Multiword Expressions

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

1. Introduction
2. Case Study 1: MWE or Construction?
3. Case Study 2: UD and MWEs
4. Summary

What are Multiword Expressions (MWEs)?

- *Definition:* A **multiword expression** (“MWE”) is:
  1. decomposable into multiple simplex words
  2. lexically, phonetically, phonologically, morphosyntactically, semantically, and/or pragmatically idiosyncratic

Adapted from Baldwin and Kim [2010]

Some Examples

- *ad hoc*, *by and large*, *The Chair*, *kick the bucket*, *part of speech*, *in step*, *trip the light fantastic*, *foundation model*, *call (someone) up*, *take a walk*, *do a number on (someone)*, *take advantage (of)*, *pull strings*, *kindle excitement*, *fresh air*, ...

Lexical Idiomaticity

- Lexical idiomaticity $=$ one or more of the elements of the MWE does not have a usage outside of MWEs
- Examples of lexical idiomaticity:

  ad hominem, bok choy, a la mode, to and fro
- Complications of lexical idiomaticity:
  - cross-linguistic effects, e.g. *ad* is unmarked in Latin
  - simple lexical occurrence outside of MWEs not sufficient, e.g. *a la mode*

Source(s): Bauer [1983], Trawiński et al. [2008]

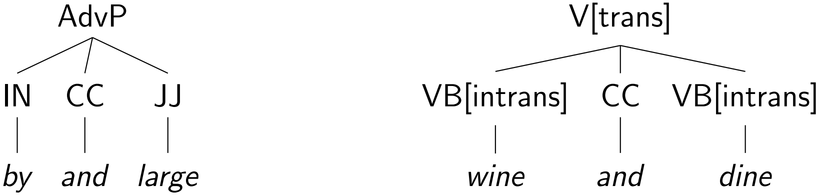
Phonetic and Phonological Idiomaticity

- Phonetic idiomaticity = one or more component elements of the MWE are pronounced in a manner specific to the MWE
- Examples of phonetic idiomaticity:

  bon voyage, 一期一会 *(ichi-go ichi-e)*

Morphosyntactic Idiomaticity

- Morphosyntactic idiomaticity = the morphosyntax of the MWE differs from that of its components
- Examples of morphosyntactic idiomaticity:
  cat's cradle, yin hry "evil eye"
- Examples of syntactic idiomaticity:



Source(s): Katz and Postal [2004], Chafe [1968], Bauer [1983], Sag et al. [2002], Al-Haj et al. [2014]

Semantic Idiomaticity

- Semantic idiomaticity = the meaning of the MWE is not the simple sum of its parts, in that:
  - there is a mismatch in simplex and MWE semantics for one or more of the components, e.g.

birds of a feather, blow hot and cold

OR

  - there is extra semantic content in the MWE not encoded in the parts, e.g.

designated driver vs. backseat driver vs. bus driver

Source(s): Katz and Postal [2004], Chafe [1968], Bauer [1983], Sag et al. [2002]

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
- The contexts/real-word information/expectations vary a lot in their generality and also strength:
    - societal norms (e.g. *all aboard*, *gin and tonic*)
    - sub-community norms (e.g. the Monty Python effect)
    - idiolectal norms

Source(s): Kastovsky [1982], Jackendoff [1997], Sag et al. [2002]

MWE Markedness

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

Challenges in Pinning down MWEs

- What is a “word”?
  - complications with non-segmenting languages (Japanese, Thai, ...) and languages without a pre-existing writing system (Walpiri, Mohawk, ...)
  - even in English: *houseboat* vs. *house boat*, *trade off* vs. *trade-off* vs. *tradeoff*
- What is (lexical|phonetic|phonological|morphosyntactic|semantic|pragmatic) idiosyncrasy?
- What is an MWE and what is (purely) constructional?
- How should MWEs be represented to capture their (cross-linguistic) idiosyncrasies (but also their compositionality)?

Source(s): Bond and Baldwin [2016], Mansfield [to appear]

Talk Outline

1. Introduction
2. Case Study 1: MWE or Construction?
3. Case Study 2: UD and MWEs
4. Summary

Definition

- Determinerless PPs (PP−Ds) are MWEs comprising a preposition (P) and a singular noun (N_{Sing}) without a determiner:

  in gaol, off screen, on ice, in town, by train, per student

- Most languages with articles have PP−Ds ... interesting to consider what the equivalent of a PP−D is for languages without articles (but does have rich morphology)
- Same basic semantic types attested in English, Albanian, Tagalog, German, et al. [Himmelmann, 1998]

Source(s): Baldwin et al. [2006], Stvan [1998]

The Syntax of PP−Ds

- Variability in syntactic markedness, productivity and nominal modifiability for different PP−D constructions
- Non-productive, non-modifiable PP−Ds: *ex cathedra*, *ad hominem*, *ad nauseum*
- Fully-productive, highly-modifiable PP−Ds: *per recruited student that finishes the project*
- Most PP−Ds lie between these two extremes

Source(s): Ross [1995]

Syntactic Markedness

- Syntactically-unmarked PP−Ds: N_{Sing} is uncountable

  *E.g. Institutions: in school, in gaol, but *in library (cf. school finished vs. *library finished)*

- Syntactically-marked PP−Ds: N_{Sing} is strictly countable

  *E.g. PPs headed by per: per person, but *per information (c.f. by bus/public transport)*

Nominal Modifiability

- No modification: _in *mental/*small hospital_
- Idiosyncratic modification: _at long/*lengthy/*short last_
- Relatively free modification: _at great/considerable/tedious length_
- Modification can be:
  - impossible, optional, or obligatory
  - nominal, adjectival, or both (or none)

	Obligatory	Optional	Impossible
Noun	_at *(eye) level_	_on (summer) vacation_	_on (*very) top_
Adjective	_at *(long) range_	_in (sharp) contrast_	
Either	_at *(company) expense_ _at *(considerable) expense_	_in (family) court_ _in (open) court_	

The Semantics of PP−Ds

- All PP−Ds show a certain degree of (generally systematic) semantic markedness on the noun:
  - institutional: *in hospital*, *at school*
  - metaphoric: *on ice*
  - generic uses: *by car*
- Some semantic systematicity with particular prepositions

Source(s): Stvan [1998]

PP–D: MWE or Construction?

- “Words with spaces” PP–Ds (*ad hominen*, *on top*) are clearly MWEs, fully-productive PP–Ds (*by* MEANS, *per* N_{+count}) are purely constructional
- Examples that lie between these extremes:

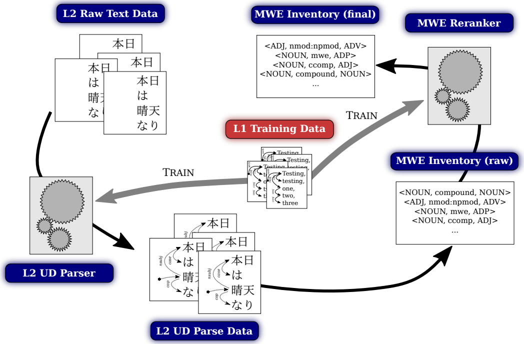
	Markedness				**MWE?**
	Lex	**MorphoSyn**	**Sem**	**Prag**	
on ice	☒	☒	☑	☒	☑
at large	☒	☑	☑	☒	☑
up front	☒	☑	☒	☒	☑
in winter	☒	☒	☒	☒	☒
at X level	☒	☒	☒	☒	☒
in gaol	☒	☒	?	☒	?

Talk Outline

1. Introduction
2. Case Study 1: MWE or Construction?
3. Case Study 2: UD and MWEs
4. Summary

Determining a Language’s MWE Inventory

- **Question:** given a raw text corpus in a given language and a Universal Dependency (“UD”) parser, can we automatically learn the inventory of MWE constructions?



<NOUN, compound, NOUN>
<ADJ, nmod:npmod, ADV>
<NOUN, mwe, ADP>
<NOUN, ccomp, ADJ>
...

<ADJ, nmod:npmod, ADV>
<NOUN, mwe, ADP>
<NOUN, ccomp, ADJ>
<NOUN, compound, NOUN>
...

Source(s): Salehi et al. [2016]

Talk Outline

1. Introduction
2. Case Study 1: MWE or Construction?
3. Case Study 2: UD and MWEs
4. Summary

Summary

- *Definition:* A **multiword expression** ("MWE") is:
  1. decomposable into multiple simplex words
  2. lexically, phonetically, phonologically, morphosyntactically, semantically, and/or pragmatically idiosyncratic
- BUT:
  - what is a word?
  - what is idiomaticity?
  - how to represent both idiomaticity and systematicity (e.g. in UD)?
  - what can we learn about MWEs based on a systematic typological investigation?

References

Hassan Al-Haj, Alon Itai, and Shuly Wintner. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170, 2014.

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

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier, editor, *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*, pages 163–180. Springer, Dordrecht, Netherlands, 2006.

Laurie Bauer. *English Word-formation*. Cambridge University Press, Cambridge, UK, 1983.

Francis Bond and Timothy Baldwin. Introduction to Japanese computational linguistics. In Francis Bond, Timothy Baldwin, Kentaro Inui, Shun Ishizaki, Hiroshi Nakagawa, and Akira Shimazu, editors, *Readings in Japanese Natural Language Processing*. CSLI Publications, Stanford, USA, 2016.

Wallace L. Chafe. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language*, 4:109–127, 1968.

Chitra Fernando and Roger Flavell. On idioms. In *Critical views and perspectives, volume 5 of Exeter Linguistic Studies*. Exeter: University of Exeter, 1981.

Nikolaus P. Himmelmann. Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology*, 2: 315–353, 1998.

References

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

Dieter Kastovsky. *Wortbildung und Semantik*. Bagel/Francke, Dusseldorf, Germany, 1982. (in German).

Jerrold J. Katz and Paul M. Postal. Semantic interpretation of idioms and sentences containing them. In *Quarterly Progress Report (70), MIT Research Laboratory of Electronics*, pages 275–282. MIT Press, 2004.

John Mansfield. The word as a unit of internal predictability. *Linguistics*, to appear.

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. Idioms. *Language*, 70:491–538, 1994.

Háj Ross. Defective noun phrases. In *Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, pages 398–440, 1995.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico, 2002.

Bahar Salehi, Paul Cook, and Timothy Baldwin. Determining the multiword expression inventory of a surprise language. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 471–481, Osaka, Japan, 2016.

Richard W. Sproat and Mark Y. Liberman. Toward treating English nominals correctly. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, USA, 1987.

References

Laurel Smith Stvan. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. PhD thesis, Northwestern University, 1998. URL `http://ling.uta.edu/~laurel/stvan98_overview.html`.

Beata Trawiński, Manfred Sailer, Jan-Philipp Soehn, Lothar Lemnitzer, and Frank Richter. Cranberry expressions in English and in German. In *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 35–38, Marrakech, Morocco, 2008.