Multiword Expressions at the Grammar–Lexicon Interface

Timothy Baldwin

THE UNIVERSITY OF
MELBOURNE

Talk Outline

What are Multiword Expressions (MWEs)?

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

Adapted from Baldwin and Kim [2010]

Some Examples

- *Mt Rokko*, *ad hoc*, *by and large*, *Toy Story*, *kick the bucket*, *part of speech*, *in step*, *Hanshin Tigers*, *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*, ...
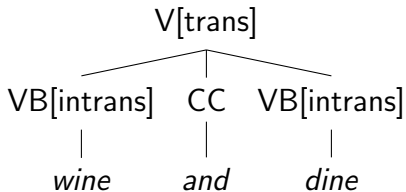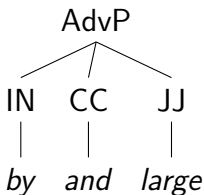
Lexical Idiomaticity

- 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]

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:

```
        AdvP                              V[trans]
       /  |  \                          /    |    \
     IN   CC   JJ                 VB[intrans] CC  VB[intrans]
     |    |    |                       |      |       |
     by  and  large                  wine   and     dine
```

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

Talk Outline

1. Introduction

2. MWEs at the Grammar–Lexicon Interface

3. A Case Study: English Determinerless PPs
   - Problem Statement
   - Analyses

4. Summary

MWEs and Productivity

- MWEs populate a spectrum of productivity, from completely unproductive, lexicalised MWEs such as *by and large* to highly productive (but still semi-lexicalised) MWEs such as English verb–particle constructions

- The first question in terms of the grammar–lexicon interface when it comes to MWEs, is the degree of productivity of a given MWE, to determine whether it is possible to enumerate all of the instances of a given MWE class, or if not, what the rules that govern the productivity of the MWE are

MWEs and Syntactic Flexibility

- The next question is what degree of morphosyntactic flexibility the (open or closed) class of MWEs has:
  - ▶ if none, treat as a "word-with-spaces" and map onto a pre-existing type in the grammar, where possible
  - ▶ if only morphological flexibility (e.g. *attorney general*), capture this appropriately and map to a type in the grammar
  - ▶ if morphosyntactic flexibility, ascertain the level of flexibility; most MWEs are somewhat constrained in their morphosyntactic flexibility, but working out the precise limits of flexibility is notoriously difficult (suggest erring on the side of under-constraint)

- Beware metalinguistic markers such as *proverbial*

Idioms

- Idioms are a fascinating class of MWE, in that they take many different forms lexically and syntactically, with syntactic flexibility determined through "decomposability" [Nunberg et al., 1994, Baldwin et al., 2003]:
    - if the semantics of the idiom can be (idiosyncratically) ascribed to parts of the MWE, those parts will tend to undergo (almost) unlimited syntactic flexibility
      spill the beans $=$ reveal′(secret′) $\approx$ spill′(beans′)
    - one approach of dealing with this in the grammar is to have an idiomatic lexical entry associated with each of the decomposed components, and constrain idiomatic parses to have all of the idiom parts in particular relational configurations

Talk Outline

Contents

Definition

- Determinerless PPs ("PP$-$Ds") are MWEs comprising a preposition (P) and a singular noun (N$_{Sing}$) without a determiner:

Institution	Media	Metaphor	Temporal	Means/Manner
at school	*on film*	*on ice*	*at breakfast*	*by car*
in church	*on TV*	*at large*	*at lunch*	*by train*
in gaol	*to video*	*at hand*	*on break*	*by hammer*
on campus	*off screen*	*at leave*	*by night*	*by computer*
at temple	*in radio*	*at liberty*	*by day*	*via radio*
. . .	. . .	. . .	. . .	. . .

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 seldom unrestricted, except for fully productive prepositions (e.g. *by*)

Types of Modification in PP−Ds

- Modification can be:
  - ▸ impossible, optional or obligatory
  - ▸ nominal, adjectival or both (or none)

	Obligatory	Optional	Impossible
Noun	*at *(eye) level*	*on (summer) vacation*	
Adjective	*at *(long) range*	*in (sharp) contrast*	*on (*very) top*
Either	*at *(company) expense* *at *(considerable) expense*	*in (family) court* *in (open) court*	

Contents

Analysis 1: Lexical Listing

- **Analysis:** all PP−Ds fully lexicalised
- **Pros:**
  - ▶ effective at capturing syntactically- and semantically-marked PP−Ds
- **Cons:**
  - ▶ unable to handle nominal modification
  - ▶ data sparseness effects with productive classes
  - ▶ inability to capture semantic compositionality

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

Analysis 2: Simple Combination

- **Analysis:** use head complement rule to combine simplex P and N lexical entries
- **Pros:**
  - ▶ Effective at capturing syntactically-unmarked PP−Ds
  - ▶ Licenses unconstrained nominal modification
- **Cons:**
  - ▶ Inability to capture non-compositional semantics
  - ▶ Overgeneration (*by/on/... school*)
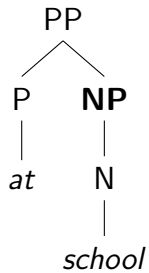
Analysis 3: Idiosyncratic Selection

- **Analysis:** introduce idiosyncratic lexical types for head nouns which specify constraints such that:
  1. phrases that it projects will only appear as complements of (certain) Ps;
  2. its specifier (determiner) will never be expressed
  3. it must combine with a (pre-head) modifier before it can combine with the P; and
  4. it can only appear with particular modifier types (e.g. only nouns)

Analysis 3: Idiosyncratic Selection

- Example for *at eye/street/... level*:

```
level_i_n1 := n_-_c-brn_le &
 [ STEM < "level" >,
   SYNSEM [ LKEYS.KEYREL.PRED "_level_n_1_rel",
            PHON.ONSET con ] ].

at+level := detless_pp_idiom_mtr &
  [ INPUT.RELS <! [ PRED _at_p_rel ],
                  [ PRED "_level_n_1_rel" ],
                  [ PRED idiom_q_i_rel ] !> ].
```

Analysis 3: Idiosyncratic Selection

- **Pros:**
  - ▶ captures idiosyncracies of modifiability, P-N collocation
- **Cons:**
  - ▶ lexical redundancy
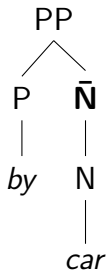 - ▶ overgeneration

Analysis 4: N̄ Selection

- **Analysis:** allow particular P senses (e.g. *by* MEANS) to select for unsaturated NPs (N̄s)
- **Pros:**
  - ▸ allows for nominal modification and full productivity
- **Cons:**
  - ▸ semantic restriction of the noun?

    (Dutch) *in* CLOTHING vs. *by* MEANS

$$\left[\text{SYN} \ \left[\text{CAT} \ \left[\text{HEAD} \ \textit{prep} \ \text{VAL} \ \left[\text{COMPS} \ \left\langle \left[\text{SPR} \ \langle \ \text{Det} \ \rangle \right] \right\rangle \right] \right] \right] \right]$$

Analysis 4: N̄ Selection

```
              PP
             /  \
            P    N̄
            |    |
           by    N
                 |
                car
```

Determining the Appropriate Analysis

- Consider:
  - ▶ degree and type of nominal modifiability
  - ▶ productivity relative to a given P
  - ▶ semantic markedness
  - ▶ NP saturation (syntactic markedness)

Talk Outline

Summary

- MWEs populate a broad spectrum, in terms of productivity, constructional composition, morphosyntactic flexibility, …
- Important to understand the full complexities of the data, to be able to customise the syntactic analysis to the MWE

References

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, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan, 2003.

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.

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

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.

References

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.

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.