
<td>148</td>
<td>645.33558940887</td>
</tr>
<tr>
<td>7</td>
<td>with_PREP</td>
<td>659765</td>
<td>146</td>
<td>111</td>
<td>583.72213506699</td>
</tr>
<tr>
<td>8</td>
<td>spade_SUBST</td>
<td>465</td>
<td>29</td>
<td>25</td>
<td>439.51377275601</td>
</tr>
<tr>
<td>9</td>
<td>into_PREP</td>
<td>157873</td>
<td>79</td>
<td>66</td>
<td>436.4903626444</td>
</tr>
<tr>
<td>10</td>
<td>plastic_SUBST</td>
<td>4383</td>
<td>33</td>
<td>24</td>
<td>358.93386077881</td>
</tr>
<tr>
<td>11</td>
<td>size_SUBST</td>
<td>14236</td>
<td>32</td>
<td>3</td>
<td>323.8816986084</td>
</tr>
<tr>
<td>12</td>
<td>to_PREP</td>
<td>2598966</td>
<td>165</td>
<td>105</td>
<td>290.37604326603</td>
</tr>
<tr>
<td>13</td>
<td>mop_SUBST</td>
<td>212</td>
<td>18</td>
<td>13</td>
<td>284.18669303051</td>
</tr>
<tr>
<td>14</td>
<td>fill_VERB</td>
<td>10571</td>
<td>31</td>
<td>24</td>
<td>278.68590784073</td>
</tr>
<tr>
<td>15</td>
<td>slop_SUBST</td>
<td>51</td>
<td>14</td>
<td>8</td>
<td>256.87542724609</td>
</tr>
<tr>
<td>16</td>
<td>ice_SUBST</td>
<td>3999</td>
<td>22</td>
<td>18</td>
<td>225.28485774994</td>
</tr>
<tr>
<td>17</td>
<td>large_ADJ</td>
<td>47548</td>
<td>33</td>
<td>16</td>
<td>202.39562368393</td>
</tr>
<tr>
<td>18</td>
<td>record_SUBST</td>
<td>19505</td>
<td>26</td>
<td>1</td>
<td>192.85549688339</td>
</tr>
<tr>
<td>19</td>
<td>seat_SUBST</td>
<td>10463</td>
<td>21</td>
<td>17</td>
<td>172.76644775736</td>
</tr>
<tr>
<td>20</td>
<td>shop_SUBST</td>
<td>15056</td>
<td>22</td>
<td>8</td>
<td>167.1304030457</td>
</tr>
<tr>
<td>21</td>
<td>hold_VERB</td>
<td>46780</td>
<td>28</td>
<td>16</td>
<td>163.4971307419</td>
</tr>
<tr>
<td>22</td>
<td>throw_VERB</td>
<td>10948</td>
<td>20</td>
<td>18</td>
<td>160.78088808066</td>
</tr>
<tr>
<td>23</td>
<td>kick_VERB</td>
<td>3493</td>
<td>16</td>
<td>11</td>
<td>157.91067552567</td>
</tr>
<tr>
<td>24</td>
<td>for_PREP</td>
<td>867278</td>
<td>69</td>
<td>32</td>
<td>143.98433303833</td>
</tr>
<tr>
<td>25</td>
<td>from_PREP</td>
<td>425899</td>
<td>50</td>
<td>39</td>
<td>137.9228811264</td>
</tr>
<tr>
<td>26</td>
<td>full_ADJ</td>
<td>28809</td>
<td>21</td>
<td>19</td>
<td>130.64289283752</td>
</tr>
<tr>
<td>Noun</td>
<td>$f$</td>
<td>local MI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>water</td>
<td>183</td>
<td>1023.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spade</td>
<td>31</td>
<td>288.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plastic</td>
<td>36</td>
<td>225.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>size</td>
<td>41</td>
<td>195.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>record</td>
<td>38</td>
<td>163.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slop</td>
<td>14</td>
<td>162.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mop</td>
<td>16</td>
<td>155.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ice</td>
<td>22</td>
<td>125.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bucket</td>
<td>18</td>
<td>125.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seat</td>
<td>21</td>
<td>89.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coal</td>
<td>16</td>
<td>77.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>density</td>
<td>11</td>
<td>63.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>brigade</td>
<td>10</td>
<td>62.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sand</td>
<td>12</td>
<td>61.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>algorithm</td>
<td>9</td>
<td>60.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shop</td>
<td>17</td>
<td>59.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>container</td>
<td>10</td>
<td>59.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>champagne</td>
<td>10</td>
<td>56.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shovel</td>
<td>7</td>
<td>56.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>oats</td>
<td>7</td>
<td>54.93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verb</th>
<th>$f$</th>
<th>local MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>throw</td>
<td>36</td>
<td>168.87</td>
</tr>
<tr>
<td>fill</td>
<td>30</td>
<td>139.45</td>
</tr>
<tr>
<td>empty</td>
<td>14</td>
<td>96.73</td>
</tr>
<tr>
<td>randomize</td>
<td>9</td>
<td>96.11</td>
</tr>
<tr>
<td>hold</td>
<td>31</td>
<td>78.93</td>
</tr>
<tr>
<td>put</td>
<td>37</td>
<td>77.96</td>
</tr>
<tr>
<td>carry</td>
<td>26</td>
<td>71.95</td>
</tr>
<tr>
<td>tip</td>
<td>10</td>
<td>59.30</td>
</tr>
<tr>
<td>kick</td>
<td>12</td>
<td>59.28</td>
</tr>
<tr>
<td>chuck</td>
<td>7</td>
<td>44.85</td>
</tr>
<tr>
<td>use</td>
<td>31</td>
<td>42.31</td>
</tr>
<tr>
<td>weep</td>
<td>7</td>
<td>41.73</td>
</tr>
<tr>
<td>pour</td>
<td>9</td>
<td>40.73</td>
</tr>
<tr>
<td>take</td>
<td>42</td>
<td>37.57</td>
</tr>
<tr>
<td>fetch</td>
<td>7</td>
<td>35.13</td>
</tr>
<tr>
<td>get</td>
<td>46</td>
<td>34.73</td>
</tr>
<tr>
<td>douse</td>
<td>4</td>
<td>33.03</td>
</tr>
<tr>
<td>store</td>
<td>7</td>
<td>31.82</td>
</tr>
<tr>
<td>drop</td>
<td>10</td>
<td>31.49</td>
</tr>
<tr>
<td>pick</td>
<td>11</td>
<td>28.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adjective</th>
<th>$f$</th>
<th>local MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>large</td>
<td>37</td>
<td>114.79</td>
</tr>
<tr>
<td>single-record</td>
<td>5</td>
<td>64.53</td>
</tr>
<tr>
<td>full</td>
<td>21</td>
<td>63.23</td>
</tr>
<tr>
<td>cold</td>
<td>13</td>
<td>55.52</td>
</tr>
<tr>
<td>small</td>
<td>21</td>
<td>45.61</td>
</tr>
<tr>
<td>galvanized</td>
<td>4</td>
<td>43.47</td>
</tr>
<tr>
<td>ten-record</td>
<td>3</td>
<td>40.17</td>
</tr>
<tr>
<td>empty</td>
<td>9</td>
<td>38.41</td>
</tr>
<tr>
<td>old</td>
<td>20</td>
<td>35.67</td>
</tr>
<tr>
<td>steaming</td>
<td>4</td>
<td>31.89</td>
</tr>
<tr>
<td>clean</td>
<td>7</td>
<td>27.47</td>
</tr>
<tr>
<td>leaky</td>
<td>3</td>
<td>25.91</td>
</tr>
<tr>
<td>wooden</td>
<td>6</td>
<td>25.50</td>
</tr>
<tr>
<td>bottomless</td>
<td>3</td>
<td>25.17</td>
</tr>
<tr>
<td>galvanised</td>
<td>3</td>
<td>24.70</td>
</tr>
<tr>
<td>big</td>
<td>12</td>
<td>23.86</td>
</tr>
<tr>
<td>iced</td>
<td>3</td>
<td>22.62</td>
</tr>
<tr>
<td>warm</td>
<td>6</td>
<td>19.55</td>
</tr>
<tr>
<td>hot</td>
<td>6</td>
<td>17.05</td>
</tr>
<tr>
<td>pink</td>
<td>3</td>
<td>11.15</td>
</tr>
</tbody>
</table>
Relevance for lexical semantics

• Idioms: *kick the bucket, red herring*
  ‣ completely opaque interpretation
  ‣ homomorphic interpretation vs. computability

• Proper names: *Rhino Bucket*
  ‣ a 1990s hard rock band that sounded very much like AC/DC

• Solution: list in dictionary as “complex words”
Relevance for lexical semantics

• Terminology & lexicalised compounds
  ‣ plastic bucket, fire bucket, bucket shop, bucket seat
  ‣ bus stop, apple pie, motion sickness, support vector machine

• Lexical collocations (semi-compositional)
  ‣ weep buckets (where buckets acts as an intensifier)
    *I used to weep buckets because I wanted to touch him again.*

• Productivity: “complex words” approach not sufficient
  ‣ meaning is at least partially computable
  ‣ regular patterns: *make a mistake, argument, point, statement, …*
The goal of multiword extraction is to identify new MWE and determine their semantic, syntactic, … properties automatically based on corpus data.

Let us take a look at current research in this area …
A series of workshops on MWE

- Identification, Interpretation, Disambiguation and Applications (ACL 2008)
- Towards a Shared Task for Multiword Expressions (LREC 2008)
- A Broader Perspective on Multiword Expressions (ACL 2007)
- MWE: Identifying and Exploiting Underlying Properties (ACL 2006)
- Multiword Expressions in a Multilingual Context (EACL 2006)
- Collocations and Idioms 2006: Linguistic, computational, and psycholinguistic perspectives (Berlin, 2006)
- Multiword Expressions: Integrating Processing (ACL 2004)
- Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003)
- Collocations and Idioms 2003: Linguistic, computational, and psycholinguistic perspectives (Berlin, 2003)
- Computational Approaches to Collocations (Vienna, 2002)
- Workshop on Collocations (ACL 2001)
The state of the art in multiword extraction

**Special issues** of scientific journals

- *Computer Speech and Language* 19(4), 2005
  Multiword Expressions: Having a crack at a hard nut

- *Language Resources and Evaluation*, to appear
  Multiword Expressions: Hard going or plain sailing?

**Online bibliographies**

- MWE project, Stanford (ca. 2001)
- Idioms & Collocations in German, Berlin (ca. 2006)
- Help us build new resources at [http://multiword.sf.net/](http://multiword.sf.net/)
Multiword extraction tasks

- Collocation (LWC) identification
- MWE detection
- Semantic interpretation
- Compositionality
- Morphosyntactic preferences
- Variability & modifiability
- Token recognition
Approaches: LWC identification

Goal: identify lexicalised word combinations
- traditionally **word pairs** (→ collocations)
- also combinations of 3 or more words (*eat humble pie*)
- often restricted to a particular **syntactic relation** or construction

Cooccurrence and statistical association
- exploits overlap between empirical collocations & lexicalisation
- see e.g. [http://www.collocations.de/](http://www.collocations.de/) for details

Additional filters: distance, syntactic patterns, variability, synonym substitution test, lexical resources, …

LWC often form seed (or other part) of a larger MWE
Approaches: MWE detection

What are the essential components of a MWE?

- number of components: *get cold feet* vs. *eat humble pie*
- optional elements: *keep a (small) fortune*
- internal structure:
  - *carry emotional baggage* = *carry baggage* + *emotional baggage*
  - *wish a happy birthday* = *wish* + ( *happy* + *birthday* )

Hierarchical models of statistical association

- model selection techniques from mathematical statistics
- heuristic formulae that determine “best” combination
  (relatively easy for contiguous n-grams)
- massive sparse data problems in $n$-dimensional contingency tables

Additional criteria: e.g. variability & boundary entropy
Approaches: token recognition

- Most MWE also have “literal”, i.e. compositional reading
  → distinguish between MWE and literal *instances* (tokens)
  
  *Did you think I'd kicked the bucket, Ma?*
  vs. *It was as if God had kicked a bucket of water over.*

- British National Corpus: 8 x literal meaning, 3 x idiom (all in reported speech), 9 x metalinguistic (discussion of the idiom)

- Use knowledge about restricted variability of specific MWE

- Can be thought of as a form of *word sense disambiguation*
  
  - classification with *machine learning* algorithms
  - requires separate training data for each distinct MWE
  - are generalisations possible (indications for “literal” context)?
Approaches: morphosyntactic preferences

- MWE often put restrictions on certain morphosyntactic features, or have strong preferences
  - *kick the bucket*: definite article required, only active voice
  - *eat humble pie*: strong preference for null article and singular number, weak preference for active voice

Statistical analysis of **morphosyntactic distributions**

- e.g. proportion of instances in singular, or without article
- corpus with (automatic) morphosyntactic annotation is needed

Problem: often not enough data for significant results

- most MWE have relatively few instances even in gigaword corpora
- exacerbated by low accuracy of morphosyntactic tagging
Approaches: compositionality

- Related to token recognition and WSD
  - **machine learning** approaches are promising

- Determine **semantic compatibility** with context
  - assumption: non-compositional MWE belongs to different semantic field than component words (e.g. metaphor *fig leaf* ≠ *fig* / *leaf*)
  - uses lexical databases such as WordNet or Roget's Thesaurus (similar to Lesk algorithm for word sense disambiguation)

- Distributional semantic models (**DSM**)
  - vector representation of word meaning & compositional meaning
  - compare vector of *humble pie* with vector obtained by composition of *humble* and *pie*
Approaches: semantic interpretation

Can meaning of semi-compositional MWE be predicted?

- **noun compounds**: semantic relation or paraphrasing verb
  - *corpus researcher* = researcher who studies corpora
  - *apple juice*: MATERIAL (juice made from apples)

- **particle verbs**: entailment, specialised senses for each particle
  - *John put up the picture* vs. *John put up his friend over the weekend*
  - *Goal-up (deadline is coming up)*, *Compl-up (drink up)*, *Refl-up (curl up)*, ...

- **lexical collocations and light verbs**: lexical functions
  - *INTENSIFIER(smoker) = heavy*

Lexical collocations vs. word senses

- classical example: *emotional baggage* vs. *emotional luggage*
- metaphorical sense of *baggage* combines with *cultural* (15), *emotional* (13), *historical* (6), *ideological* (5), *political* (4), ...
Approaches: semantic interpretation

- Supervised machine learning (classification problems)
  - yes/no-classification (entailment) or multiple classes
  - training data often specific to particular lexical item (e.g. up)

- Exploit **semantic similarity** of components
  - apple juice → orange juice, mint tea, …
  - using WordNet or distributional models (word space)

- Search for possible **paraphrases** in large corpora
  - often in the form of Google queries
  - e.g. for interpretation of corpus researcher:
    - ? researcher … studies … corpus
    - ? researcher … made of … corpus
    - ? researcher … contains … corpus
Problems & challenges

- Collocation identification (LWC)
  - accuracy still unsatisfactory, only semi-automatic extraction
  - methods do not always generalise to other languages, genres, …

- MWE detection: sparse-data problem

- Morphosyntactic preferences
  - high degree of ambiguity & noise → more corpus data needed

- Semantic interpretation
  - formalisation of non-compositional meaning aspects still unclear
  - no direct comparison of current approaches possible

- Compositionality: DSM still not well-understood
Questions?

Thank you for listening!