Multiword Expressions

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Notice Anything Strange?

(1) *Modern Australian politicians drive me up the wall*

(2) *Sandy walked by and large straight*

(3) *It's time to up stumps and move on*

INTRODUCTION

What are Multiword Expressions (MWEs)?

- *Definition:* A **multiword expression** (MWE) is:

1. decomposable into multiple simplex words
2. lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic

(Baldwin and Kim 2010)

Some Examples

- *Fitzroy North*, *ad hoc*, *by and large*, *Toy Story*, *kick the bucket*, *part of speech*, *in step*, *the Woodville Warriors*, *trip the light fantastic*, *telephone box*, *call (someone) up*, *take a walk*, *do a number on (someone)*, *take advantage (of)*, *pull strings*, *kindle excitement*, *fresh air*, ...

MWE or not MWE?

... there is no unified phenomenon to describe but rather a complex of features that interact in various, often untidy, ways and represent a broad continuum between non-compositional (or idiomatic) and compositional groups of words. (Moon 1998)

MWE Markedness

MWE	***Markedness***				
	Lex	***Syn***	***Sem***	***Prag***	***Stat***
ad hominem	☑	?	?	?	☑
at first	☒	☑	☒	☒	☒
first aid	☒	☒	☑	☒	?
salt and pepper	☒	☒	☒	☒	☑
good morning	☒	☒	☒	☑	☑
cat's cradle	☑	☑	☑	☒	?

Lexicographic Concept of "Multiword"

- *Rough and ready definition:* a lexeme that crosses word boundaries
- Complications with non-segmenting languages (Japanese, Thai, ...) and languages without a pre-existing writing system (Walpiri, Mohawk, ...)
- Also, in English: *houseboat* vs. *house boat*, *trade off* vs. *trade-off* vs. *tradeoff*

Lexical Idiomaticity

- Lexical idiomaticity = one or more of the elements of the MWE does not have a usage outside of the MWE

- Examples of lexical idiomaticity:

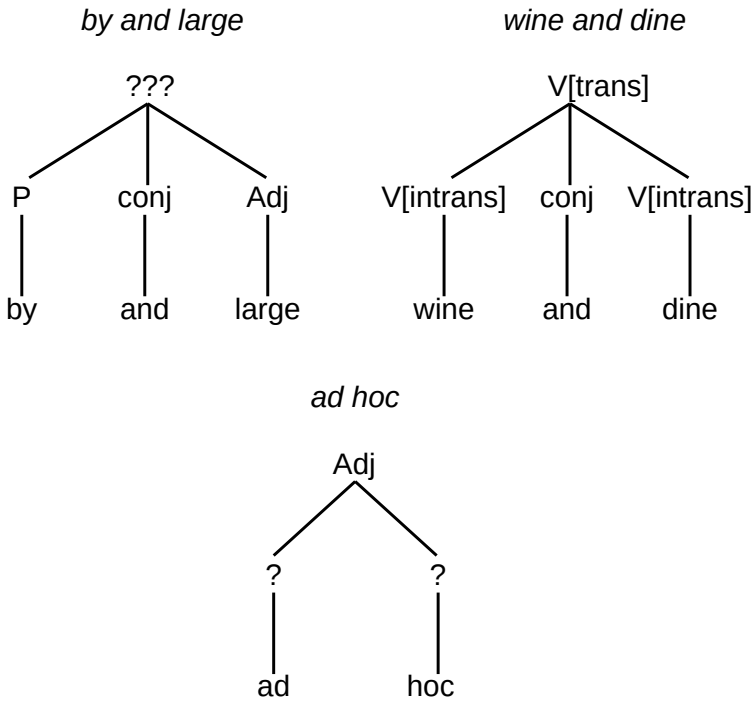
  ad hominem, *bok choy*, *a la mode*, *Moonee Ponds*

- Complications of lexical idiomaticity:

  ⋆ cross-linguistic effects, e.g. *ad* is unmarked in Latin
  ⋆ simple lexical occurrence not sufficient, e.g. *a la mode*

Syntactic Idiomaticity

- Syntactic idiomaticity = the syntax of the MWE is not the sum of its parts



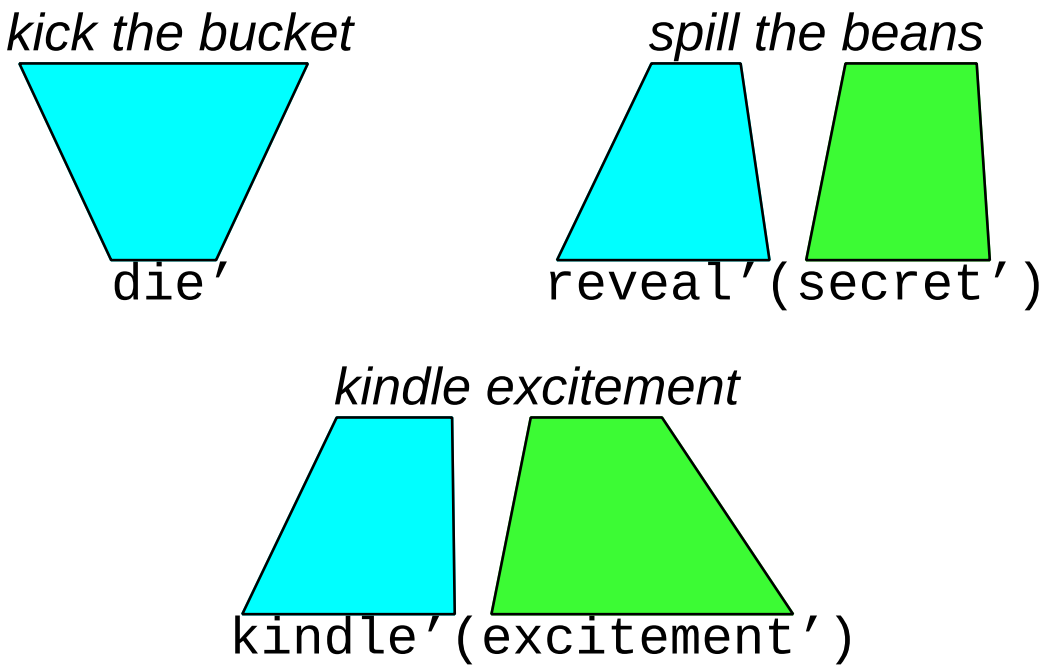
Mapping the Boundaries of Flexibility

- Cline between full flexibility and full rigidity, e.g.:

  Can/could you tell?
  Are you able to tell?
  **They might/ought to tell.*
  How might you tell?
  **How ought they to tell?*

Semantic Idiomaticity

- Semantic idiomaticity = the meaning of the MWE is not the simple sum of its parts



Decomposability and Syntactic Flexibility

- Decomposability = *degree to which the semantics of an MWE can be ascribed to those of its parts*

- Consider:

  **the bucket was kicked by Kim*
  Strings were pulled to get Sandy the job.
  The FBI kept closer tabs on Kim than they kept on Sandy.
  ... the considerable advantage that was taken of the situation

- The syntactic flexibility of an idiom can generally be explained in terms of its decomposability

Pragmatic idiomaticity

- Pragmatic idiomaticity = the MWE is associated with a fixed set of situations or a particular context, or with real-world information or expectations about the MWE (Kastovsky 1982; Jackendoff 1997)
- The Monty Python factor — mish-mash of language fragments which evoke particular events/individuals/memories
- The Wheel of Fortune factor — there's a lot of encyclopaedic junk stored in our heads!

Statistical Idiomaticity I

- Statistical idiomaticity = a particular combination of words has a high lexical affinity, or preferred lexical configuration relative to alternative phrasings of the same expression (Cruse 1986)

Statistical Idiomaticity II

	unblemished	*spotless*	*flawless*	*immaculate*	*impeccable*
eye	−	−	−	−	+
gentleman	−	−	?	−	+
home	?	+	−	+	?
lawn	−	−	?	+	−
memory	−	−	+	−	?
quality	−	−	−	−	+
record	+	+	+	+	+
reputation	+	−	−	+	+
taste	−	−	−	−	+

Adapted from Cruse (1986)

Statistical Idiomaticity and Dialect

- The arbitrariness of some MWEs is brought out well in dialect differences (e.g. OzE vs. AmE):
  - ⋆ *phone box* vs. *phone booth*
  - ⋆ *mail man* vs. *post man*
  - ⋆ *no through road* vs. *not a through street*
  - ⋆ *toss up a wrong un* vs. *throw a curve ball*

MWE Markedness

MWE	***Markedness***				
	Lex	***Syn***	***Sem***	***Prag***	***Stat***
ad hominem	☑	?	?	?	☑
at first	☒	☑	☒	☒	☒
first aid	☒	☒	☑	☒	?
salt and pepper	☒	☒	☒	☒	☑
good morning	☒	☒	☒	☑	☑
cat's cradle	☑	☑	☑	☒	?

Exercise #1: Spot the MWE

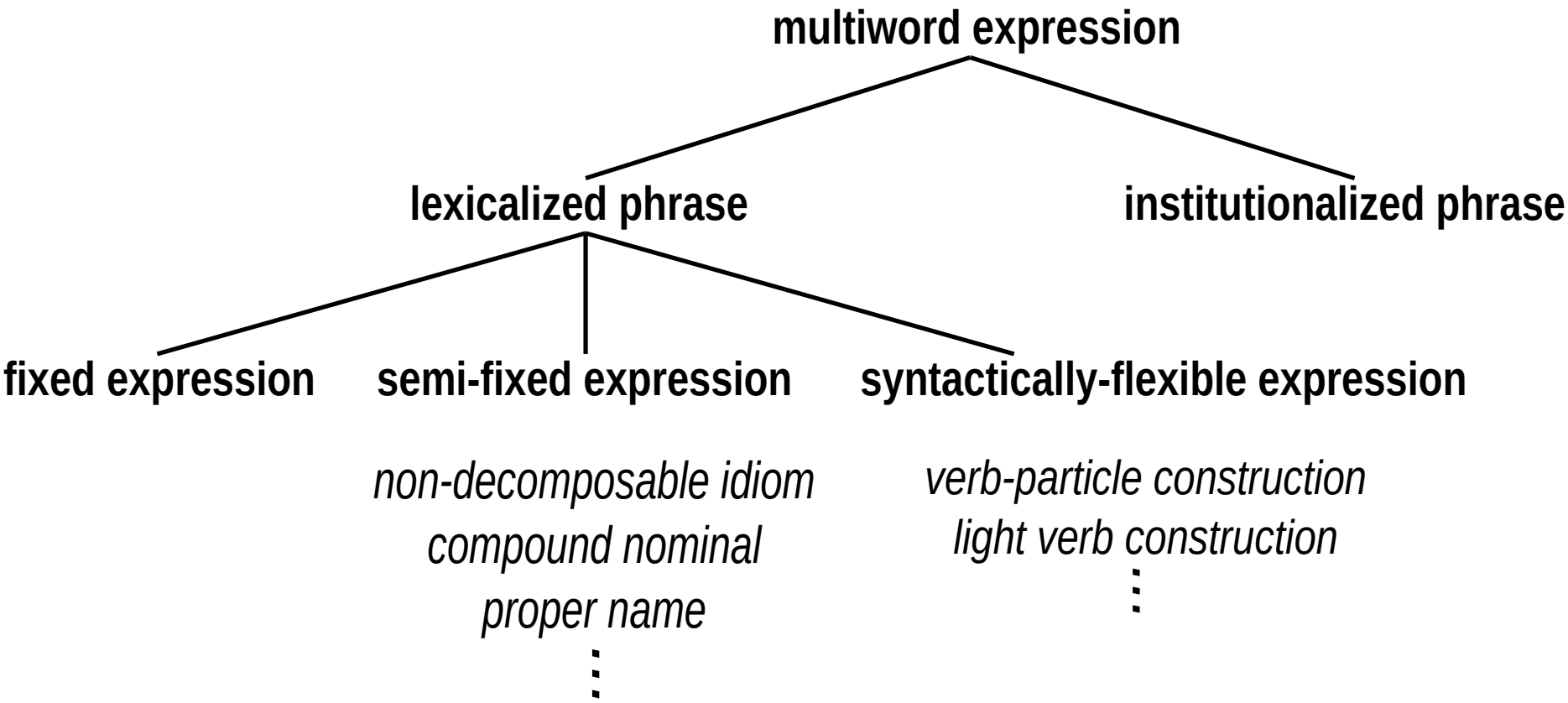
Expression	***MWE?***	***Markedness***				
		Lex	***Syn***	***Sem***	***Prag***	***Stat***
library card						
at arm's length						
old tree						
kick the bucket						
once upon a time						
at home						
to read Shakespeare						

COMPUTATIONAL SYNTAX OF MWEs

Basic Syntactic Approach

- Classify different MWE types according to their syntactic flexibility and productivity, and determine the appropriate analysis accordingly

MWE Types


multiword expression
lexicalized phrase
institutionalized phrase
fixed expression
semi-fixed expression
syntactically-flexible expression
non-decomposable idiom
compound nominal
proper name
⋮
verb-particle construction
light verb construction
⋮

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 string which undergoes neither morphosyntactic variation (**in shorter*) nor internal modification (**in very short*)
- Simple words-with-spaces representation is sufficient

Semi-fixed Expressions

- *kick the bucket*, *prostrate oneself*, *part of speech*, *San Francisco 49ers*
- Adhere to strict constraints on word order and composition
- BUT undergo some lexical variation, e.g.:
  - ⋆ **inflectional**: *kick/kicks/kicking/kicked the bucket*
  - ⋆ **reflexive pronominal**: *prostrate him/.../herself*
  - ⋆ **determiner selection**: *the/those 49ers*

Compound Nouns

- Fully productive = any sequence of nouns can combine to form a MWE (within pragmatic bounds)

- Underspecified semantic relation between the noun modifier and head:

  newspaper archive
  school bus
  apple juice seat

Syntactically-flexible Expressions

- *write up*, *let the cat out of the bag*, *have a shower*, ...
- Variable level of flexibility for different expressions
- Basic mechanism of lexical selection

Verb-Particle Constructions

- **Verb-Preposition Combinations**:
  - ⋆ *It was like falling off a log/*falling a log off.*
  - ⋆ *[Off how many logs] did the drunk fall?*
  - ⋆ *They fell quietly off the log.*

- **Verb-Particle Combinations**:
  - ⋆ They *wrote up the memo/wrote the memo up*
  - ⋆ **Up how many memos did they write?*
  - ⋆ **They wrote quietly up the memos.*

Verb-Particle Constructions

- **Compositional:** *write up*, *eat/gobble up* (activity → accomplishment)
- **Noncompositional**: *look up*, *throw up*
- *write up the memo/write the memo up*
  look up the answer/look the answer up

Representing Institutionalized Phrases

- Store matrix of dependency pairs, with the (smoothed) corpus-based frequency of each
- Statistically disprefer rather than symbolically rule out certain word combinations
- Principal use in natural language generation

MWES IN IR AND NLP

MWEs in IR

- The IR response to MWEs is generally one of the following:

1. ignore the problem, and assume that the combination of query terms + document statistics will solve the problem
2. support phrasal querying (user-disambiguated MWEs)
   e.g. `"multiword expression" solutions`

3. use term co-occurrence statistics to probabilistically predict (sequential or full) "term dependence" (Metzler and Croft 2005), and bias the document ranking function based on the relative dependence/proximity of the terms in the retrieved documents (i.e. represent MWEs purely statistically)

- Most MWEs in search queries are compound nouns + AdjN combinations, so syntactic flexibility not a big problem

MWEs in POS Tagging

- POS (Part of Speech) Tagging = the task of assigning POS tags to tokens in context

- MWEs that cause problems for POS tagging are primarily:
  - ⋆ lexically idiosyncractic MWEs (no or misleading lexical prior, e.g. *a la mode*)
  - ⋆ syntactically idiosyncratic MWEs (e.g. VPCs, cf. *play* <u>*up*</u> *the difference* vs. *play* <u>*up*</u> *the road*)

- The solution is largely to learn lexical priors as well as possible, and capture as much context as possible

MWEs in Machine Translation

- Given sufficient training data, SMT systems do a remarkably good job of capturing (higher-frequency) lexically and semantically idiomatic (semi-)contiguous MWEs, as the alignment tends to be easier

- MWEs that cause problems for MT are primarily:
  - ⋆ (morpho-)syntactically idiosyncratic MWEs (e.g. VPCs, cf. *rampant advantage was constantly taken of Kim*)

- ⋆ pragmatically idiosyncratic MWEs (SMT still doesn't really "do" context/situatedness)

- Grammar-based SMT methods tend to be somewhat better at capturing syntactically idiosyncratic MWEs

MWEs in Document Categorisation

- For document-level categorisation tasks, MWEs that cause the most problems are named entities (person names, company names, etc.)
- Named entity recognition (as pre-processing step or in tandem with document categorisation) tends to improve results

WRAP-UP

Overall Conclusions

- MWEs are frequent, fun and funky in all sorts of ways
- There's much, much more to MWEs than our old friend *kick the bucket*
- The impact of MWEs on IR and NLP applications varies widely
- Most of the research problems are far from resolved: lots of room for all to play!

References

Baldwin, Timothy, and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, ed. by Nitin Indurkhya and Fred J. Damerau. Boca Raton, USA: CRC Press, 2nd edition.

Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. Cambridge, USA: MIT Press.

Kastovsky, Dieter. 1982. *Wortbildung und Semantik*. Dusseldorf: Bagel/Francke.

Metzler, Donald, and W Bruce Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of 28th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 472–479, Salvador, Brazil.

Moon, Rosamund E. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford, UK: Oxford University Press.