Multiword Expressions

Timothy Baldwin

CSSE, University of Melbourne

THE UNIVERSITY OF MELBOURNE

Structure of Course

a. Introduction

b. Computational syntax

c. Extraction/identification

d. Computational semantics/interpretation

e. Decomposability/compositionality

INTRODUCTION

Introduction: Structure

- Definitions, properties of MWEs
- Computational challenges
- Component tasks
- MWEs in NLP applications

But First ...

- Research background:
  - ⋆ Multiword Expression (MWE) Project (CSLI, NTT CS Labs. and Cambridge University)
  - ⋆ Jointly funded by NSF and NTT CS Labs.
  - ⋆ Primary project aim is to investigate different means for encoding a variety of MWEs in precision grammars
  - ⋆ Visit us online at: `mwe.stanford.edu`

What are Multiword Expressions (MWEs)?

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

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

Some Examples

- *San Francisco, ad hoc, by and large, Where Eagles Dare, kick the bucket, part of speech, in step, the Oakland Raiders, trip the light fantastic, telephone box, call (someone) up, take a walk, do a number on (someone), take (unfair) 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	☑	☑	☑	☒	?

Indicators of MWE-hood

- Institutionalisation/conventionalisation

- Lexicogrammatical fixedness: formal rigidity, preferred lexical realisation, restrictions on aspect, mood, voice, etc.

  - ⋆ lexicogrammatically fixed MWE: *kick the bucket*
  - ⋆ lexicogrammatically fixed non-MWE: *look like*
  - ⋆ lexicogrammatically non-fixed MWE: *keep tabs on*

- Semantic/pragmatic non-compositionality: there is a mismatch between the semantics/pragmatics of the parts and the whole
  - ⋆ non-compositional MWE: *kick the bucket*
  - ⋆ compositional MWE: *at first*

- Syntactic irregularity:
  - ⋆ syntactically-irregular MWEs: *all of a sudden*, *the be all and end all of*
  - ⋆ syntactically regular MWEs: *kick the bucket*, *fly off the handle*

- Non-identifiability: meaning cannot be predicted from surface form
  - ⋆ idiom of decoding (non-identifiable): *kick the bucket*, *fly off the handle*
  - ⋆ idiom of encoding (identifiable): *wide awake*, *plain truth*

- Situatedness: the expression is associated with a fixed pragmatic point
  - ⋆ situated MWEs: *good morning, all aboard*
  - ⋆ non-situated MWEs: *first off, to and fro*

- Figuration: the expression encodes some metaphor, metonymy, hyperbole, etc
  - ⋆ figurative expressions: *bull market, beat around the bush*
  - ⋆ non-figurative expressions: *first off, to and fro*

- Single-word paraphrasability: the expression has a single word paraphrase
  - ⋆ paraphrasable MWEs: *leave out* = *omit*
  - ⋆ non-paraphrasable MWEs: *look up*
  - ⋆ paraphrasable non-MWEs: *take off clothes* = *undress*

- Proverbiality: the expression is used "to describe—and implicitly, to explain—a recurrent situation of particular social interest ... in virtue of its resemblance or relation to a scenario involving homely, concrete things and relations" (Nunberg *et al.* 1994)
  - ⋆ informality: the expression is associated with more informal or colloquial registers
  - ⋆ affect: the expression encodes a certain evaluation of affective stance toward the thing it denotes

- Prosody: the expression has a distinctive stress pattern which diverges from the norm
  - ⋆ prosodically-marked MWE: *soft spot*
  - ⋆ prosodically-unmarked MWE: *first aid, red herring*
  - ⋆ prosodically-marked non-MWE: *dental operation*

- Substitutability: MWEs characteristically stand in opposition to **anti-collocations**, i.e. expressions derived through synonym/word order substitution which occur with markedly lower frequency than the base MWE (or not at all):
  - non-substitutable MWEs: *many thanks* (cf. **several thanks*, **many gratitudes*)
  - substitutable MWEs: *salt and pepper* vs. *pepper and salt*
  - non-substitutable non-MWEs: *common platypus*

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*

Exercise: Spot the MWE

Expression	***MWE?***	***Markedness***				
		Lex	***Syn***	***Sem***	***Prag***	***Stat***
library card						
at arm's length						
old tree						
foreign direct investment						
the sun						
at [nine] o'clock						
to go bush						
give a demo						
kick the bucket						
once upon a time						
at home						
in the meantime						
to read Shakespeare						

MWEs vs. Collocations

- A collocation is an arbitrary and recurrent word combination
- Collocations can be semantically-marked (e.g. *dark horse*) but tend to be compositional (e.g. *strong coffee*)
- Collocations are generally contiguous (binary) word sequences (more often than not $\bar{\text{N}}$s)
- Word order variation/flexibility effects generally ignored in collocation research

MWEs vs. Terms

- (Technical) term = a lexical unit consisting of one or more words which represents a concept inside a domain

- Terminology research primarily interested in the synchronic dynamics of terminology, term formation and terminological variation

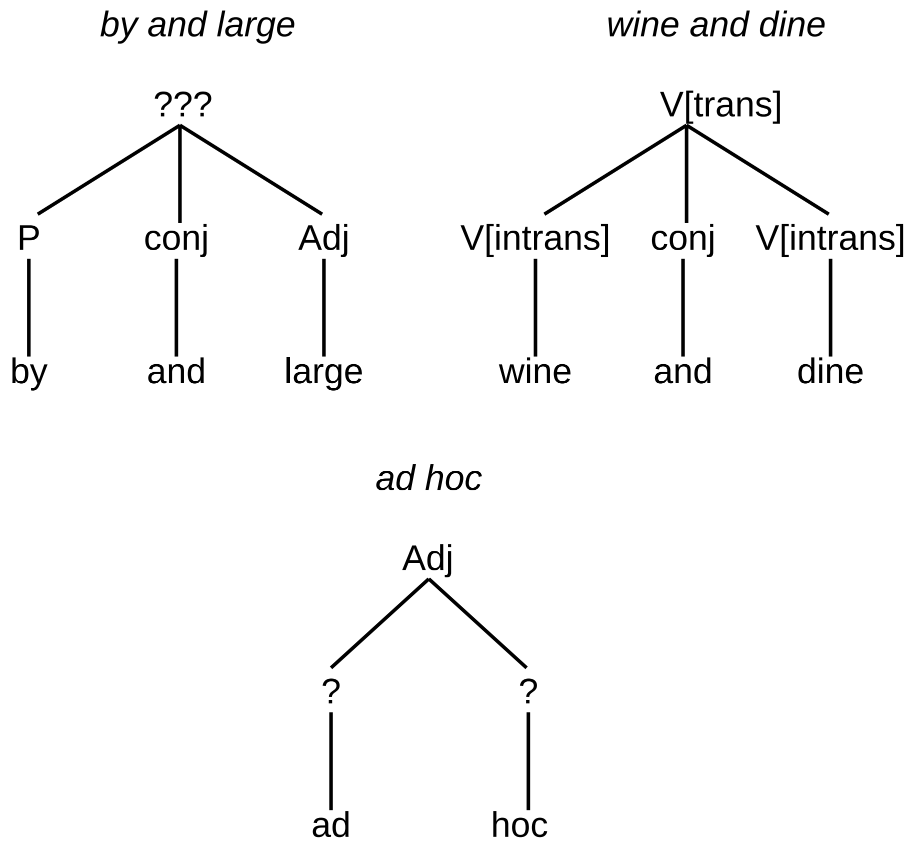
Challenges Posed by MWEs

- **Syntactic idiomaticity**
- **Semantic idiomaticity**
- **Pragmatic idiomaticity**
- **Statistical idiomaticity**
- **Flexibility**
- **Productivity**

Challenges Posed by MWEs

- **Syntactic idiomaticity**
- **Semantic idiomaticity**
- **Pragmatic idiomaticity**
- **Statistical idiomaticity**
- **Flexibility**
- **Productivity**

Syntactic Idiomaticity

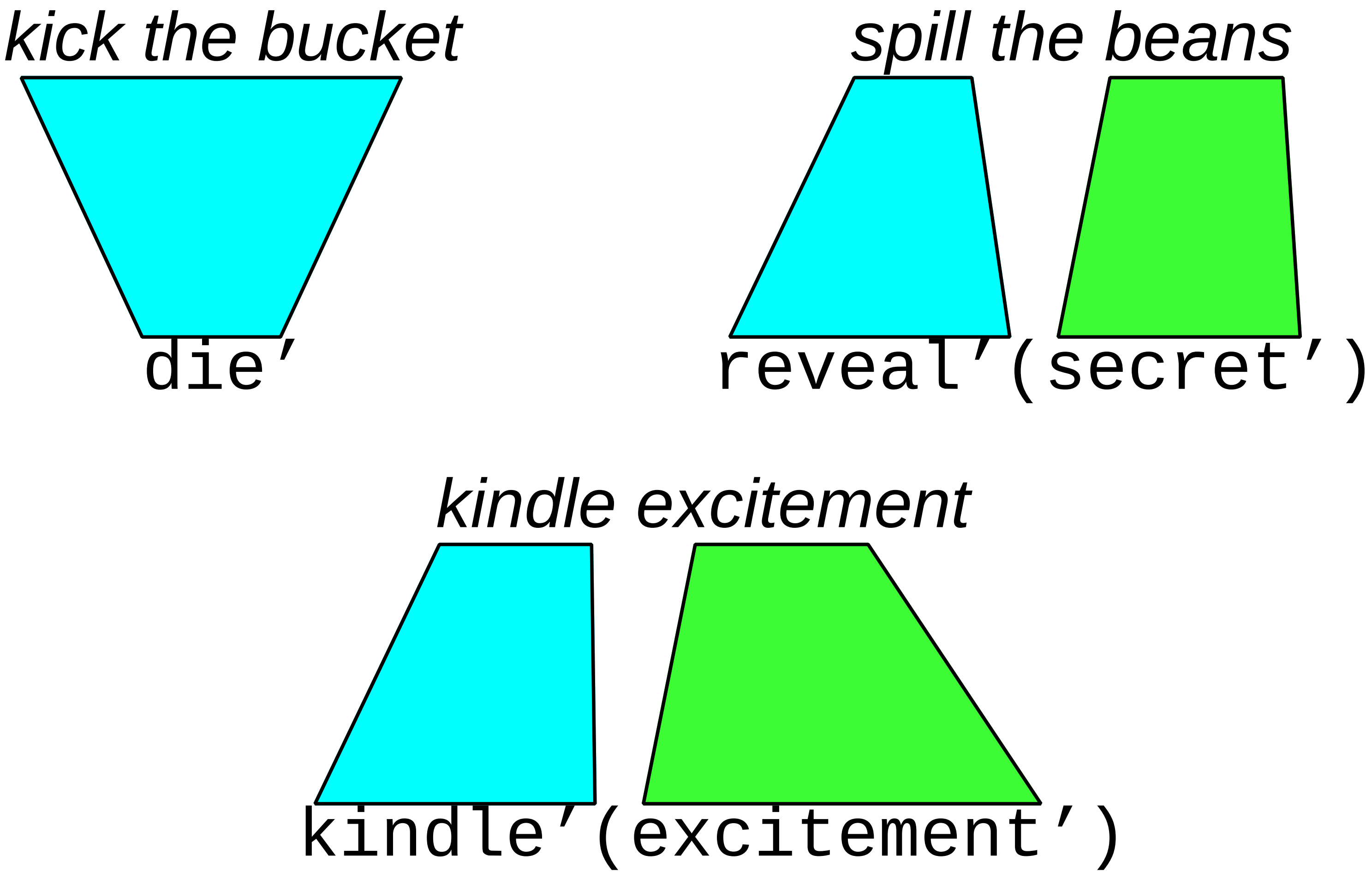


www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Challenges Posed by MWEs

- Syntactic idiomaticity
- Semantic idiomaticity
- Pragmatic idiomaticity
- Statistical idiomaticity
- Flexibility
- Productivity

Semantic Idiomaticity



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

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

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Challenges Posed by MWEs

- Syntactic idiomaticity
- Semantic idiomaticity
- **Pragmatic idiomaticity**
- Statistical idiomaticity
- Flexibility
- Productivity

Pragmatic idiomaticity

- The Wheel of Fortune factor — how to represent the jumble of phrases stored in the mental lexicon?

- The Monty Python factor — mish-mash of language fragments which evoke particular events/individuals/memories

Challenges Posed by MWEs

- Syntactic idiomaticity
- Semantic idiomaticity
- Pragmatic idiomaticity
- **Statistical idiomaticity**
- Flexibility
- Productivity

Statistical Idiomaticity

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

Adapted from Cruse (1986)

`www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf`

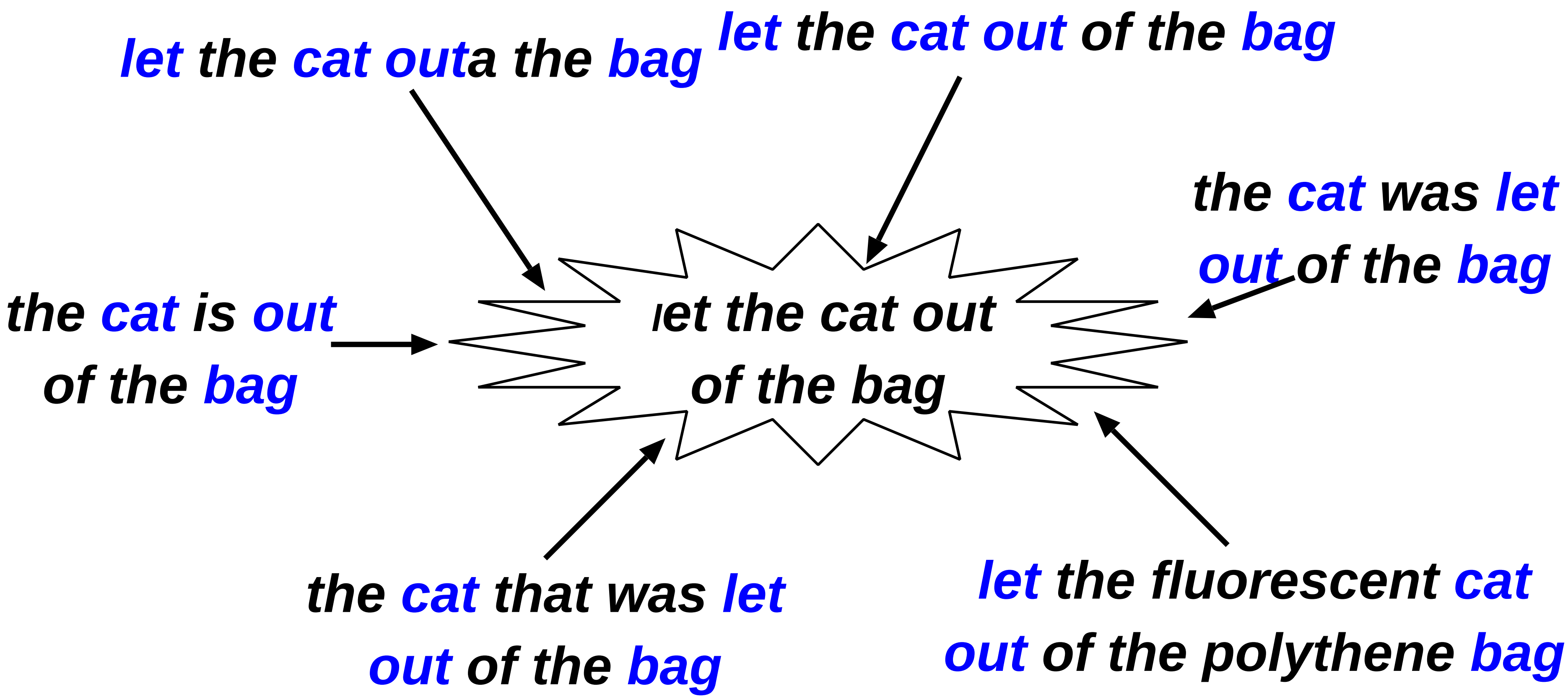
Statistical Idiomaticity and Dialect

- 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*

Challenges Posed by MWEs

- Syntactic idiomaticity
- Semantic idiomaticity
- Pragmatic idiomaticity
- Statistical idiomaticity
- **Flexibility**
- Productivity

Flexibility



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

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?*

Variation in Flexibility

- There is considerable variation in syntactic flexibility between constructions and also within a given construction type:

  a green pepper $\approx$ *a pepper which is green*
  a red herring $\neq$ *a herring which is red*
  the night is young $\neq$ *the young night*

  I handed in the paper $=$ *I handed the paper in*
  Kim ran over the dog $\approx^?$ *Kim ran the dog over*

Challenges Posed by MWEs

- Syntactic idiomaticity
- Semantic idiomaticity
- Pragmatic idiomaticity
- Statistical idiomaticity
- Flexibility
- **Productivity**

Productivity

- Varying level of productivity for different MWEs:

  *ad/post/*pre/*in/*apple/... hoc*
  *call/ring/phone/*telephone up*
  Melbourne train driver, human language technology, apple juice seat, ...

Motivation: Why Multiword Expressions?

- Pervasiveness in language
  - ⋆ MWEs estimated to be equivalent in number to simplex words in mental lexicon
- Volatility (domain tuning, terminology, ...)
  - ⋆ *axis of evil*, *make the pie higher*, *private equity*, ...

- Challenge to NLP systems
  - ⋆ language understanding
  - ⋆ fluency
  - ⋆ robustness
- Nice interaction between linguistics, statistics and computational linguistics
- MWEs in language learning environments

- Lots of interesting crosslingual commonalities/divergences
  - lexical equivalence: *in the red* vs. *no vermelho*
  - structural equivalence: *in the black* vs. *no azul*
  - semantic equivalence: *in a corner* vs. *encurralado*

Computational Tasks/Issues

- Parsing/identification
- Extraction
- Syntactic classification
- Semantic classification
- Representation
- Crosslingual approaches to MWEs

MWEs in NLP Applications

- IR (N-grams)
  - ⋆ phrase-based retrieval: mixed results (Salton and Smith 1990; Lewis and Croft 1990)
  - ⋆ query expansion: mixed results (Mandala *et al.* 2000)
  - ⋆ compound nominals more effective than simplex nominals as index terms (Wacholder and Song 2003)
- Tagging
  - ⋆ virtuous circle between MWE identification and

tagging accuracy (Piao *et al.* 2003)

- Parsing
  - ⋆ MWEs account for 8% of parsing errors with precision grammar (Baldwin *et al.* 2004)
  - ⋆ perfect knowledge of adverbial MWEs shown to enhance parser accuracy
- Information extraction
  - ⋆ collocations used extensively in IE tasks (Lin 1998b)
- Machine translation

* MWEs integral component of symbolic MT systems (Gerber and Yang 1997; Bond and Shirai 1997)

Summary

- What is an MWE?
- What properties are associated with MWEs?
- Why are MWEs challenging for NLP?
- What NLP applications do MWEs feature in?

COMPUTATIONAL SYNTAX OF MWEs

Case Study: English Resource Grammar

- HPSG-based linguistically-precise open-source grammar
- Compositional semantics based on MRS
- Reversible (parsing and generation)
- Medium coverage
- 8,218 types and 10,625 lexical entries (v06-jun-03)

English Resource Grammar (ERG) in Action

- *Leave the report on the desk*

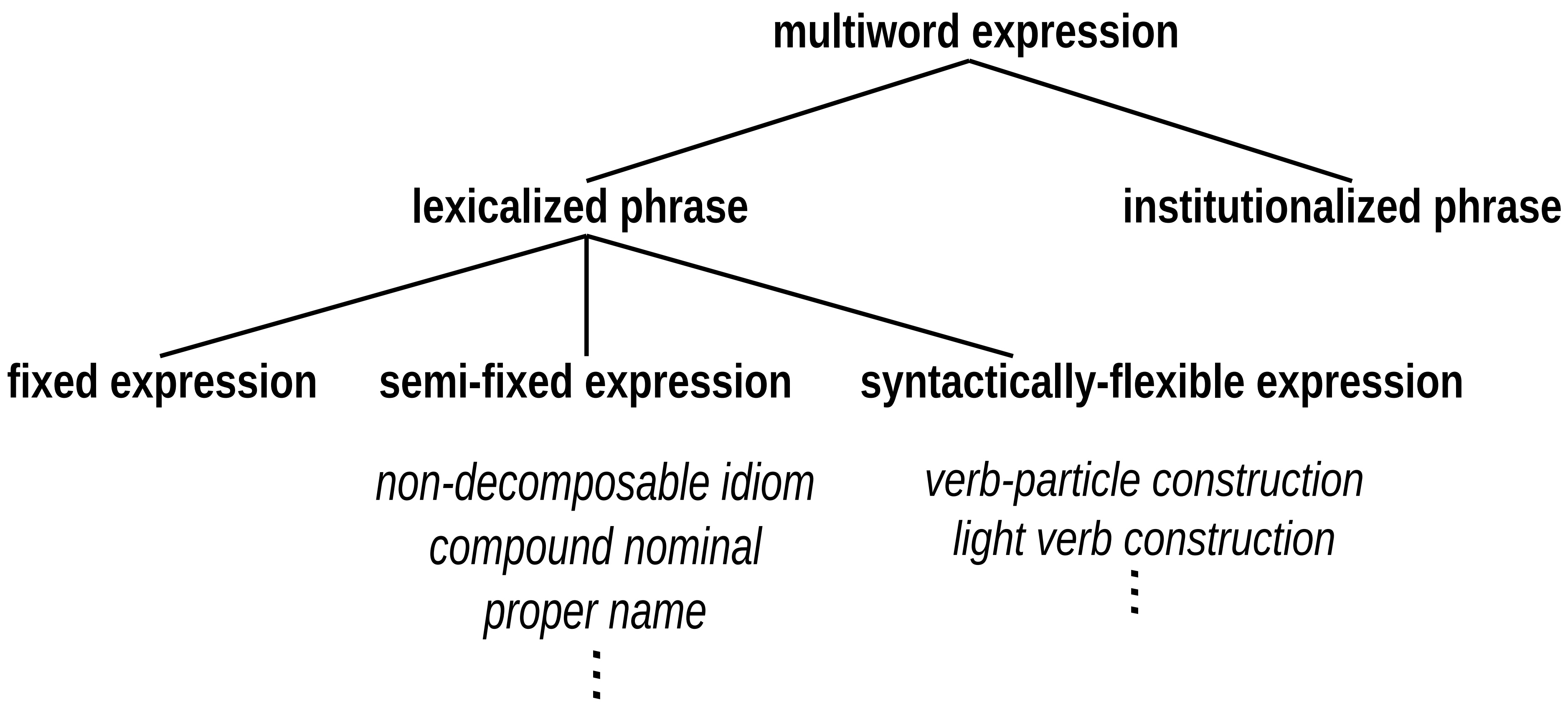
```
<h1,e2:PRESENT*:NO_ASPECT*:MOOD:BOOL,
 {h1:imp_m_rel(h3),
  h4:pronoun_q_rel(x6:2PER:REAL_GENDER:ZERO_PRON:-*, h5, h7),
  h8:pron_rel(x6),
  h9:_leave_rel(e2, x6, x11:-:REAL_GENDER:3SG*, v10),
  h12:_def_q_rel(x11, h14, h13),
  h15:_report_rel(x11, v16:BOOL),
  h15:_on_rel(e17:BOOL:NO_TENSE:ASPECT:MOOD, x11, x18:REAL_GENDER:3SG*:-),
  h19:_def_q_rel(x18, h21, h20),
  h22:_desk_rel(x18, v23:BOOL)},
 {h3 QEQ h9,
  h5 QEQ h8,
  h14 QEQ h15,
  h21 QEQ h22}>
```

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Basic Syntactic Approach

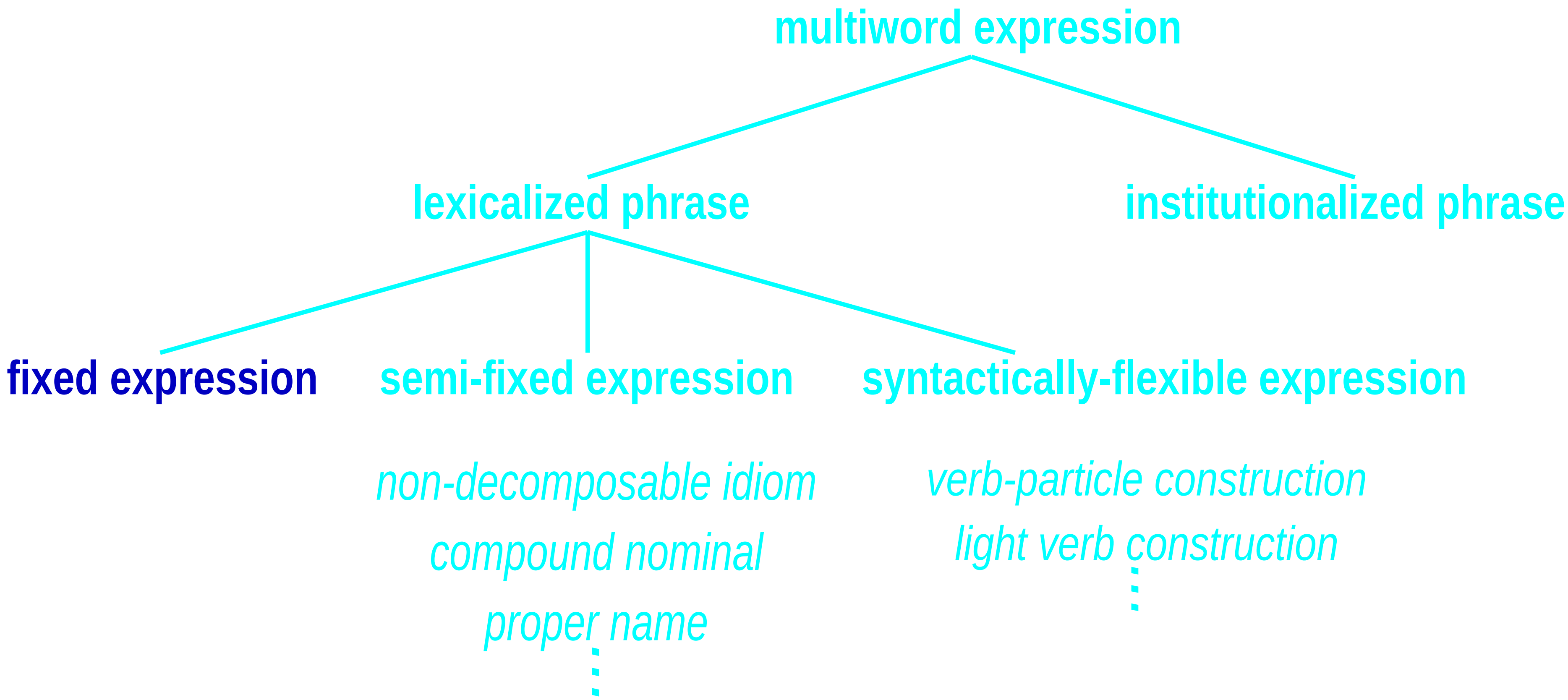
- Classify different MWE types according to their syntactic flexibility and productivity, and determine the appropriate analysis accordingly

MWE Types



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Fixed Expressions



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

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

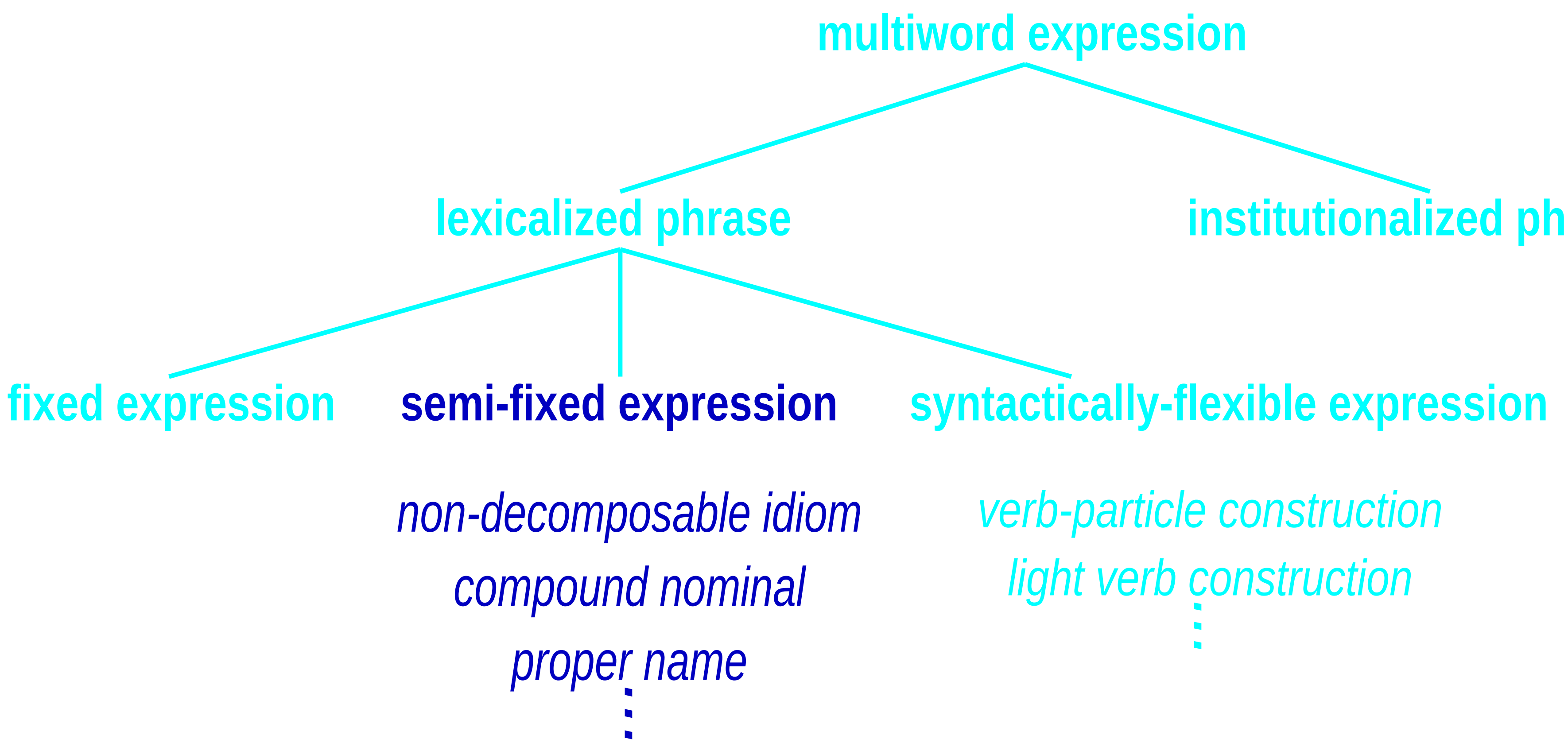
Fixed Expressions: Analysis

- Lexical entry for *ad hoc*:

```
ad_hoc_1 := intr_adj_le &
  [ STEM < "ad", "hoc" >,
    SEMANTICS [KEY ad_hoc_rel ]].
```

- Allows *very ad hoc*, but not **ad very hoc*.

Semi-fixed Expressions



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

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*

Semi-fixed Expressions: Analysis

- Treat as word complex which is lexically variable at particular positions:

```
part_of_speech_1 := n_intr_le &
  [ STEM < "part", "of", "speech" >,
    INFL-POS "1",
    SEMANTICS [KEY part_of_speech_rel ]] & / part_n1.

kick_the_bucket := v_unacc_le &
  [ STEM < "kick", "the", "bucket" >, INFL-POS "1",
    SEMANTICS [KEY kick_the_bucket_i_rel ]] &
    / kick_v_np*_trans_le.
```

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

U.S. Sports Team Names

- *the (Oakland) Raiders*
- *an/the [[(Oakland)Raiders] player]*
- *the [Raiders and 49ers].*
- *the league-leading (Oakland) Raiders.*
- *an [[(Oakland) Raider] spokesman]*
- **the Oakland 49ers*

U.S. Sports Team Names: Analysis

- Name: $\left[\texttt{SPR} \quad / \langle\ \rangle\right]$

- $\texttt{USTeamName:} \begin{bmatrix} \texttt{SPR} & \left\langle \begin{bmatrix} \texttt{Det} \\ \texttt{definite} \end{bmatrix} \right\rangle \\ \texttt{NUM} & / \ \texttt{plural} \end{bmatrix}$

- oakland_raiders_1 := $\texttt{USTeamName}$ &
$\begin{bmatrix} \text{LEX-SIGNS} & / \langle \text{ oakland\_1, raiders\_1 } \rangle \\ \text{SEMANTICS} & \langle \text{ oakland\_raiders\_rel } \rangle \end{bmatrix}$.

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 selection
  school bus
  orange juice seat

Compound Nouns: Analysis

- Constructional analysis: $\bar{\mathtt{N}} \longrightarrow \mathtt{N}\ \mathtt{N}$

- Underspecified `compound_rel` relation between nominal elements, e.g. *cardboard box*:

$$\mathtt{cardboard\_rel}(x) \wedge \mathtt{box\_rel}(y) \wedge \mathtt{compound\_rel}(x, y)$$

Syntactically-flexible Expressions

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

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*.
  - ⋆ *They fell quietly off the log*.
  - ⋆ *[Off how many logs] did the drunk fall?*
- **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 $\longrightarrow$ accomplishment)

- **Noncompositional**: *look up*, *throw up*

- *write up the memo/write the memo up*
  look up the answer/look the answer up

Verb-Particle Constructions: Analysis

- Verb selects for particle:

```
hand_out_v1 := v_particle_np_le &
  [ STEM < "hand" >,
    SEMANTICS [ KEY hand_out_rel,
                --COMPKEY out_rel ] ].
```

- Assume "joined" word order is canonical (Dehé 2002) and derive "split" word order by way of lexical rule

Verb-Particle Constructions: Productivity

- For fully/semi-productive verb-particles, avoid enumeration through use of lexical rules

- E.g., movement verb + directional particle:

  run/walk/... up/down/around/in/...

Light Verbs

- **Idiosyncracy:**

  make a mistake, **do a mistake*, **give a mistake*
  give a demo, *do a demo*, **make a demo*

- **Flexibility:**

  How many demos did Kim give?
  ...give a revealing demo
  A demo was given.

Light Verbs

- make_v1 := v_lite_l &

$$\begin{bmatrix} \text{STEM} & \langle \text{“make”} \rangle \\ \text{COMPLEMENTS} & \langle \text{NP[KEY } \mathit{make\_arg\_rel} \text{]} \rangle \end{bmatrix}.$$

$\mathit{make\_arg\_rel}$

$\mathit{mistake\_rel}$ $\mathit{argument\_rel}$ ...

- *make a mistake/error/boo-boo...*

- *make an argument/point/statement...*

Decomposable Idioms

- *take advantage (of), pull strings, keep tabs on*
- **Flexibility**:

  They regretted the considerable advantage that had been taken of the unfortunate situation.

  Strings had been pulled to get Sandy the job.

  The FBI kept closer tabs on Kim than they kept on Sandy.
- Flexibility is highly variable.

Decomposable Idioms: Analysis

- cat_out_of_bag :=

$$\left[\text{SEMANTICS} \left\{ \begin{bmatrix} \mathit{cat\_i\_rel} \\ \text{INST} \quad \text{x} \end{bmatrix}, \begin{bmatrix} \mathit{bag\_i\_rel} \\ \text{INST} \quad \text{y} \end{bmatrix}, \begin{bmatrix} \mathit{out\_rel} & \\ \text{ARG1} & \text{x} \\ \text{ARG2} & \text{y} \end{bmatrix} \ldots \right\} \right].$$

- cat_i :=

$$\left[\text{SEMANTICS} \left\langle \left[\mathit{cat\_i\_rel}\right] \right\rangle \right] \text{ \& / cat\_n1.}$$

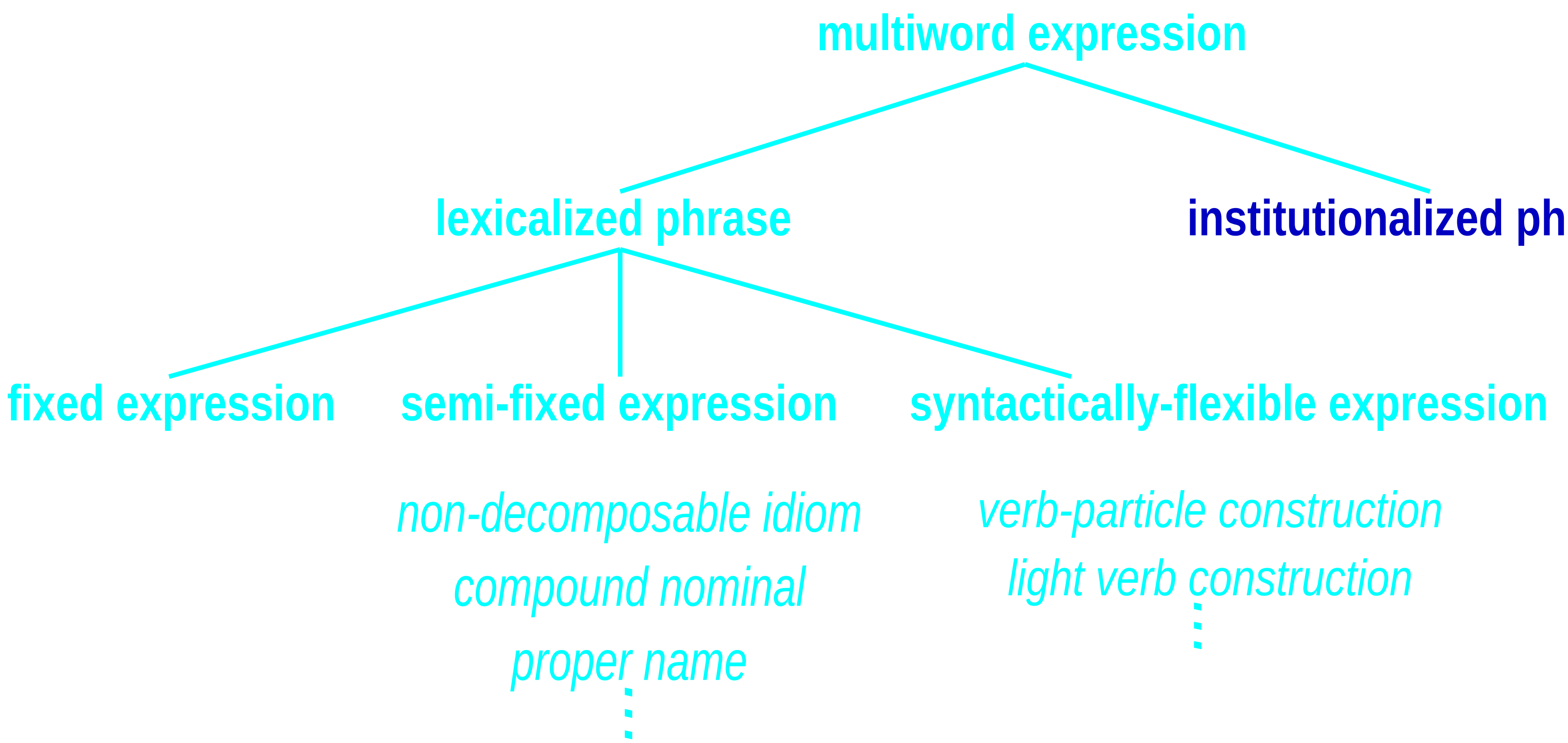
- bag_i :=

$$\left[\text{SEMANTICS } \left\langle \left[bag\_i\_rel \right] \right\rangle \right] \,\&\, / \text{ bag\_n1.}$$

- Use root conditions to ensure that all elements are present:

$$\left[\texttt{cat\_i}(x) \wedge \texttt{bag\_i}(y) \wedge \texttt{out}(x, y) \right]$$

Institutionalized Phrases



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

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 generation

Simple! ... or then Again?

- To date, we have proposed 4 basic analyses and categorised constructions according to the best fit with those 4 analysis types

- Not all constructions are this compliant!

Test Case: Determinerless PPs

Definition

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

Institutional	Media	Metaphor	Temporal	Means/Manner
at school	*on film*	*on ice*	*at breakfast*	*by car*
in church	*on TV*	*at large*	*on holiday*	*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*

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Crosslinguistic Occurrence of PP–Ds

- Most languages with articles have PP–Ds
- Same semantic types attested in English, Albanian, Tagalog, German, et al. (Himmelman 1998):
  - ⋆ Institution/Location: *at school*
  - ⋆ Metaphor/Abstract: *at large*
  - ⋆ Temporal: *in winter*
  - ⋆ Means/Manner: *by car*
- Focus of our research on English and Dutch

Corpus- and Lexicon-Occurrence of PP–Ds

- PP–Ds described statically in **COMLEX** and **WordNet**, but account for only around 30% and 15%, respectively, of the token occurrences of PP–Ds occurring $\geq$20 times in the BNC

- $\approx$0.3% of words in BNC are PP–Ds

- $\approx$0.2% of parse errors over a sample of the BNC caused by syntactically-marked PP–Ds

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

Syntactic Markedness

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

  E.g. Institutions: *in school*, *in gaol*, but **in office* (cf. *school is over* vs. **office is over*)

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

  E.g. PPs headed by *per*: *per person*, but **per information*

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 in productive constructions

Empirical Analysis of Modifiability

PP	Divergence $D(PP\|\|PP)$	$D(PP\|\|NP)$
on horseback	0.00	0.04
before dawn	0.00	0.16
to hospital	0.02	0.32
up front	0.03	0.26
on record	0.10	0.76
in diameter	0.14	0.54
in school	0.18	0.26
on loan	0.18	0.71
by decree	1.62	2.07
on analysis	4.29	2.81

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Modification Types

	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*	

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Coordination quirks

- Coordinate constructions: *from mother to child*, *room by room*.
- Partial selectional mismatches: *in brush and ink*
- Full selectional mismatches: *over mens en wereld* "about human being and world", *van stadion en hotel* "of stadium and hotel"

The semantics of PP–Ds

- All PP–Ds show a certain degree of (generally systematic) semantic markedness on the noun:
  - ⋆ institutional: *at school*
  - ⋆ metaphoric: *on ice*
  - ⋆ generic uses: *by car*

- Some semantic systematicity to the prepositions

PP−Ds with institutional nouns

- Activity enrichment: *in gaol* "while being a prisoner" and *in school* "while attending school", cf. *in a/the gaol*

- Familiarity enrichment: *John is in town* "John is in (my/his/this) town", cf. *in a/the town*

- Overlap between the two: *at work* "while at (my/his) work/working"

Metaphorical PP−Ds

- Examples:
  - ⋆ English – *at large*, *on ice*, *at last*
  - ⋆ Dutch – *op zak* "in pocket/possession", *aan zet* "(lit.) on turn"
- Non-compositional, but some degree of morphological systematicity in English: *lastly/at last*, *edgy/on edge*

PP–Ds with generic readings

- Examples: *by car*, *by hand*, *via email*
- Means/manner semantics of noun rare in subject/object position
- Resist referential uses and familiarity enrichments, but allow generic and activity readings:

  I travelled to San Francisco by car. They're/It's a great way to travel/#It rattled a lot.

Analysis 1: fixed expression

- Word-with-spaces analysis for fully lexicalised PP–Ds:

  at large, *on track*

- Prevents nominal modification, coordination

- Effective at capturing syntactically- and semantically-marked PP–Ds

- Hard to capture optional preposition selection properties (e.g. *on top (of)*, *in front (of)*)

Analysis 2: simple combination

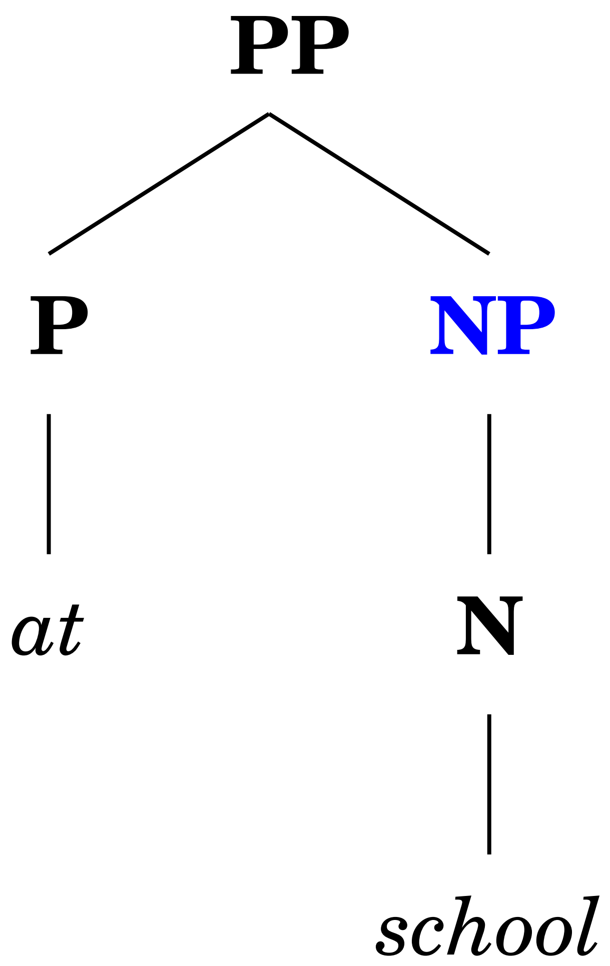
- Use head-complement rule to combine simplex P and N lexical entries:

  at church, in/after/during/... school

- Effective at capturing productive syntactically-unmarked PP−Ds

- Licenses unconstrained nominal modification

- Captures compositional semantics

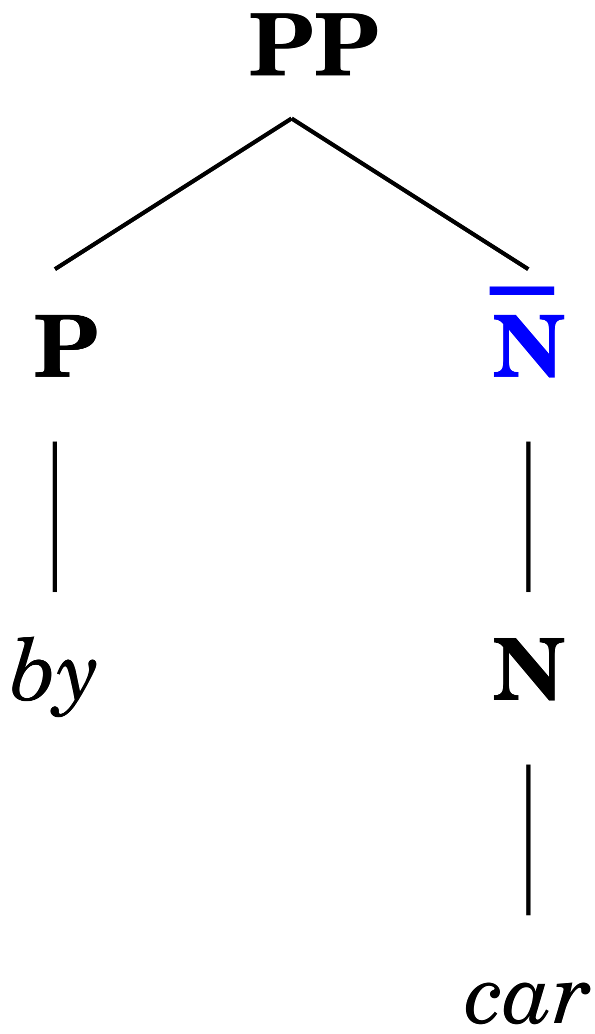

PP
P
NP
at
N
school

Analysis 3: $\bar{\mathrm{N}}$ selection

- License the preposition to select for an unsaturated NP ($\bar{\mathrm{N}}$):

  by train/plane/bus/hydrofoil/pogo stick...

- Allows for nominal modification and full productivity



www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Analysis 4: idiosyncratic modification

- Use idiomatic nominal lexical entries and unary rules to constrain the nominal modifier type (noun, adjective or neither):

  *at *(eye) level, on summer vacation, on (*very) top.*

- Specify preposition–noun combinatorics by way of root conditions (a lá decomposable idioms)

- Captures idiosyncratic modification effects

Determining the appropriate analysis

- Consider:
  - ⋆ nominal modifiability
  - ⋆ productivity
  - ⋆ semantic markedness
  - ⋆ NP saturation (syntactic markedness)

An Analytical Challenge

- Some strictly-countable nouns occur prolifically in PP$-$Ds without marked semantics, e.g.:

  $hand$, and Dutch zee “sea” and $huis$ “house”

Exercise: Pick the Right Analysis

MWE	***Analysis***				
	Fixed	***Semi-fixed***	***Syn-flex***	***Inst***	***Other***
push up the daisies					
break the ice					
in time					
pecking order					
from scratch					
bone of contention					
short shrift					
make short work of					
true candour					
blow one's top					
run up					

Summary

- What basic syntactic types of MWE are there?
- What issues do the various analyses address?
- How do we determine the correct analysis for a given MWE?
- Are MWEs really that well behaved?

EXTRACTING MWEs

Basic Task Description

- Identify the multiword expression (MWE) types in tagged (or raw) text from observation of the token distribution

- MWE (token) identification vs. (type) extraction

In American romance , almost nothing rates higher than what the movie men have called " meeting cute " – that is , boy-meets-girl seems more adorable if it does n't take place in an atmosphere of correct and acute boredom . Just about the most enthralling real-life example of meeting cute is the Charles MacArthur-Helen Hayes saga : reputedly all he did was give her a handful of peanuts , but he said simultaneously , " I wish they were emeralds " . Aside from the comico-romantico content here , a good linguist-anthropologist could readily pick up a few other facts , especially if he had a little more of the conversation to go on . The way MacArthur said his line – if you had the recorded transcript of a professional linguist – would probably have gone like this : A[fj] Primary stresses on emeralds and wish ; ; note pitch 3 (pretty high) on emeralds but with a slight degree of drawl , one degree of oversoftness . Conclusions : The people involved (and subsequent facts bear me out here) knew clearly the relative values of peanuts and emeralds , both monetary and sentimental . And the drawling , oversoft voice of flirtation , though fairly overt , was still well within the prescribed gambit of their culture . In other words , like automation machines designed to work in tandem , they shared the same programming , a mutual understanding not only of English words , but of the four stresses , pitches , and junctures that can change their meaning from black to white . At this point , unfortunately , romance becomes a regrettably small part of the picture ; ; but consider , if you can bear it , what might have happened if MacArthur , for some perverse , undaunted reason , had made the same remark to an Eskimo girl in Eskimo . To her peanuts and emeralds would have been just so much blubber

In American romance , almost nothing rates higher than what the movie men have called " meeting cute " – that is , boy-meets-girl seems more adorable if it does n't take place in an atmosphere of correct and acute boredom . Just about the most enthralling real-life example of meeting cute is the Charles MacArthur-Helen Hayes saga : reputedly all he did was give her a handful of peanuts , but he said simultaneously , " I wish they were emeralds " . Aside from the comico-romantico content here , a good linguist-anthropologist could readily pick up a few other facts , especially if he had a little more of the conversation to go on . The way MacArthur said his line – if you had the recorded transcript of a professional linguist – would probably have gone like this : A[fj] Primary stresses on emeralds and wish ; ; note pitch 3 (pretty high) on emeralds but with a slight degree of drawl , one degree of oversoftness . Conclusions : The people involved (and subsequent facts bear me out here) knew clearly the relative values of peanuts and emeralds , both monetary and sentimental . And the drawling , oversoft voice of flirtation , though fairly overt , was still well within the prescribed gambit of their culture . In other words , like automation machines designed to work in tandem , they shared the same programming , a mutual understanding not only of English words , but of the four stresses , pitches , and junctures that can change their meaning from black to white . At this point , unfortunately , romance becomes a regrettably small part of the picture ; ; but consider , if you can bear it , what might have happened if MacArthur , for some perverse , undaunted reason , had made the same remark to an Eskimo girl in Eskimo . To her peanuts and emeralds would have been just so much blubber

Complications in MWE Extraction

- Working out the extent of the collocation (phrase boundary detection)

  trip the light ☒

  trip the light fantastic ☑

  trip the light fantastic at ☒

- Fine line between collocations and simple default lexical combinations

  buy a car/purchase power

Statistical Tests Commonly Used

- Simple frequency: $f(x, y)$
- Pointwise/specific mutual information: $\log \frac{P(x,y)}{P(x)P(y)}$
- Dice's coefficient: $\frac{2\ f(x,y)}{f(x)f(y)}$
- (Student's) t score
- (Pearson's) chi-square (χ^2)
- Z score

`www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf`

- Log likelihood
- Selectional association

⋮

Finding of Evert and Krenn (2001) that simple frequency is as good as a wide range of collocation extraction measures over German Adj-N and P-N-V triple extraction tasks

Bigram Results from the WSJ

Rank	Frequency	Mutual information	χ^2	t test
1	*of the*	*Quadi Doum*	*Posse Comitatus*	*of the*
2	*in the*	*Wrongful Discharge*	*LORIMAR TELEPICTURES*	*in the*
3	*to NUMB*	*Seh Jik*	*Petits Riens*	*to NUMB*
4	*for the*	*Noo Yawk*	*Wrongful Discharge*	*on the*
5	*to the*	*WESTDEUTSCHE LANDESBANK*	*Tupac Amaru*	*the company*
6	*of NUMB*	*Naamloze Vennootschap*	*Sary Shagan*	*about NUMB*
7	*on the*	*Caisses Regionales*	*Outlaw Biker*	*said it*
8	*NUMB to*	*Centenaire Blanzy*	*GEMINI SOGETI*	*for the*
9	*that the*	*Guillen Landrau*	*Centenaire Blanzy*	*to be*
10	*the company*	*Ea Matsekha*	*Smith-Corona Typewriters*	*a share*
⋮				

Why Statistics?

- Pick up on word combinations which occur with "significantly" high relative frequency when compared to the frequencies of the individual words (i.e. $f(x, y)$ as compared to $f(x)$ and $f(y)$)

- Why so many different statistical tests?

  ⋆ complications in evaluation (hard to say which is the "best" test, conflicting results from different researchers)

- different corpora have different distributional idiosyncracies
- different tests have different statistical idiosyncracies

Collocation Extraction: Xtract (Smadja 1993)

Outline

- Automatic method for extracting collocations from raw text based on n-gram statistics
- **Basic intuition:** collocations are more rigid syntactically and more frequent than other word combinations
- **Method:** attempt to capture this intuition using the basic statistics of word combinations

Stage 1: Extract Significant Bigrams

- w and w_i **co-occur** (w_i is a **collocate** of w) if they are found in a single sentence separated by fewer than 5 words

- a bigram (w, w_i) is **significant** iff:
  - $\star$ w and w_i co-occur more frequently than chance
  - $\star$ w and w_i appear in a relatively rigid configuration

- Divide up the set of collocates according to POS

Example Corpus

$\ldots$ $multiword_{(-1)}$ $\underline{expressions}$ $\ldots$

$\ldots$ $multiword_{(-1)}$ $\underline{expressions}$ $\ldots$

$\ldots$ $dialect_{(-1)}$ $\underline{expressions}$ $\ldots$

$\ldots$ $dialect_{(-2)}$ and $\underline{expressions}$ $\ldots$

$\ldots$ $\underline{expressions}$ of $interest_{(2)}$ $\ldots$

$\ldots$ $multiword_{(-1)}$ $\underline{expressions}$ $\ldots$

$\ldots$ $collocation$ $extraction$ $\ldots$

$\ldots$ $\underline{expressions}$ $dialect_{(1)}$ $\ldots$

$\ldots$ $multiword_{(-1)}$ $\underline{expressions}$ $\ldots$

$\ldots$ $multiword_{(-1)}$ $\underline{expressions}$ $\ldots$

Noun Co-occurrence Table

$w = expressions$, $\mathcal{D} = \{-2, -1, 1, 2\}$

collocate	$multiword$ w_1	$dialect$ w_2	$collocation$ w_3	$interest$ w_4
p_i^{-2}	0	1	0	0
p_i^{-1}	5	1	0	0
p_i^{1}	0	1	0	0
p_i^{2}	0	0	0	1
$freq_i$	5	3	0	1
$w_i \in \mathcal{C}$	yes	yes	no	yes

Statistics of Expectation

- $\overline{f} = \frac{\sum_{w_i \in \mathcal{C}} freq_i}{|\mathcal{C}|} = \frac{5+3+1}{3} = 3$ (frequency average)
- $\sigma = \sqrt{\frac{\sum_{w_i \in \mathcal{C}} (freq_i - \overline{f})^2}{|\mathcal{C}|}} = \sqrt{\frac{(5-3)^2+(3-3)^2+(1-3)^2}{3}} \approx 1.63$
- $k_i = \frac{freq_i - \overline{f}}{\sigma}$ (strength)
- $\overline{p}_i = \frac{\sum_{j \in \mathcal{D}} p_i^j}{|\mathcal{D}|}$ (pair count average)
- $U_i = \frac{\sum_{j \in \mathcal{D}} (p_i^j - \overline{p}_i)^2}{|\mathcal{D}|}$ (pair count variance)

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Collocation Filters

- Strength: $k_i > k_\alpha (= 1)$

  ➔ select frequent collocates

- Spread: $U_i > U_0 (= 1)$

  ➔ select spiky distributions

- Peakiness: $p_i^j \geq \overline{p}_i + (k_\beta \times \sqrt{U_i})$ $k_\beta = 0.5$[1]

  ➔ identify interesting spikes

[1] Value of 10 suggested for k_β in Smadja (1993)

Back to our Example: Strength

- w_1 $(multiword)$
  - $\star$ $k_1 = \frac{5-3}{1.63} = 1.22 > 1$ ☑
- w_2 $(dialect)$
  - $\star$ $k_2 = \frac{3-3}{1.63} < 1$ ☒
- w_4 $(interest)$
  - $\star$ $k_3 = \frac{1-3}{1.63} < 1$ ☒

Spread and Peakiness

- w_1 $(multiword)$
  - $\overline{p}_1 = \frac{0+5+0+0}{4} = 1.25$
  - $U_1 = \frac{(0-1.25)^2+(5-1.25)^2+(0-1.25)^2+(0-1.25)^2}{4} \approx 20.31 > 1$ ☑

$\overline{p}_i + (k_\beta \times \sqrt{U_i})$	$1.25 + (0.5 \times \sqrt{20.31}) \approx 3.50$
p_1^{-2}	$0 < 3.50$ ☒
p_1^{-1}	$5 \geq 3.50$ ☑
p_1^{1}	$0 < 3.50$ ☒
p_1^{2}	$0 < 3.50$ ☒

Stage 2: Bigrams to N-grams

- Independent filter to detect larger N-grams

- Method: for each fixed-distance collocate (w, w_i^j), extract out contiguous word sequences where $\max(p(word[i])) > T(= 0.75)$

Example

$w = resistance$, $w_i^{-3} = path$

Concordances from the BNC:

$$\begin{array}{r}
\ldots\ trod\ the\ path_{(-3)}\ of\ least\ resistance\ ,\ \ldots \\
\ldots\ finding\ the\ path_{(-3)}\ of\ least\ \underline{resistance}\ will\ \ldots \\
\ldots\ along\ the\ path_{(-3)}\ of\ least\ \underline{resistance}\ . \\
\ldots\ the\ safest\ path_{(-3)}\ of\ least\ \underline{resistance}\ through\ \ldots \\
\ldots\ took\ the\ path_{(-3)}\ of\ least\ \underline{resistance}\ and\ \ldots
\end{array}$$

➔ *the path of least resistance* is a rigid noun phrase

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Reflections

- (At the time) groundbreaking research on collocation extraction

- Not effective at extracting out low-frequency words

- Difficulties in evaluating the results of collocation extraction (applies to this day)

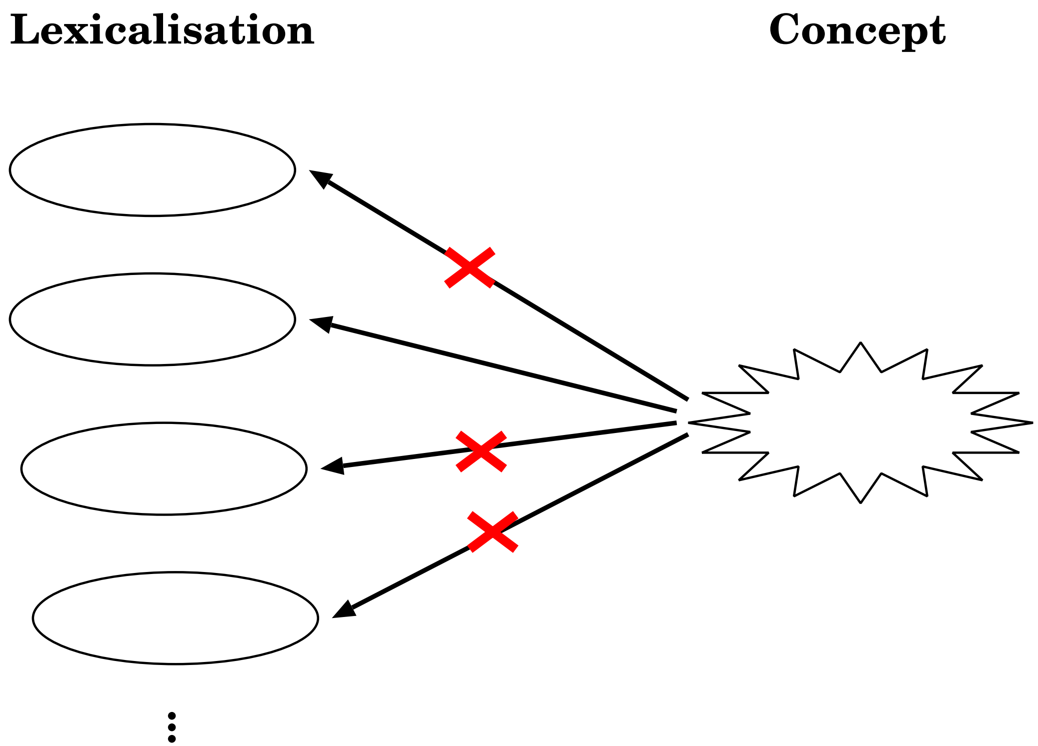
- Difficulties in extracting non-contiguous (predicative) collocations such as verb particles

Linguistics in Collocation Extraction

- Apply statistical measures to (head) bigrams in a given dependency relation (e.g. subject-verb)
  - ⋆ filters out stop words, produces "collocations" of pre-defined type for direct use in parsing, etc
- Look beyond contiguous bigrams, to bigrams occurring within a "collocational window" of fixed size (e.g. within 3-4 words of each other)
- Utilise linguistic qualities of collocations:

- limited internal modifiability (applicable as a post-filter)
- limited substitutability (contrast with anti-collocations, e.g. *(strong/*powerful) coffee*)
- non-compositional semantics

Substitutability



Substitutability

- Most immediate means of testing substitutability via synonyms (Pearce 2001b)
- Synonyms accessible from thesauri, but word sense disambiguation is generally needed to isolate which synset(s) over which to apply substitution test
- Possibilities of getting at synonyms via distributional analysis (possibly based on dependency pairs)

Collocation Extraction and Evaluation

- Difficulties in evaluation collocation extraction techniques due to lack of gold-standard datasets (what is MWE?)
- Precision generally evaluated according to pre-compiled LR or relative to corpus
- How to evaluate recall?
- How much is good enough?

Verb-particle Extraction (Baldwin and Villavicencio 2002; Baldwin (to appear))

Verb-particle Constructions (VPCs)

- VPC = verb + obligatory particle(s)
  - **intransitive:**

    Kim ***calmed down*** `v_particle_le`

  - **transitive:**

    Kim ***handed in*** *the paper* | `v_particle_np_le`
    Kim ***handed*** *the paper* ***in*** | `v_particle_np_le`

    Kim ***gets*** *Sandy* ***down*** `v_np_prep_particle_only_le`

Linguistic Properties of VPCs

- Transitive VPCs undergo the particle alternation (***hand in*** *the paper* vs. ***hand*** *the paper* ***in***)
- With transitive VPCs, pronominal objects must be expressed in the split configuration (***hand*** *it* ***in*** vs. *****hand in*** *it*)
- Manner adverbs cannot occur between the verb and particle (*****hand*** *it promptly* ***in***)

Extracting VPCs: Task Description

- Extract out full list of VPCs attested in a given corpus (cf. generation of independent list of VPCs)
- Make no assumptions about corpus annotation (use only information from pre-processors)
- Base extraction method on basic linguistic properties of VPCs
- Develop technique to be robust over low-frequency VPCs

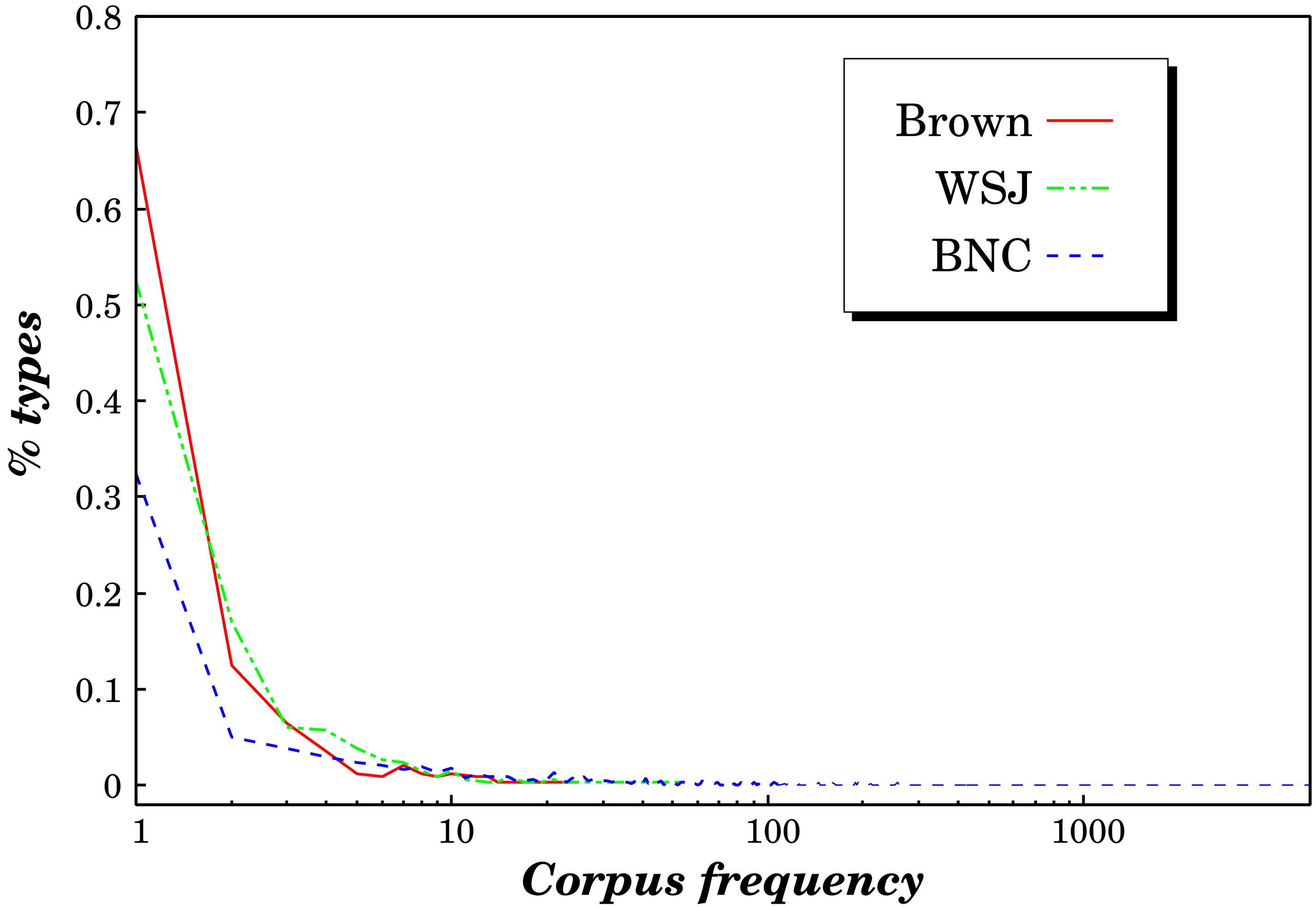
The Joys of VPC Extraction

- Limited coverage of linguistic tests
- Variable word order
- Variable window length
- Structural/analytical ambiguity:
  - ⋆ *hand [the paper] [in] [here]* vs. *hand [the paper] [in here]* vs. *hand [the paper in here]*
  - ⋆ *hand [in] [the paper]* vs. *hand [in the paper]*

Corpus Analysis of VPCs

- Generate gold-standard VPC data by taking intersection of VPCs in **Alvey Tools data**, **COMLEX v3.0** and **ERG** (total of 3,205 entries)
- Take random sample of 1,000 VPCs (1,577 LEs) and manually check for occurrences of each lexical entry (valence) in the **Brown Corpus**, **WSJ** and **BNC**
- Estimate the frequency of attested VPCs by voting across a range of extraction methods (explained later)

VPC Frequency Distribution



Analysis of VPC Results

Corpus	*Attested LEs*	*Ave. freq*	*Median freq*
Brown	21.4%	2.3	1
WSJ	21.2%	3.4	1
BNC	69.9%	89.6	7

Why do We Need Extraction?

- Rather then extracting VPCs, why not just use a pre-compiled broad-coverage, general-purpose VPC dictionary?

Resources	VPCs	Verbs	Particles
A+C+E	3,156	1,400	45
BNC	7,070	2,542	48
A+C+E ∩ BNC	2,014	1,149	28
A+C+E - BNC	1,138	251	17
BNC - A+C+E	5,056	1,393	20
A+C+E+BNC	8,208	2,793	65

Tasks

a. **Shallow lexical acquisition:** extraction of VPC types without valence information (e.g. *calm down*, *hand in*)

b. **Deep lexical acquisition:** extraction of VPC lexical entries (e.g. *calm down* = `v_particle_le` $\wedge$ `v_particle_np_le`, *blaze away* = `v_particle_le`)

Evaluation

- Use standard measures of **precision**, **recall** and **F-score** ($\beta = 1$)

- Calculate relative to the manually-determined corpus attestations of the 1,000 VPCs (for each corpus), cast in terms of the relative task

Classifier design

a. Generate feature vectors based on various statistics of VPC occurrence from training and test corpora, and build classifier using TiMBL v4.2 (k-NN)

b. Evaluation according to 10-fold cross validation

- hold out test VPCs in training corpus data
- test corpus annotations only used in evaluation (not in training data)

Training data

a. **Corpus attestation data:** manually-annotated corpus attestation of 1,000 VPCs/1,577 LEs

b. **Gold-standard dictionary data:** 3,205 valence-annotated VPC types (4,597 LEs)

- apply closed-world assumption in classifying VPC training data

Method-1: Simple POS-based Extraction

- Identify particles using dedicated POS tag (`RP` in Penn and CLAWS2 tagsets)
- PROCEDURE:

  a. tag the data using a tagger and lemmatise using morph
  b. for each particle, search back to the left up to 6 words to find governing verb
  c. filter data according to set of 73 canonical particles
  d. classify as transitive if split or immediately followed by NP, otherwise intransitive

Method-1: Feature Representation

- Features describing frequency of intransitive and transitive VPC types:

  `INTRANS TRANS`

Method-1: Example

country_NN fund_NNS offer_VBP an_DT easy_JJ way_NN to_TO get_VB a_DT taste_NN of_IN foreign_JJ stock_NNS without_IN the_DT hard_JJ research_NN of_IN seek_VBG out_RP individual_JJ company_NNS ._.

Method-1: Example

country_NN fund_NNS offer_VBP an_DT easy_JJ way_NN to_TO get_VB a_DT taste_NN of_IN foreign_JJ stock_NNS without_IN the_DT hard_JJ research_NN of_IN seek_VBG out_RP individual_JJ company_NNS ._.

Method-1: Taggers

mxpost: Penn-based MaxEnt tagger

fnTBL: Penn-based TBL tagger

RASP: CLAWS2-based HMM tagger used in RASP

Method-1: Results (Corpus-based)

Tagger	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
mxpost	Shallow	.973	.877	.922
	$\text{Deep}_{intrans}$	.570	.447	.502
	Deep_{trans}	.886	.824	.854
fnTBL	Shallow	.979	.825	.896
	$\text{Deep}_{intrans}$	.573	.447	.503
	Deep_{trans}	.894	.776	.831
RASP	Shallow	.971	.735	.837
	$\text{Deep}_{intrans}$	.600	.525	.560
	Deep_{trans}	.834	.707	.765

Method-1: Results (Dictionary-based)

Tagger	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
mxpost	Shallow	.973	.876	.922
	$\text{Deep}_{intrans}$	.645	.658	.651
	Deep_{trans}	.871	.842	.857
fnTBL	Shallow	.979	.822	.894
	$\text{Deep}_{intrans}$	.663	.627	.644
	Deep_{trans}	.856	.832	.844
RASP	Shallow	.963	.537	.690
	$\text{Deep}_{intrans}$	.652	.451	.533
	Deep_{trans}	.829	.442	.577

Method-1: Results

- Good results for the shallow task and transitive VPCs, lesser so for intransitive VPCs
- mxpost and fnTBL roughly equivalent, RASP significantly worse
- Corpus-based training data generally produces higher precision, dictionary-based training data higher recall

Method-2: Simple Chunk-based Extraction

- Identify particles using dedicated CoNLL-2000 chunk tag (`PRT`)

- PROCEDURE:

  a. chunk-parse tagged/lemmatised data using fnTBL
  b. for each (canonical) particle, search back to the left up to 6 words to find governing verb
  c. only allow noun, preposition and adverb chunks between verb and particle
  d. valence determination similar to POS-based extraction

Method-2: Feature Representation

- Features describing frequency of intransitive and transitive VPC types:

  $\texttt{INTRANS}\ \texttt{INTRANS}_L\ \texttt{INTRANS}_{\%}\ \texttt{TRANS}\ \texttt{TRANS}_L\ \texttt{TRANS}_{\%}\ \texttt{MI}$

  where:

  $$\texttt{(IN)TRANS}_L = \text{freq(linguistic test data)}$$
  $$\texttt{(IN)TRANS} = \text{freq(other VPC instances)}$$
  $$\texttt{(IN)TRANS}_{\%} = \frac{\text{freq((IN)TRANS)}}{\text{freq(INTRANS)+freq(TRANS)}}$$
  $$\texttt{MI} = MI(V;P)$$

Method-2: Example

$[_O$ “$]$ $[_{PP}$ instead of$]$ $[_{VP}$ buy$]$ $[_{NP}$ mask$]$ $[_{PP}$ for$]$ $[_{NP}$ your kid$]$ $[_O$,$]$ $[_{ADVP}$ just$]$ $[_{VP}$ cut$]$ $[_{PRT}$ out$]$ $[_{NP}$ the columnist$]$ $[_{NP}$ ’ picture$]$ $[_O$ ...$]$ $[_O$.$]$

Method-2: Example

[$_{O}$ “] [$_{PP}$ instead of] [$_{VP}$ buy] [$_{NP}$ mask] [$_{PP}$ for] [$_{NP}$ your kid] [$_{O}$,] [$_{ADVP}$ just] [$_{VP}$ cut] [$_{PRT}$ out] [$_{NP}$ the columnist] [$_{NP}$ ’ picture] [$_{O}$ ...] [$_{O}$.]

Method-2: Results

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
Corpus	Shallow	.991	.736	.845
	Deep$_{intrans}$	.596	.489	.537
	Deep$_{trans}$	.936	.724	.817
Dict	Shallow	.987	.735	.842
	Deep$_{intrans}$	.634	.614	.624
	Deep$_{trans}$	.881	.751	.811

Method-2: Results

- Higher precision than Method-1, but recall goes down considerably

  cause: low chunk recall over particles

- Slightly disappointing results

Method-3: Chunk Grammar-based

- Improve recall by looking also at canonical particles occurring as non-particle (`PP`, `ADV`) chunks
- Use chunk grammar to determine the syntactic relation between verbs and "particle candidates"
- Classify instances as:
  - ⋆ unambiguously intransitive/transitive VPC
  - ⋆ unambiguously intransitive/transitive non-VPC
  - ⋆ possible intransitive/transitive VPC

Method-3: Identifying VPCs

- Use chunk grammar to:
  - ⋆ check that the chunks which occur between the verb and particle are maximally an NP and particle pre-modifier adverb chunk (*back*, *right*, ...)
  - ⋆ check for a clause boundary or NP immediately after the particle/preposition/adverb chunk
  - ⋆ check the clause context of the verb chunk for possible extraposition of an NP verbal complement
- Check congruity with linguistic properties of VPCs

Method-3: Structural Ambiguity

[$_{NP}$ we] [$_{VP}$ may ask] [$_{NP}$ question] [$_{SBAR}$ as] [$_{NP}$ you] [$_{VP}$ go] [$_{ADVP}$ along] [$_{O}$,] ... ✔

[$_{NP}$ it] [$_{VP}$ wo n't do] [$_{NP}$ any good] [$_{PP}$ for] [$_{NP}$ anybody] [$_{SBAR}$ unless] [$_{NP}$ employee] [$_{VP}$ know] [$_{PP}$ about] [$_{NP}$ it] [$_{O}$.] ✗

[$_{VP}$ nonperform] [$_{NP}$ loan] [$_{VP}$ will make] [$_{PP}$ up] [$_{NP}$ only about 0.5 %] [$_{PP}$ of] [$_{NP}$ the combine bank] [$_{NP}$'s total loan] [$_{ADJ}$ outstanding] ... ???

Method-3: Attachment Disambiguation

- For cases of structural ambiguity, attempt to resolve using log likelihood ratio (verb–particle (VP), verb–NP$_1$ head (VN_1), NP$_1$ head–particle (N_1P) and particle–NP$_2$ head (PN_2):

  [$_{VP}$ hand] [$_{NP_1}$ the paper] [$_{PP}$ in] [$_{NP_2}$ here]

  VPC realised iff:

  $$VP \times VN_1 > N_1P \times PN_2$$
  $$VP \times VN_1 > VP \times PN_2$$

Method-3: Feature Representation

- Features describing frequency of positive/negative diagnostics for each (intrans/trans) VPC type:

$\texttt{INTRANS}_{+}$ $\texttt{INTRANS}_{-}$ ($\texttt{INTRANS}_{ATT}$) $\texttt{TRANS}_{+}$ $\texttt{TRANS}_{-}$ ($\texttt{TRANS}_{ATT}$)

Method-3: Results

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
Corpus	Shallow	.984	.891	.935
	Deep$_{intrans}$	.670	.848	.748
	Deep$_{trans}$	.889	.821	.853
Dict	Shallow	.982	.790	.876
	Deep$_{intrans}$	.753	.672	.710
	Deep$_{trans}$	.877	.762	.815

Method-3: Results

- Appreciable gain in recall over Method-1 and Method-2 (greater robustness over low-frequency data)
- More credible results over deep processing tasks
- Corpus-based training data markedly better than dictionary-based training data

Method-4: Parser-based

- Use full parser to resolve attachment ambiguity
- **Parser:** RASP (tag sequence-based parser)
- Read VPCs off RASP output directly
- Valence determination directly from RASP output (presence of `dobj` for head verb)

Method-4: Example (Full Parse)

```
(|He:1_PPHS1| |be+ed:2_VBDZ| |wound+ed:3_VVN| |,:4_,|
|but:5_CCB| |fight+ed:6_VVD| |on:7_RP|) 1 ; (-7.655)

  (|ncsubj| |fight+ed:6_VVD| |He:1_PPHS1| _)
  (|ncsubj| |wound+ed:3_VVN| |He:1_PPHS1| |obj|)
  (|aux| _ |wound+ed:3_VVN| |be+ed:2_VBDZ|)
  (|ncmod| _ |fight+ed:6_VVD| |on:7_RP|)
  (|conj| _ |wound+ed:3_VVN| |fight+ed:6_VVD|)
```

Method-4: Example (Partial Parse)

```
... |to:15_TO| |bring:16_VV0| |out:17_RP|
|the:18_AT| |sheen:19_NN1|) 0 ; ()

(|ncsubj| |bring:16_VV0| |child+s:12_NN2| _)
(|dobj| |bring:16_VV0| |sheen:19_NN1| _)
(|ncmod| _ |bring:16_VV0| |out:17_RP|)
(|detmod| _ |sheen:19_NN1| |the:18_AT|)
(|xcomp| |to:15_TO| |hair:14_NN1| |bring:16_VV0|)
```

Method-4: Feature Representation

- Features describing frequency of each (intrans/trans) VPC type, for full and partial parses:

$$\texttt{INTRANS}_{\mathit{full}}\ \texttt{INTRANS}_{\mathit{partial}}\ \texttt{TRANS}_{\mathit{full}}\ \texttt{TRANS}_{\mathit{partial}}$$

Method-4: Results

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
Corpus	Shallow	.975	.715	.825
	Deep$_{intrans}$	.632	.656	.644
	Deep$_{trans}$	.861	.705	.775
Dict	Shallow	.975	.715	.825
	Deep$_{intrans}$	.643	.639	.641
	Deep$_{trans}$	.865	.705	.777

Method-4: Results

- Recall down as compared to Method-3
- Results superior to RASP tagger (parser pre-processor) but below those of the other taggers
- Very little difference between corpus- and dictionary-based training data

Method Combination

- Combine methods to consolidate on relative strengths (precision/recall)
- Combination by concatenating feature vectors for individual methods

Full Combination Results

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
Corpus	Shallow	.978	.957	.967
	Deep$_{intrans}$	.638	.830	.721
	Deep$_{trans}$	.893	.890	.891
Dict	Shallow	.979	.857	.914
	Deep$_{intrans}$	.660	.781	.715
	Deep$_{trans}$	.845	.796	.820

Results of Method Combination

- System combination produces F-score superior to individual methods

- Still disappointing results for intransitive VPCs

  $\longrightarrow$ try selective system combination (chunker, chunk grammar, RASP)

Selective Combination Results

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
Corpus	Shallow	.978	.942	.960
	Deep$_{intrans}$	.651	.893	.753
	Deep$_{trans}$	.899	.874	.886
Dict	Shallow	.979	.824	.895
	Deep$_{intrans}$	.651	.717	.683
	Deep$_{trans}$	.863	.754	.805

- Best F-score for intransitive VPCs (corpus-based training data)

Full Combination Results (Brown)

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
	Shallow	.956	.888	.921
Corpus	Deep$_{intrans}$	.783	.692	.735
	Deep$_{trans}$	.840	.766	.801
	Shallow	.973	.555	.706
Dict	Deep$_{intrans}$	.675	.314	.429
	Deep$_{trans}$	.765	.571	.654

Full Combination Results (WSJ)

Training data	*Task*	*Precision*	*Recall*	$F_{\beta=1}$
Corpus	Shallow	.938	.916	.927
	Deep$_{intrans}$	.664	.669	.667
	Deep$_{trans}$	.875	.812	.843
Dict	Shallow	.983	.559	.713
	Deep$_{intrans}$	.561	.274	.368
	Deep$_{trans}$	.830	.612	.704

Reflections

- A range of methods proposed for shallow/deep lexical acquisition of VPCs from unannotated corpora

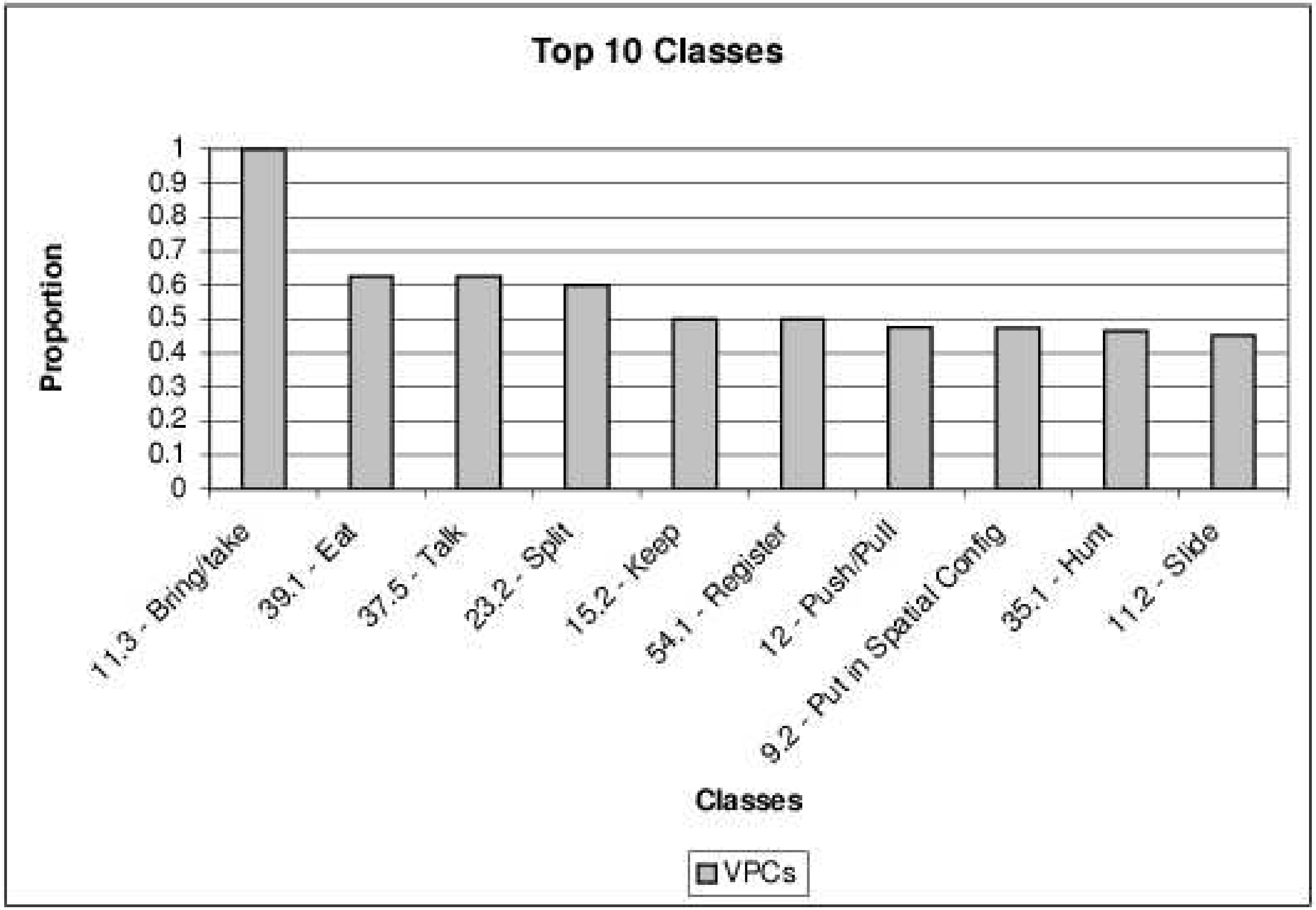
- Blurring of the token/type distinction used to boost performance

- Method robust over extremely low-frequency data (vital for many MWE types)

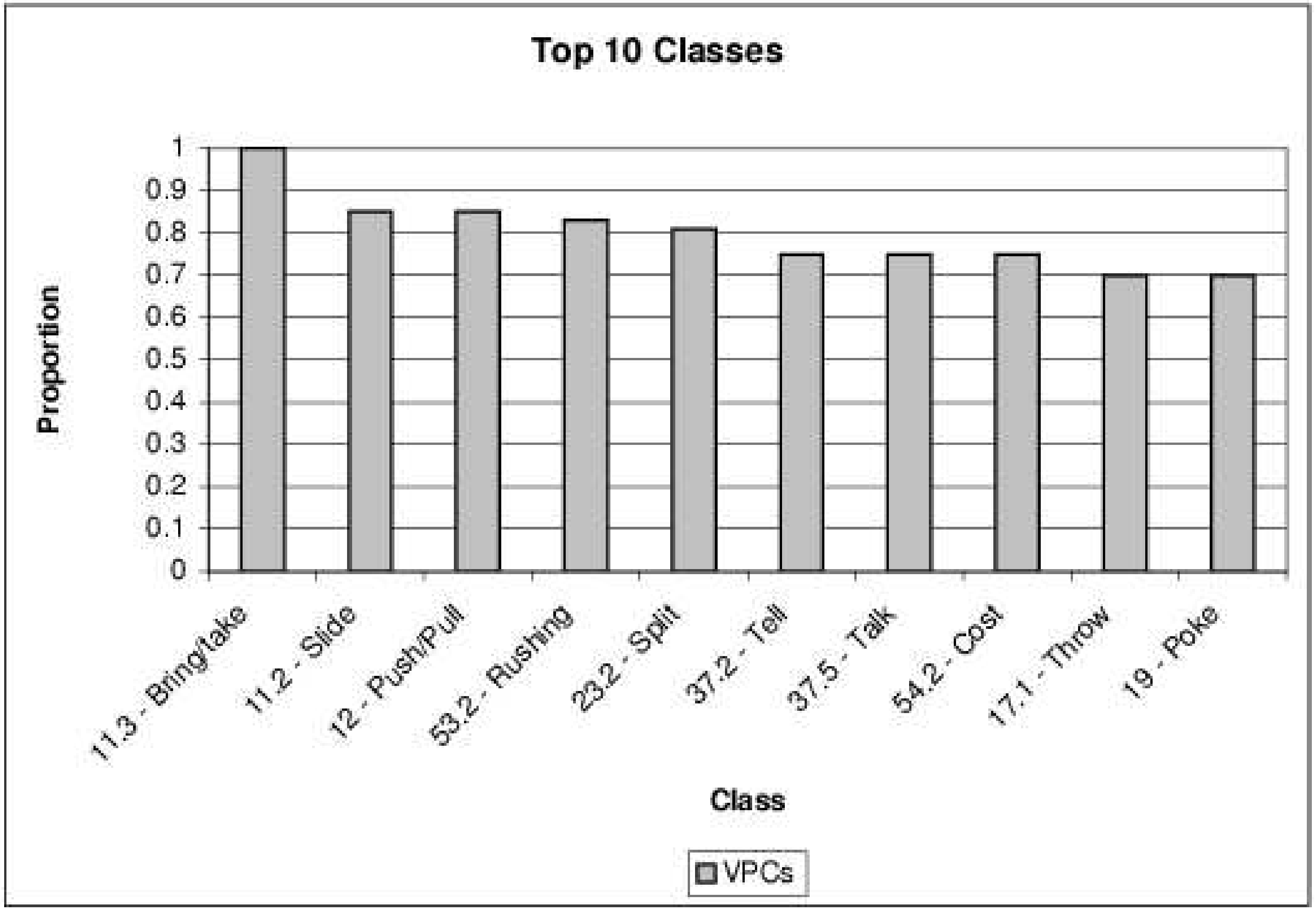
Predicting VPC Productivity

- Verb semantics are often a good predictor of verb–particle combinatorics (for compositional VPCs)
- Try using Levin classes to predict productive verb–particle combinations (e.g. aspectual *up*)

Levin-based Productivity (vs. Dictionaries)



Levin-based Productivity (vs. Dict + BNC)



Summary

- What is the basis of collocation extraction methods?
- How do collocation and MWE extraction methods differ?
- What properties of MWEs make collocation extraction techniques unsuitable?
- In what way can MWE extraction circumvent the issue of lexical coverage?

MWE INTERPRETATION: COMPOUND NOUNS

Compound Nominals and Nominalisations

- **Compound nominal:** $\bar{\text{N}}$ made up of two or more nouns, e.g.:

  telephone box/booth, river bed, radar footprint, chest X-ray

- **Nominalisation:** subclass of compound nominals in which the head noun is deverbal, e.g.:

  machine performance, museum construction, family worker, student education, satellite observation

Compound Nominals and NLP

- Compound nominals generally processed in three steps:

  a. **Identification** of compound nominals in some corpus

     A film interpretation of the book which satirises black assimilation into white society.

  b. **Syntactic analysis** of the structure

     engine oil filter ➔ *[[engine oil] filter]*

  c. **Interpretation** of the semantics

     film interpretation ➔ OBJ

- We will focus on interpretation (Step 3)

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Identification

- Compound nominals easily detectable from the output of a tagger:

  A_DT film_NN interpretation_NN of_IN the_DT book_NN which_WDT satirises_VBZ black_NN assimilation_NN into_IN white_NN society_NN ._.

Identification

- Compound nominals easily detectable from the output of a tagger:

  A_DT film_NN interpretation_NN of_IN the_DT book_NN which_WDT satirises_VBZ black_NN assimilation_NN into_IN white_NN society_NN ._.

Identification

- Largely a question of POS tagger post-correction (Lapata and Lascarides 2003)
- Subtle questions about how to detect compound nominals which are part of larger lexical items (e.g. *social services committee*)

NN Corpus Occurrence

- Estimate of English and Japanese NN corpus occurrence:

	BNC	*Reuters*	*Mainichi*
Token coverage	2.6%	3.9%	2.9%
Total no. types	265K	166K	889K
Ave. token freq.	4.2	12.7	11.1
Singletons	60.3%	44.9%	45.9%

- Highly productive, high frequency of occurrence

Syntactic Analysis

- **Adjacency** (Resnik 1993) vs. **dependency** (Lauer 1995a) in syntactic analysis, e.g. *woman aid worker*:

$$woman\ aid > aid\ worker \qquad \text{[Adj]}$$
$$aid\ worker > woman\ worker \qquad \text{[Dep]}$$
$$\rightarrow [[woman\ aid]\ worker]$$

$$woman\ aid < aid\ worker \qquad \text{[Adj]}$$
$$aid\ worker < woman\ worker \qquad \text{[Dep]}$$
$$\rightarrow [woman\ [aid\ worker]]$$

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Interpretation

- Compound nominals are largely unrestricted semantically

  diesel truck/oil/tanker, phone book, cloud bus, apple juice seat

- Nominalisations tend to occur with subject or object interpretation:

  machine performance, museum construction, student education BUT ALSO *soccer competition*

Why Compound Nominal Interpretation?

- Component of language understanding
- (Partial) interpretation required for MT into certain languages:

  cf. Italian: *coltello da pane* "bread knife", *porta a vetri* "glass door", *succo di limone* "lemon juice"

- Interface between semantics and stress assignment:

  pint jar/mile run/six-figure salary/...
  pantry shelf/garage door/bedroom furniture/...

  wood box/water bucket/gin bottle/...
  daisy chain/cable network/sugar cube/...

  BUT *rubber boots/steel plate/gold medal/...*

Interpreting Compound Nominals

- Possible to interpret compound by way of:
  - ⋆ system of "semantic relations"

    $\textsc{activity}$, $\textsc{change}$, $\textsc{person-afflicted}$, ...

    $steel\ can = \textsc{made-of}(can, steel)$
  - ⋆ paraphrasing with prepositional "hidden variable"

    $P(n2, p, n1) \approx P(n2, p, *)P(*, p, n1)$

    $baby\ chair = chair\ \textsc{for}\ babies$

Semantic Theories

- Every linguist has her own theory, but with commonalities
- Import of syntax, semantics, discourse and knowledge representation in different theories
- Claims that finite enumeration of semantic relations are psychologically untenable (Downing 1977)

Example Theory 1: Levi (1978)

- 4 roles for nominalisations:
  - ⋆ ACT, PRODUCT, AGENT, PATIENT

    truck driver = AGENT

    student discontinuation = ACT

- 9 **recoverably deletable predicates** for compound nominals:
  - ⋆ IN, FOR, FROM, ABOUT (prepositional)
  - ⋆ CAUSE, MAKE, HAVE, USE, BE (relative clauses)

    power station = MAKE

    steel box = USE

    baby crocodile = BE

Example Theory 2: Lauer (1995a)

- Interpret compound nominals according to 7 prepositions:
  - ⋆ ***of***: *state law* = *law* OF *state*
  - ⋆ ***for***: *baby chair* = *chair* FOR *baby*
  - ⋆ ***in***: *morning prayer* = *prayer* IN *morning*
  - ⋆ ***at***: *airport food* = *food* AT *airport*
  - ⋆ ***on***: *Sunday television* = *television* ON *Sunday*
  - ⋆ ***from***: *reactor waste* = *waste* FROM *reactor*
  - ⋆ ***with***: *gun man* = *man* WITH *gun*
  - ⋆ ***about***: *war story* = *story* ABOUT *war*

Example Theory 3: Copestake (2003)

- Cateogrise compounds as first category that "fits":

  a. **listed compounds**: *home secretary*
  b. **hypernmic compounds**: *tuna fish, oak tree*
  c. **deverbal compounds**: *satellite observation*
  d. **relational compounds**: *jazz fan*
  e. **made-of compounds**: *steel sword, polystyrene box*
  f. **prepositional compounds**: *airshow accident*
  g. **non-deverbal verb compounds**: *oil town*
  h. **non-paraphrasable compounds**: *listeria society*

Exercise: Analyse the NN

NN	***Interpretation***		
	Levi	***Lauer***	***Copestake***
machine translation			
cold virus			
cardboard box			
state premier			
tax module			
disk cylinder			
relaxation class			
darts competition			
tennis coach			
city protest			
telephone number			

Open Questions

- Is there a definitive categorical system of compound nominal interpretation types? (splitters and lumpers)
- Can any one system work for all domains and compound nominal types?
- What systems of interpretation work in different domains?
- To what degree is interpretation required?

The Disambiguation of Nominalisations (Lapata 2002)

Basic Outline

- **Task:** binary classification of nominalisations as having a SUBJ or OBJ interpretation (ignore nominalisations such as *soccer competition* — i.e. constrain the space in such a way that interpretation is a well-defined task)

- **Assumption:** $P(rel|n_1, n_2) \approx P(rel|v_{n_2}, n_1)$

- **Problem:** getting accurate estimates of $P(rel|v_{n_2}, n_1)$

Basic Model

$$RA(rel, n_1, n_2) = \log_2 \frac{P(\textsc{obj}|n_1, n_2)}{P(\textsc{subj}|n_1, n_2)}$$

$$P(rel|n_1, n_2) \approx \frac{f(v_{n_2}, rel, n_1)}{\sum_i f(v_{n_2}, rel_i, n_1)}$$

Resources

- Derive frequency estimates from the BNC
- Estimate $f(v_{n_2}, rel, n_1)$ from output of dependency parser (Cass)
- Determine base verb form of nominalisation based on NOMLEX and CELEX
- Hand-annotate/filter 796 nominalisations extracted from BNC

Observation

- Of 796 items in gold-standard nominalisation set, 47% not attested in BNC in either a verb-object or verb-subject relation
- How to get accurate estimates of $f(v_{n_2}, rel, n_1)$?
- **Answer:** smoothing based on the frequencies of observed verb-argument pairs

Smoothing

a. **Discounting:** redistribute probability from observed events to unobserved events

b. **Class-based smoothing:** word-to-class distributional similarity

c. **Distance-weighted averaging:** word-to-word distributional similarity

Discounting

- Katz's backing-off:

$$P(rel|n_1, n_2) = \begin{cases} \alpha\frac{f(v_{n_2}, rel, n_1)}{\sum_i f(v_{n_2}, rel_i, n_1)} & \textbf{if } f(v_{n_2}, rel, n_1) > 0 \\ \beta\frac{f(rel, n_1)}{f(n_1)} & \textbf{if } f(rel, n_1) > 0 \\ (1 - \alpha - \beta)\frac{f(rel)}{\sum_i f(rel_i)} & \textbf{otherwise} \end{cases}$$

- Estimate α and β by Good-Turing estimation

Class-based Smoothing

- Map observed verb-argument tuples onto the WordNet/Roget classes of the noun, distributing equally across all synsets the noun is categorised as belonging to

- Calculate $f(v_{n_2}, rel, n_1)$ by averaging across the classes that n_1 occurs in

- Closed world assumption for nouns

Distance-weighted Averaging

- Use **confusion probability** or **Jensen-Shannon divergence** to estimate the distributional similarity between v_{n_2} and each verb w'_1, and estimate $f(v_{n_2}, rel, n_1)$ according to:

$$f_s(v_{n_2}, rel, n_1) = \sum_{w'_1} \mathrm{sim}(v_{n_2}, w'_1) f(w'_1, rel, n_1)$$

Confusion Probability

$$
\begin{aligned}
P_C(w_1|w_1') &= \sum_{rel,w_2} P(w_1|rel,w_2)P(rel,w_2|w_1') \\
&= \sum_{rel,w_2} \frac{f(w_1,rel,w_2)}{f(rel,w_2)} \frac{f(w_1',rel,w_2)}{f(w_1')}
\end{aligned}
$$

Jensen-Shannon Divergence

$$
\begin{aligned}
J(w_1, w_1') &= \frac{1}{2}\left[D\Big(m(w_1)||n(w_1, w_1')\Big) + D\Big(m(w_2)||n(w_1, w_1')\Big)\right] \\
W_J(w_1, w_1') &= 10^{-\beta J(w_1, w_1')}
\end{aligned}
$$

where

$$m(w) = P(rel, w_2|w)$$

$$n(w_1, w_1') = \frac{1}{2}\Big(m(w_1) + m(w_1')\Big)$$

$$D\Big(m(w_1)||n(w_1, w_1')\Big) =$$

$$\sum_{rel, w_2} P(rel, w_2|w_1) \log \frac{P(rel, w_2|w_1)}{\frac{1}{2}\Big(P(rel, w_2|w_1) + P(rel, w_2|w_1')\Big)}$$

Evaluation

- Annotator agreement $=$ 89.7%
- Take 2,000 nearest neighbour verbs w_1' distance-weighted averaging methods, $\beta = 5$
- Baseline accuracy of 61.5% (OBJ interpretation)

Results

- Confusion probability and WordNet-based smoothing tend to do the best overall
- Good results for system classification, combined with context modelling in the form of the right word context of the compound nominal (85% test accuracy)

Reflections

- Interesting task-oriented smoothing experiment
- What to do with non-SUBJ/OBJ nominalisations?
- What to do with prepositional verbs, verb particles?
- Influence of pragmatics on interpretation

Classifying the Semantic Relations in Noun Compounds (Rosario and Hearst 2001)

Basic Outline

- **Task:** interpretation of (2-word) compound nominals within the biomedical domain
- **Method:** use lexical or conceptual knowledge about the component nouns to interpret the whole (context-independent)
- **Resource:** MeSH (biomedical thesaurus)

Semantic Roles

- Compound nominals interpreted via 18 (out of 38) relations:
  - ⋆ more specific than case roles, and less specific than IE template fillers
  - ⋆ customised to the biomedical domain (e.g. *polio survivors* ➔ PERSON-AFFLICTED)
  - ⋆ thresholded for frequency
  - ⋆ overlapping (multiclass classification possible: *cell growth* ➔ ACTIVITY + CHANGE)

Method

- **Class-based model:** describe NN according to the concatenation of the MeSH representations of N_1 and N_2 (up to level N)

- **Lexical model:** describe NN by its component words *(closed-word assumption)*

- **Learner:** neural network (feed-forward network with one hidden layer)

Results

- Over closed data, the lexical and class-based models perform equivalently ($\approx$ 60%)
- Over open data, the class-based model performs better (unsurprisingly)
- Suggestion that N_2 has a stronger impact on the interpretation than N_1

Reflections

- Question of interpretation system sidestepped to some degree by picking a technical domain
- Multiclassification awkward effect, which raises questions about the appropriateness of the interpretation system
- Possibility for a hybrid approach combining the class-based and lexical models?
- No systematic treatment of lexicalised nominals

Integrating Symbolic and Statistical Representations (Copestake and Lascarides 1997)

Basic Outline

- **Basic method:**

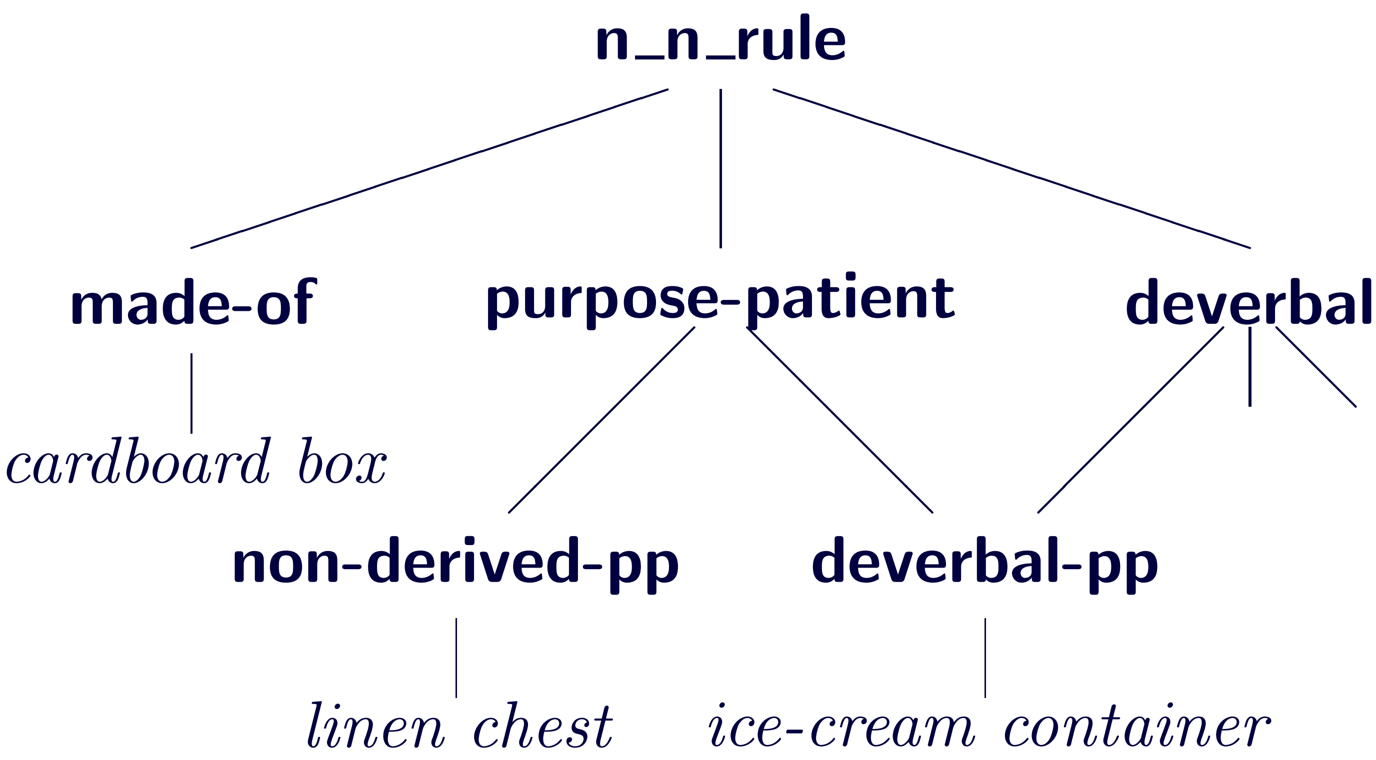
  a. use the grammar/lexicon to delimit the range of potential interpretations of a given NN

  b. use "productivity" probabilities to rank the individual interpretations

  c. use pragmatics to filter out interpretations which produce discourse incoherence within a given context

- Possible to derive non-standard interpretations for a compound nominal (e.g. *garbage man*)

Semantic Hierarchy



Estimating Productivity

- Estimate productivity based on the number of attested forms of a given schemata:

$$\mathrm{Prod}(cmp\_schema) = \frac{M+1}{N}$$

  where N is the number of pairs of senses which match cmp_schema and M is the number of attested forms

- Cf. substitution tests for collocations/compositionality

Applying the Productivity Estimates

- Interpretations for *cotton bag* based on analysis of fabric/container NNs in the BNC (based on WordNet):

MADE-OF	$P = 0.84$
PURPOSE-PATIENT	$P = 0.14$
GENERAL-NN	$P = 0.02$

- Prediction that the default interpretation for *cotton bag* is MADE-OF

Interface with Pragmatics

- Model pragmatics with SDRT and world knowledge with DICE
- Use SDRT and DICE to filter out interpretations that produce discourse incoherence:
  a. *Mary sorted her clothes into various bags made from plastic*
  b. *She put her skirt into the cotton bag*

Reflections

- Rare instance of method which provides direct handling of the lexicon-pragmatics interface
- Implausible interpretations supported explicitly, but dispreferred
- Difficulties in collecting productivity statistics
- Question of real-world applicability of SDRT/pragmatic reasoning

Summary

- What basic types of compound noun are there, and how do they differ?
- What types of theory are there for interpreting compound nouns?
- What are their strengths and weaknesses?
- What computational techniques can be employed to interpret compound nouns?

SEMANTIC COMPOSITIONALITY

Semantic Decomposability and Compositionality

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

  kick the bucket $\longrightarrow$ `die'`

  spill the beans $\longrightarrow$ `reveal'(secret')`

- Compositionality = *degree to which the semantics of the parts of an MWE contribute towards those of the whole*

- Domain considerations: *monosodium glutamate* in chemistry vs. health domains

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Syntactic Compositionality

- *Degree to which the syntactic properties of the parts of an MWE combine to make up the syntax of the whole*
  - ⋆ Fixed expressions: *by and large*, *San Francisco*
  - ⋆ Verb particles: *eat up* vs. *chicken out*
- Syntactic compositionality binary effect; non-compositional MWEs lexicalised
- Semantic decomposability continuum of regularity with more subtle effects and syntactic corollaries

Decomposability and Syntactic Flexibility

- 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

Ideal Research Objective

- Automatically decompose a given MWE/align component words with semantic primitives

- Classification of MWEs into 3 classes:

  a. **non-decomposable MWEs** (e.g. *kick the bucket*, *shoot the breeze*, *hot dog*)

  b. **idiosyncratically decomposable MWEs** (e.g. *spill the beans*, *let the cat out of the bag*, *radar footprint*)

  c. **simple decomposable MWEs** (e.g. *kindle excitement*, *traffic light*)

Realistic Short-term Objective

- Demarcate simple decomposable MWEs from idiosyncratically decomposable and non-decomposable MWEs (roughly equivalent to **endocentric** vs. **exocentric** distinction)

- Binary distinction vs. mapping onto continuum of relative decomposability

Approaches to Evaluation

- **Dictionary based**: binary evaluation, based on prediction that non-compositional MWEs will be lexically listed

- **Similarity based**: relative similarity of the parts to the whole (e.g. relative to WordNet)

$$\mathrm{sim}(\textit{pig metal}\ ,\textit{metal}) \gg \mathrm{sim}(\textit{pig metal}\ ,\textit{pig})$$

- **Entailment based**: binary evaluation, based on whether the whole "entails" the parts or not

  Susan finished up her paper $\models$ *Susan finished her paper*

- **Ranking based**: describe MWE compositionality by way of continuous/discrete scale of compositionality

  $\mathrm{comp}(\mathit{put\ up}) \geq \mathrm{comp}(\mathit{eat\ up}) \geq \mathrm{comp}(\mathit{gun\ down})$

  ...

Exercise: Rate the Compositionality

VPC	***Compositionality***				
	Dic	***Sim***	***Ent(V)***	***Ent(P)***	***Rank***
$get\ down_{trans}$					
$piss\ off_{trans}$					
$pay\ off_{trans}$					
$lift\ out_{trans}$					
$roll\ back_{trans}$					
$dig\ up_{trans}$					
$lie\ down_{intrans}$					
$wear\ on_{intrans}$					
$chicken\ out_{intrans}$					
$hand\ out_{trans}$					

Automatic Identification of Non-Compositional Phrases (Lin 1999)

Basic Method

- Use substitution as a test of compositionality:

  $red\ tape \longrightarrow \underline{yellow}\ tape,\ red\ \underline{cassette}$
  $economic\ impact \longrightarrow \underline{political}\ impact,\ economic\ \underline{effect}$

- Evaluate based on a dictionary of idioms

System Resources

- POS-conditioned thesaurus (nouns, verbs, adjectives/adverbs)
  - ⋆ derived from dependency data (Minipar):
- Collocation data
  - ⋆ dependency tuples (H,R,M) with high log-likelihood ratio (H = head, R = relation, M = modifier)

(Point-wise) Mutual Information

- Measure of the level of association between two events A and B:

$$MI(A,B) \;=\; \log_2 \frac{P(A,B)}{P(A)P(B)}$$

- Commonly used in collocation extraction

- Not appropriate for low-frequency events

Mutual Information and Compositionality

- Scaling up to 3 events A, B and C, where B and C are conditionally independent given A:

$$
\begin{aligned}
MI(A,B,C) &= \log_2 \frac{P(A,B,C)}{P(B|A)P(C|A)P(A)} \\
MI(H,R,M) &= \log_2 \frac{\frac{|\mathtt{H\ R\ M}|}{|\mathtt{*\ *\ *}|}}{\frac{|\mathtt{H\ R\ *}|}{|\mathtt{*\ R\ *}|}\frac{|\mathtt{*\ R\ M}|}{|\mathtt{*\ R\ *}|}\frac{|\mathtt{*\ R\ *}|}{|\mathtt{*\ *\ *}|}} \\
&= \log_2 \frac{|\mathtt{H\ R\ M}||\mathtt{*\ R\ *}|}{|\mathtt{H\ R\ *}||\mathtt{*\ R\ M}|}
\end{aligned}
$$

Definition of Compositionality

- A phrase α is non-compositional iff there is no β s.t.:

(a) β can be produced by substitution of the components of α for any of 10 most-similar words, and
(b) there is an overlap between the 95% confidence interval of the MI values of α and β

- 10 most-similar words tested for each of `H` and `M` (`R` fixed)

Word Similarity: Lin (1998)

$$sim(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} MI(w_1, r, w) + MI(w_2, r, w)}{\sum_{(r,w)\in T(w_1)} MI(w_1, r, w) + \sum_{(r,w)\in T(w_2)} MI(w_2, r, w)}$$

Example 1: *spill (one's) guts*

- ($spill$,V:comp1:N,gut):
  - ⋆ $spill$: $leak$, $pour$, $spew$, ..., $spray$
  - ⋆ gut: $intestine$, $instinct$, $foresight$, ..., $charisma$

- Check for each of ($leak$,V:comp1:N,gut), ($spill$,V:comp1:N,$inte$ ... in the collocation database

- None found, so *spill (one's) guts* is non-compositional

Example 2: *red tape*

- ($tape$,N:adj:N,red):
  - ⋆ $tape$: *videotape, cassette, videocassette, ..., audio*
  - ⋆ red: *yellow, purple, pink, ..., shade*

- Find ($tape$,N:adj:N,$yellow$), ($tape$,N:adj:N,$orange$), ($tape$,N:adj:N,$black$) in the collocation database but with very different MI values

- *red tape* is non-compositional

MI Confidence Interval: the Z-test

- Possible to calculate the "true" MI of (H,R,M) according to the Z-test:

$$\overline{p} \pm z_N \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = \frac{k}{n} \pm z_N \sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} \approx \frac{k \pm z_N \sqrt{k}}{n}$$

where $\overline{p}$ is the MLE of p, n is `|* * *|`, k is `|H R M|`, and z_N is a constant determined by the confidence level N, e.g. $z_{0.95} = 1.96$

Applying the Z-test

- Determine the "fit" between two MI values by calculating the Z-score interval for the putative non-compositional MWE and determining whether the MI of the second falls into that interval

Evaluation

- Evaluate the method relative to an idiom dictionary
- OK precision, and significant numbers of the extracted MWEs not contained in the dictionary appear to be non-compositional based on manual inspection

Reflections

- Is substitution really a good test for non-compositionality?
  - ⋆ institutionalised phrases: *frying pan*, *salt and pepper*, *many thanks*
  - ⋆ productive MWEs: *call/phone/ring up*
- Look to alternative methods

A Statistical Approach to the Semantics of Verb-particles (Bannard *et al.* 2003)

www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf

Basic Method

- Define similarity in terms of distributional similarity, i.e. assume that if an MWE is compositional, it will occur in the same lexical context as its parts
- Divide up compositionality to look at verb and particle similarity independently
- Evaluate against human judgements

Verb-particle constructions (VPCs)

- VPC = A verb plus one or more obligatory (prepositional) particles

  Peter put the picture up
  Susan finished up her paper
  Philip gunned down the intruder
  Barbara and Simon made out

- **Assumption:** VPCs are not always fully compositional or fully non-compositional, but rather populate a continuum between the two extremes

Compositionality by Entailment

- *Peter put the picture up*
  $\models$ *Peter put the picture somewhere*
  $\models$ *the picture was up*
- *Susan finished up her paper*
  $\models$ *Susan finished her paper*
- *Philip gunned down the intruder*
  $\models$ *the intruder was down*
- *Barbara and Simon made out*

Obtaining Human Judgements

a. Extract VPCs from British National Corpus (Baldwin and Villavicencio 2002)

b. Randomly select 5 sentence tokens for each of 40 randomly selected VPC types

c. Present native English speakers with tokens and asked whether verb and/or particle is implied by the VPC

d. Response: **Yes**, **No** or **Don't Know**

Example Sentence Tokens: *round up*

A dog started to round up sheep.

In three years they had rounded up fifty captive orangs.

Owned by Jo Rutherford, Trigo rounded up the milking herd and brought it back to the milking parlour in Devon.

On 9 August, 349 Arrests were made as the miltary swooped to round up serving and former IRA activists.

Ten days later, when the agents moved in to round up their targets, El-Jorr checked out and returned to Cyprus, charging the hotel bill to his American Express card as instructed.

Human Judgements

- Does *round up* imply *round*?
- Does *round up* imply *up*?
- Obtain gold-standard analysis by taking majority judgement (ignore **Don't Know** responses)

Sample Judgements

VPC	*Component word*	*Yes*	*No*	*Don't Know*
dig up	dig	21	5	0
	up	18	7	1
stay up	stay	20	5	1
	up	21	5	0
brighten up	brighten	9	16	1
	up	16	10	0
add up	add	12	14	0
	up	19	6	1

Binary Classification Tasks ($\times 4$)

a. Is the item completely compositional?

b. Does the item include at least one item that is compositional?

c. Does the verb contribute its simplex meaning?

d. Does the particle contribute its simplex meaning?

Classification Methods

- Four different classifiers implemented
- Method 1 based on Lin (1999), Methods 2-4 address theoretical concerns with this model
- All methods based on co-occurrence vector representation of VPC and component words

Method 1

- Reimplemented Lin (1999) over VPCs.
- Tested over all four tasks - assuming that the substitutability of each item will give us some insight into its semantic contribution
- Reconstruct Lin's thesaurus to include all of verbs, nouns, adjectives/adverbs and prepositions.

Method 2

- Similar to Lin (1999) except for use of knowledge-free approach to obtaining thesaurus
- Very similar to Schütze (1998) "context space" method
- Similarities from pairwise comparison of all verbs, particles and VPCs
- Obtain thesaurus by taking the N most similar words of a given POS

Method 3

- Use same method of substitution

- A component is said to be contributing simplex meaning if expression formed by substitution occurs among the nearest 100 verb-particle constructions

Method 4

- Hypothesis is that if a verb or particle is contributing simplex meaning to a VPC then it will be semantically similar to the VPC according to cosine measure
  - ⋆ a verb is judged to be contributing simplex meaning if it occurs within the 20 most similar items to the VPC.
  - ⋆ a particle is judged to be contributing simplex meaning if it is in top 10 most similar items to the VPC.

Results

- Mixed at best!
- Methods 3 and 4 tend to perform better than methods 1 and 2

Detecting a Continuum of Compositionality in Phrasal Verbs (McCarthy *et al.* 2003)

Basic Method

- Compare different methods for modelling compositionality based on distributional similarity and statistical tests traditionally used in collocation extraction

- Map 111 VPCs onto a ranked list, based on human judgements over an 11-point compositionality scale

- Evaluate according to the rank order correlation with the gold-standard ranked list

System Resources

- Build thesaurus a lá Lin (1998), based on dependency tuple output of RASP

Similarity-based Methods

- **overlap:** relative overlap between the top N neighbours of the VPC and its simplex verb

- **sameparticle:** the number of VPCs which select for the same particle as the given VPC amongst the top N neighbours of that VPC

- **sameparticle** $-$ **simplex:** the value for **sameparticle** minus the number of top N neighbours of the simplex verb which select for that same particle

- **simplexasneighbour:** does the simplex verb occur in the top 50 neighbours of the VPC?

- **rankofsimplex:** what is the rank of the simplex verb in the neighbours of the VPC?

- **overlapS:** the overlap of neighbours in the top N neighbours of the VPC and simplex verb, where VPC neighbours are converted to simplex verbs in the VPC case

Statistical Methods

- χ^2
- Log-likelihood ratio
- (Point-wise) mutual information
- Simple frequency of the VPC
- Simple frequency of the simplex verb

Resource-based Method

- Binary test for the occurrence of the VPC in:
  - ⋆ WordNet
  - ⋆ Alvey Tools (ANLT) VPC data
  - ⋆ Alvey Tools (ANLT) prepositional verb data

Correlation with the Gold-standard Data

- For binary tests (**simplexasneighbour**, WordNet, ANLT), use the Mann-Whitney U test (rank sum test)
- For other methods, map each output value onto a rank and apply the Spearman Rank Correlation test (rank test)
- In each case, calculate the Z score and the probability of the null hypothesis (i.e. no correlation between the output of the method and the gold-standard data)

Results

- **same particle** – **simplex** best performer of similarity-based methods
- MI best performer of statistical tests
- Question of how to apply the results to the proposed task of lexical acquisition?

Overall Reflections

- Promising results observed for detecting compositionality/decom
but less so for determining the semantic contribution of individual words in an overall MWE
- What about MWEs where the simplex words don't occur with that same POS (e.g. *chicken out*)
- Effects of polysemy (e.g. *run down*, *run over*)
- How to move on to actually semantically decompose an MWE?

Summary

- How do decomposability and compositionality differ?
- What methods have been proposed for generating gold-standard compositionality data?
-

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*
- More work needs to be done in developing gold-standard resources to encourage others to enter the fray
- Most of the research problems are far from resolved: lots of room for everyone to play in!

MWE Resources

- MWE project page: `mwe.stanford.edu`
- On-line MWE data: `mwe.stanford.edu/resources`
- On-line bibliographies:
  - ⋆ `mwe.stanford.edu/bib.html`
  - ⋆ `www.ims.uni-stuttgart.de/euralex/bibweb/`

References

Aarts, Bas. 1989a. *Small Clauses in English: the Non-verbal Types*. Berlin: Mouton de Gruyter.

——. 1989b. Verb-preposition constructions and small clauses in English. *Journal of Linguistics* 25.277–90.

Abeillé, Anne. 1988. Light verb constructions and extraction out of NP in a tree adjoining grammar. In *Papers of the 24th Regional Meeting of the Chicago Linguistics Society*.

——. 1990. Lexical and syntactic rules in a tree adjoining grammar. In *Proc. of the 28th Annual Meeting of the ACL*, 292–8.

——. 1995. The flexibility of French idioms: A representation with Lexicalised Tree Adjoining Grammar. In (Everaert *et al.* 1995a), chapter 1.

Akimoto, Minoji. 1989. *A Study of Verbo-Nominal Structures in English*. Tokyo: Shinozaki Shorin.

Aldinger, Nadine. 2004. Towards a dynamic lexicon: Predicting the syntactic argument structure of complex verbs. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Alexander, Richard J. 1978. Fixed expressions in English: A linguistic, psycholinguistic, sociolinguistic and didactic study (part 1). *Anglistik und Englischunterricht* 6.171–88.

—— 1979. Fixed expressions in English: A linguistic, psycholinguistic, sociolinguistic and didactic study (part 2). *Anglistik und Englischunterricht* 7.181–202.

Allerton, D.J. 1984. Three (or four) levels of word cooccurrence restriction. *Lingua* 63.17–40.

Ananiadou, Sophia. 1994. A methodology for automatic term recognition. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, 1034–8, Kyoto, Japan.

BADDORF, DEBRA S., and MARTHA W. EVENS. 1998. Finding phrases rather than discovering collocations: Searching corpora for dictionary phrases. In *Proc. of the 9th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS'98)*, Dayton, USA.

BALDWIN, TIMOTHY. (to appear). The deep lexical acquisition of english verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions* .

——, COLIN BANNARD, TAKAAKI TANAKA, and DOMINIC WIDDOWS. 2003a. An empirical model of multiword expression decomposability. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 89–96, Sapporo, Japan.

——, JOHN BEAVERS, LEONOOR VAN DER BEEK, FRANCIS BOND, DAN FLICKINGER, and IVAN A. SAG. 2003b. In search of a systematic treatment of determinerless PPs. In *Proc. of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France.

——, EMILY M. BENDER, DAN FLICKINGER, ARA KIM, and STEPHAN OEPEN. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

——, and FRANCIS BOND. 2002. Multiword expressions: Some problems for Japanese NLP. In *Proc. of the 8th Annual Meeting of the Association for Natural Language Processing (Japan)*, 379–82, Keihanna, Japan.

——, and ALINE VILLAVICENCIO. 2002. Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, 98–104, Taipei, Taiwan.

BAME, KEN, 1999. Aspectual and resultative verb-particle constructions with up. Handout for talk presented at the Ohio State University Linguistics Graduate Student Colloquium.

BANNARD, COLIN, 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. *LinGO Working Paper No. 2002-06*.

——, TIMOTHY BALDWIN, and ALEX LASCARIDES. 2003. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 65–72, Sapporo, Japan.

BARKEMA, HENK. 1996a. The effect of inherent and contextual features on the grammatical flexibility of idioms. In *Synchronic Corpus Linguistics*, 69–83. Amsterdam, Netherlands: Rodopi.

——. 1996b. Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica* 50.125–60.

BARKER, KEN. 1998. A trainable bracketer for noun modifiers. In *Proc. of the 12th Canadian Conference on Artificial Intelligence*, 196–210, Vancouver, Canada.

——, and STAN SZPAKOWICZ. 1998. Semi-automatic recognition of noun modifier relationships. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, 96–102, Montreal, Canada.

BASILI, ROBERTO, MARIA TERESA PAZIENZA, and PAOLA VELARDI. 1993. Semi-automatic extraction of linguistic information for syntactic disambiguation. *Applied Artificial Intelligence* 7.339–64.

BAUER, LAURIE. 1979. On the need for pragmatics in the study of nominal compounding. *Journal of Pragmatics* 3.45–50.

——. 1983. *English Word-formation*. Cambridge, UK: Cambridge University Press.

BENNIS, HANS J., MARCEL DEN DIKKEN, PETER JORDENS, SUSAN POWERS, and JURGEN WEISSENBORN. 1995. Picking up particles. In *Proc. of the 19th Annual Boston University Conference on Language Development (BUCLD 19)*, 70–82, Somerville, USA. Cascadilla Press.

BLAHETA, DON, and MARK JOHNSON. 2001. Unsupervised learning of multi-word verbs. In *Proc. of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, 54–60, Toulouse, France.

BOLINGER, DWIGHT. 1967. Adjectives in English: Attribution and predication. *Lingua* 18.1–34.

——. 1971. *The Phrasal Verb in English*. Harvard, USA: Harvard University Press.

—— (ed.) 1972. *Degree Words*. the Hague, Netherlands: Mouton.

BOND, FRANCIS, and SATOSHI SHIRAI, 1997. *Practical and Efficient Organization of a Large Valency Dictionary*. Handout at the *Workshop on Multilingual Information Processing, held in conjunction with NLPRS'97*.

BORTHEN, KAJA, 2003. *Norwegian Bare Singulars*. Norwegian University of Science and Technology dissertation.

BOUCHER, PAUL, FRÉDÉRIC DANNA, and PASCALE SÉBILLOT. 1993. Compounds: An intelligent tutoring system for learning to use compounds in English. *Computer Assisted Language Learning (CALL)* 6.249–72.

BOUILLON, PIERRETTE. 1996. Mental state adjectives: the perspective of generative lexicon. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, 143–8, Copenhagen, Denmark.

——, KATHARINA BOESEFELDT, and GRAHAM RUSSELL. 1994. Compound nouns in a unification-based MT system. In *Proc. of the 4th Conference on Applied Natural Language Processing (ANLP)*, Stuttgart, Germany.

——, and EVELYN VIEGAS. 1999. The description of adjectives for natural language processing: Theoretical and applied perspectives. In *Proc. of the TALN'99 workshop on Description des adjectifs pour les traitements informatiques*, Cargèse, France.

BOURIGAULT, DIDIER. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of the 14th International Conference on Computational Linguistics (COLING '92)*, 977–81, Nantes, Frances.

——. 1993. An endogenous corpus-based method for structural noun phrase disambiguation. In *Proc. of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, Utrecht, Netherlands.

——, CHRISTIAN JACQUEMIN, and MARIE-CLAUDE LH́OMME (eds.) London, UK. *Recent advances in Computational Terminology*. Johns Benjamins.

BREIDT, ELISABETH. 1993. Extraction of V-N-collocations from text corpora: A feasibility study for German. In *Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, USA.

BRINTON, LAUREL. 1985. Verb particles in English: Aspect or aktionsart. *Studia Linguistica* 39.157–68.

BRUGMAN, CLAUDIA, 1981. The story of over. Master's thesis, UCLA, Berkeley.

——, 1988. *The Syntax and Semantics of HAVE and its Complements*. UCLA, Berkeley dissertation.

BUSA, FREDERICA, and MICHAEL JOHNSTON. 1996. Cross-linguistic semantics for complex nominals in the generative lexicon. In *AISB Workshop on Multilinguality in the Lexicon*, Sussex, UK.

BUTT, MIRIAM. 1995. *The structure of complex predicates in Urdu*. Stanford, USA: CSLI Publications.

CACCIARI, CRISTINA, and PATRIZIA TABOSSI (eds.) 1993a. *Idioms: Processing, Structure and Interpretation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

——, and ——. 1993b. Preface. In (Cacciari and Tabossi 1993a).

CALZOLARI, NICOLETTA, CHARLES FILLMORE, RALPH GRISHMAN, NANCY IDE, ALESSANDRO LENCI, CATHERINE MACLEOD, and ANTONIO ZAMPOLLI. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 1934–40, Las Palmas, Canary Islands.

CAO, YUNBO, and HANG LI. 2002. Base noun phrase translation using Web data and the EM algorithm. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

CHEN, A., K. KISHIDA, H. JIANG, and Q. LIANG. 1999. Automatic construction of a Japanese–English lexicon and its application in cross-language information retrieval. In *ACM DL/ACM SIGIR Workshop on Multilingual Information Discovery and Access (MIDAS)*.

CHEN, K-H., and H-H. CHEN. 1994. Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. In *Proc. of the 32nd Annual Meeting of the ACL*, 234–41.

CHEN, P. 1986. Discourse and particle movement in English. *Studies in Language* 10.79–95.

CHOUEKA, Y., S.T. KLEIN, and E. NEUWITZ. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing* 4.34–8.

COLOMBO, L. 1993. The comprehension of ambiguous idioms in context. In (Cacciari and Tabossi 1993a), chapter 8.

COPESTAKE, ANN. 2001. The semi-generative lexicon: limits on lexical productivity. In *Proceedings of the 1st International Workshop on Generative Approaches to the Lexicon. Geneva*.

——. 2002. *Implementing Typed Feature Structure Grammars*. Stanford, USA: CSLI Publications.

——. 2003. Compounds revisited. In *Proc. of the 2nd International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland.

——, and DAN FLICKINGER. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

——, FABRE LAMBEAU, ALINE VILLAVICENCIO, FRANCIS BOND, TIMOTHY BALDWIN, IVAN A. SAG, and DAN FLICKINGER. 2002. Multiword expressions: Linguistic precision and reusability. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 1941–7, Las Palmas, Canary Islands.

——, FABRE LAMBEAU, BENJAMIN WALDRON, FRANCIS BOND, DAN FLICKINGER, and STEPHAN OEPEN. 2004. A lexicon module for a grammar development environment. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1111–4, Lisbon, Portugal.

——, and ALEX LASCARIDES. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, 136–43, Madrid, Spain.

Cowie, A.P. (ed.) 1998. *Phraseology : theory, analysis, and applications*. OUP.

——, and P. Howarth. 1996. Phraseology—a select bibliography. *International Journal of Lexicography* 9.38–51.

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

Dagan, I., and K. Church. 1994. Termight: Identifying and translation technical terminology. In *Proc. of the 4th Conference on Applied Natural Language Processing (ANLP)*, 34–40, Stuttgart, Germany.

Dagut, M., and B. Laufer. 1985. Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition* 7.73–9.

Dasgupta, P. 1993. Idiomaticity and Esperanto texts: An empirical study. *Linguistics* 31.367–86.

Dehé, Nicole. 2000. On particle verbs in English: More evidence from information structure. In *Proc. of the 28th Western Conference on Linguistics (WECOL 1999)*, 92–105.

——. 2001a. Intonation patterns of particle verb constructions in English. In *Proc. of the 31st Conference of the North Eastern Linguistics Society (NELS 31)*, 183–97.

——. 2001b. Transitive particle verbs in English: The neutral order. evidence from speech production. In (Dehé and Wanner 2001), 165–89.

——. 2002. *Particle Verbs in English: Syntax, Information, Structure and Intonation*. Amsterdam, Netherlands/Philadelphia USA: John Benjamins.

——, Ray Jackendoff, Andrew McIntyre, and Silke Urban (eds.) 2002. *Verb-particle explorations*. Berlin/New York: Mouton de Gruyter.

——, and Anja Wanner (eds.) 2001. *Structural Aspects of Semantically Complex Verbs*. Peter Lang.

Den Dikken, M. 1995. *Particles: On the Syntax of Verb-particle, Triadic, and Causative Constructions*. Oxford University Press.

DIAS, G., S. GUILLORÉ, J-C. BASSANO, and J.G. PEREIRA LOPES. 2000. Combining linguistics with statistics for multiword term extraction: A fruitful association? In *Proc. of Recherche d'Informations Assistee par Ordinateur 2000 (RIAO'2000)*.

DIRVEN, RENÉ. 2001. The metaphoric in recent cognitive approaches to English phrasal verbs. *metaphorik.de* 1.39–54.

DIXON, ROBERT. 1982. The grammar of English phrasal verbs. *Australian Journal of Linguistics* 2.149–247.

DORR, BONNIE J. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation* 12.271–322.

DOWNING, PAMELA. 1977. On the creation and use of English compound nouns. *Language* 53.810–42.

ERBACH, G., and B. KRENN. 1994. Idioms and support-verb constructions in HPSG. In *German in Head-driven Phrase Structure Grammar*, ed. by J. Nerbonne, K. Netter, and C. Pollard. Stanford, USA: CSLI Lecture Notes.

EVERAERT, MARTIN, ERIK-JAN VAN DER LINDEN, ANDRÉ SCHENK, and ROB SCHREUDER (eds.) 1995a. *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates.

——, ERIK-JAN VAN DER LINDEN, ANDRÉ SCHENK, and ROB SCHREUDER. 1995b. Introduction. In (Everaert *et al.* 1995a).

EVERT, STEFAN, and BRIGITTE KRENN. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proc. of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, Toulouse, France.

FELLBAUM, CHRISTIANE. 1993. The determiner in English idioms. In (Cacciari and Tabossi 1993a), chapter 12.

——. 1998a. Towards a representation of idioms in wordnet. In *Proc. of the COLING-ACL'98 Workshop on the Usage of WordNet in Natural Language Processing Systems*, 52–7, Montreal, Canada.

—— (ed.) 1998b. *WordNet: An Electronic Lexical Database*. Cambridge, USA: MIT Press.

FERNANDO, C., and R. FLAVELL. 1978. *On Idiom: Critical Views and Perspectives*. Exeter: University of Exeter.

FERRIS, C. 1993. *The Meaning of Syntax: A Study of Adjectives of English*. London: Longman.

FILLMORE, CHARLES J., and B.T. SUE ATKINS. 1994. Starting where the dictionaries stop: The challenge of corpus lexicography. In *Computational Approaches to the Lexicon*, ed. by B.T. Sue Atkins and Antonio Zampolli, chapter 13, 349–93. Oxford, UK: Oxford University Press.

——, and PAUL KAY. to appear. *Construction Grammar*. Stanford, USA: CSLI Publications.

——, ——, and MARY C. O'CONNOR. 1988. Regularity and idiomaticity in grammatical constructions. *Language* 64.501–38.

FININ, T.W. 1980. The semantic interpretation of nominal compounds. In *Proc. of the 1st Conference on Artificial Intelligence (AAAI-80)*.

FLORES D'ARCAIS, G.B. 1993. The comprehension and semantic interpretation of idioms. In (Cacciari and Tabossi 1993a), chapter 4.

FRASER, B. 1970a. Idioms within a transformational grammar. *Foundations of Language* 6.22–42.

—— 1970b. Some remarks on the action nominalization in English. In *Readings in English Transformational Grammar*, ed. by R.A. Jacobs and P.S. Rosenbaum. Ginn and Company.

—— 1976. *The Verb-Particle Combination in English*. The Hague: Mouton.

FUNG, P., M-Y. KAN, and Y. HORITA. 1996. Extracting Japanese domain and technical terms is relatively easy. In *Proc. of the 2nd International Conference on New Methods in Natural Language Processing*, 148–59.

——, and L.Y. YEE. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics: COLING/ACL-98*, 414–20.

Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proc. of the 33rd Annual Meeting of the ACL*, 236–43, Cambridge, USA.

——. 1998. Extracting key terms from Chinese and Japanese texts. *The International Journal on Computer Processing of Oriental Language, Special Issue on Information Retrieval on Oriental Languages* 99–121.

——, and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proc. of the 5th Annual Workshop on Very Large Corpora*, 192–202, Hong Kong.

Geeraerts, Dirk. 1995. Specialization and reinterpretation in idioms. In (Everaert *et al.* 1995a), chapter 3.

Gerber, Laurie, and Jin Yang. 1997. Systran MT dictionary development. In *Proc. of the Fifth Machine Translation Summit (MT Summit V)*, San Diego, USA.

Gibbs, Raymond W., and Nandini P. Nayak. 1989. Psycholinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology* 21.100–38.

Godby, J., 1994. Two techniques for the identification of phrases in full text. http://www.oclc.org/research/publications/arr/1994/part1/twotech.htm.

Grefenstette, Gregory. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Translating and the Computer 21: ASLIB'99*, London, UK.

Gries, Stefan T. 1999. Particle movement: A cognitive and functional approach. *Cognitive Linguistics* 10.105–45.

——, 2000. *Towards multifactorial analyses of syntactic variation: The case of particle placement*. University of Hamburg dissertation.

Grishman, Ralph, Catherine Macleod, and Adam Myers, 1998. *COMLEX Syntax Reference Manual*. Proteus Project, NYU. (`http://nlp.cs.nyu.edu/comlex/refman.ps`).

Grover, Claire, John Carroll, and Edward Briscoe. 1993. The Alvey Natural Language Tools grammar (4th release). Technical Report 284, Computer Laboratory, Cambridge University, UK.

GUÉRON, J. 1986. Clause union and the verb-particle construction in English. In *Proc. of the 16th Conference of the North Eastern Linguistics Society (NELS 16)*.

HARUNO, M., S. IKEHARA, and T. YAMAZAKI. 1996. Learning bilingual collocations by word-level sorting. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark.

HASHIMOTO, CHIKARA. 2003. Japanese HPSG: Treatment of syntax and semantics of syntactically complex verbs. In *Information Processing Society of Japan SIG Notes*. (in Japanese).

HASPELMATH, MARTIN. 1997. *From Space to Time in The World's Languages*. Munich, Germany: Lincom Europa.

——. 2002. *Understanding Morphology*. Arnold Publishers.

HEKTOEN, E. 1997. Probabilistic parse selection based on semantic cooccurrences. In *5th International workshop on parsing technologies (IWPT-97)*, 113–122.

HIMENO, MASAKO. 2001. The nature of compound verbs. *Nihongogaku* 240.6–15.

HIMMELMANN, NIKOLAUS P. 1998. Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology* 2.315–353.

HOSHI, H., 1994. *Passive, Causive, and Light Verbs: A Study of Theta Role Assignment*. University of Connecticut dissertation.

HOWARTH, P.A. 1996. *Phraseology in English Academic Writing*. Tb̈ingen: Max Niemeyer.

HUDDLESTON, RODNEY, and GEOFFREY K. PULLUM. 2002. *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

HULSTIJN, J.H., and E. MARCHENA. 1989. Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition* 11.241–55.

ichiro Kamei, Shin. 1993. *Nihongo no yōgen-ku sōtō kan'yō-hyōgen no bunrui to sono ōyō [A classification of Japanese verbal idioms]*. In *Proc. of the 47th National Conference of the Information Processing Society of Japan*, volume 3, 87–8. (In Japanese).

——, Shinko Tamura, and Kazunori Muraki. 1997. Nihongo no yōgen-sōtō kan'yō-hyōgen no imi-kūkan ni-okeru bunpu-zu [A classification of Japanese verbal idioms over a semantic space]. In *Proc. of the Third Annual Meeting of the Japanese Association for Natural Language Processing*, 51–4. (In Japanese).

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Nihongo Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten.

IPA, 1987. *IPA Lexicon of the Japanese Language for Computers*. Tokyo, Japan. (In Japanese).

Isabelle, P. 1984. Another look at nominal compounds. In *Proc. of the 10th International Conference on Computational Linguistics (COLING '84)*, Stanford, USA.

Ishikawa, K. 2000. A local relation between particles and verbal prefixes in English. *English Linguistics* 17.249–75.

J. Yoon, K-S. Choi, M. Song. 2001. A corpus-based approach for Korean nominal compound analysis based on linguistic and statistical information. *Natural Language Engineering* 7.

Jackendoff, Ray. 1983. *Semantics and Cognition*. Cambridge, USA: MIT Press.

——. 1995. The boundaries of the lexicon. In (Everaert *et al.* 1995a), chapter 7.

——. 1996. The proper treatment of measuring out, telicity and perhaps event quantification in English. *Natural Language and Linguistic Theory* 14.305–54.

——. 1997a. *The Architecture of the Language Faculty*. Cambridge, USA: MIT Press.

——. 1997b. Twistin' the night away. *Language* 73.534–59.

JOHANSSON, C. 1996. Good bigrams. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, 592–7, Copenhagen, Denmark.

JOHNSON, K. 1991. Object positions. *Natural Language and Linguistic Theory* 9.577–636.

JOHNSON-LAIRD, PHILIP N. 1993. Foreword. In (Cacciari and Tabossi 1993a).

JOHNSTON, MICHAEL, and FREDERICA BUSA. 1996. Qualia structure and the compositional interpretation of compounds. In *ACL SIGLEX workshop on breadth and depth of semantic lexicons*, Santa Cruz, USA.

——, and ——. 1998. The compositional interpretation of nominal compounds. In *The Breadth and Depth of Semantics Lexicons*, ed. by Evelyn Viegas. Kluwer Academic.

JONES, DANIEL, and MELINA ALEXA. 1997. Towards automatically aligning German compounds with English word groups. In *New Methods in Language Processing*, ed. by Daniel Jones and Harold Somers. London, UK: University College Press.

JUSTESON, JOHN S., and SLAVA M. KATZ. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1.9–27.

KAALEP, HEIKI-JAAN, and KADRI MUISCHNEK. 2002. Using the text corpus to create a comprehensive list of phrasal verbs. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 101–5, Las Palmas, Canary Islands.

KAGEURA, KYO. 1997. On intra-term relations of complex terms in the description of term formation patterns. In *Melanges de Linguistique Offerts a R. Kocourek*, 105–11. Halifax: Les Presses d'ALFA.

——. 1998. A statistical analysis of morphemes in Japanese terminology. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, 638–45, Montreal, Canada.

——, Fuyuki Yoshikane, and Takayuki Nozawa. 2004. Parallel bilingual paraphrase rules for noun compounds: Concepts and rules for exploring Web language resources. In *Proc. of the Fourth Workshop on Asian Language Resources*, 54–61, Sanya, China.

Kageyama, Taro. 1993. *Grammar and Word Formation*. Tokyo, Japan: Hitsuji Shobo.

——. 1999. Word formation. In *The Handbook of Japanese Linguistics*, 297–325. Blackwell Publishers.

Kaluza, H. 1984. English verbs with prepositions and particles. *International Review of Applied Linguistics in Language Teaching (IRAL)* 12.109–13.

Kanzaki, Kyoko. 1997. Lexical semantic relations between adnominal constituents and their head nouns. *Mathematical Linguistics* 21.53–68. (In Japanese).

——, and Hitoshi Isahara. 1997. Lexical semantics for adnominal constituents in Japanese. In *Proc. of the 4th Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, 573–6, Phuket, Thailand.

——, Qing Ma, and Hitoshi Isahara. 2000. Similarities and differences among semantic behaviors of japanese adnominal constituents. In *Proc. of ANLP/NAACL2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*, 59–68, Seattle, USA.

Kayne, Richard. 1985. Principles of particle constructions. In *Grammatical Representation*, ed. by J. Guéron, H-G. Obenauer, and J-Y. Pollock, 101–40. Dordrecht, Netherlands: Foris.

Kita, K., Y. Kato, T. Omoto, and Y. Yano. 1994. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing* 1.21–33.

——, and H. Ogata. 1997. Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning: An International Journal* 10.229–38.

KOBAYASI, Y., T. TOKUNAGA, and H. TANAKA. 1994. Analysis of Japanese compound nouns using collocational information. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, 865–9, Kyoto, Japan.

LANGKILDE, IRENE, and KEVIN KNIGHT. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, 704–710, Montreal, Canada.

LAPATA, M., F. KELLER, and S. MCDONALD. 2001. Evaluating smoothing algorithms against plausibility judgements. In *Proc. of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, Toulouse, France.

——, S. MCDONALD, and F. KELLER. 1999. Determinants of adjective-noun plausibility. In *Proc. of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, 30–6.

LAPATA, MARIA. 2002. The disambiguation of nominalizations. *Computational Linguistics* 28.357–88.

LAPATA, MIRELLA, and ALEX LASCARIDES. 2003. Detecting novel compounds: The role of distributional evidence. In *Proc. of the 10th Conference of the EACL (EACL 2003)*, Budapest, Hungary.

LAUER, MARK. 1995a. Corpus statistics meet the noun compound: Some empirical results. In *Proc. of the 33rd Annual Meeting of the ACL*.

——, 1995b. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Macquarie University dissertation.

——, and MARK DRAS. 1994. A probabilistic model of compound nouns. In *Proc. of the 7th Joint Australian Conference on Artificial Intelligence*.

LAUFER, B., and S. ELIASSON. 1993. What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition* 15.35–48.

LAUFER, BATIA. 2000. Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia Linguistica* 54.186–96.

LEE, HYUN AH, and GIL CHANG KIM. 2002. Translation selection through source word sense disambiguation and target word selection. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, 530–536, Taipei, Taiwan.

LEHRBERGER, J. 1982. Automatic translation and the concept of sublanguage. In *Sublanguage: Studies of Language in Restricted Semantic Domains*, ed. by R. Kittredge and J. Lehrberger. de Gruyter.

—— 1986. Sublanguage analysis. In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, ed. by Ralph Grishman and Richard Kittredge, chapter 2, 19–38. Lawrence Erlbaum Associates.

LEVI, JUDITH N. 1978. *The Syntax and Semantics of Complex Nominals*. New York, USA: Academic Press.

LEWIS, DAVID D., and W. BRUCE CROFT. 1990. Term clustering of syntactic phrases. In *Proc. of 13th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'90)*, 385–404, Brussels, Belgium.

LI, WEI, XIUHONG ZHANG, CHENG NIU, YUANKAI JIANG, and ROHINI K. SRIHARI. 2003. An expert lexicon approach to identifying English phrasal verbs. In *Proc. of the 41st Annual Meeting of the ACL*, 513–20, Sapporo, Japan.

LIBERMAN, MARK, and RICHARD SPROAT. 1992. The stress and structure of modified noun phrases in English. In *Lexical Matters – CSLI Lecture Notes No. 24*, ed. by Ivan A. Sag and A. Szabolcsi. Stanford, USA: CSLI Publications.

LICERAS, J.M., and L. DÍAZ. 2000. Triggers in L2 acquisition: The case of Spanish N-N compounds. *Studia Linguistica* 54.197–211.

LIN, DEKANG. 1998a. Extracting collocations from text corpora. In *Proc. of the COLING-ACL'98 Workshop on Computational Terminology*, Montreal, Canada.

——. 1998b. Using collocation statistics in information extraction. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*.

——. 1999. Automatic identification of non-compositional phrases. In *Proc. of the 37th Annual Meeting of the ACL*, 317–24, College Park, USA.

Lindner, Susan. 1982. What goes up doesn't necessarily come down: The ins and outs of opposites. In *Papers of the 18th Regional Meeting of the Chicago Linguistics Society*, 305–23.

——. 1983. *A Lexico-semantic Analysis of Verb-particle Constructions with Up and Out*. Indiana, USA: Indiana University Linguistics Club.

Liontas, John I. 2002. Exploring second language learners' notions of idiomaticity. *System* 30.289–313.

Lipka, L. 1972. *Semantic Structure and Word-Formation: Verb-Particle Constructions in Contemporary English*. Wilhelm Fink Verlag.

Lohse, Barbara, John A. Hawkins, and Tom Wasow, in preparation. Processsing domains in English verb-particle constructions.

Lüdeling, A. 2001. *Particle verbs and similar constructions in German*. Stanford, USA: CSLI Publications.

Makai, A. 1972. *Idiom Structure in English*. Mouton.

Mandala, Rila, Takenobu Tokunaga, and Hozumi Tanaka. 2000. Query expansion using heterogeneous thesauri. *Information Processing and Management* 36.361–78.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, USA: MIT Press.

Marsh, E. 1984. A computational analysis of complex noun phrases in navy messages. In *Proc. of the 10th International Conference on Computational Linguistics (COLING '84)*, 505–8, Stanford, USA.

MATSUMOTO, YO. 1996. *Complex Predicates in Japanese: A Syntactic and Semantic Study of the Notion 'Word'*. Stanford, USA: CSLI & Kurosio Publishers.

——. 1998. The combinatory possibilities in Japanese V-V lexical compounds. *Gengo Kenkyu* 114.37–83.

MAYNARD, DIANA, and SOPHIA ANANIADOU. 1999. Identifying contextual information for multi-word term extraction. In *5th International Congress on Terminology and Knowledge Engineering (TKE 99)*, 212–21.

MCCARTHY, DIANA, BILL KELLER, and JOHN CARROLL. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

MCINTYRE, ANDREW, 2001. Introduction to the verb-particle experience. Ms, Leipzig.

MCKEOWN, KATHLEEN, and DRAGOMIR RADEV. 2000. Collocations. In *Handbook of Natural Language Processing*, ed. by Robert Dale, Hermann Moisl, and Harold Somers, chapter 21. Marcel Dekker.

MELAMED, I.D. 1998. Empirical methods for MT lexicon development. In *Proc. of AMTA'98: Conference of the Association for Machine Translation in the Americas*, 18–30.

—— 2000. Models of translational equivalence among words. *Computational Linguistics* 26.221–49.

MEL'ČUK, IGOR. 1995. Phrasemes in language and phraseology in linguistics. In (Everaert *et al.* 1995a), chapter 8.

MEL'ČUK, IGOR. 1998. Collocations and lexical functions. In *Phraseology: Theory, Analysis, and Applications*, 23–54. Oxford: Clarendon Press.

——, and ALAIN POLGUÈRE. 1987. A formal lexicon in meaning-text theory (or how to do lexica with words). *Computational Linguistics* 13.261–275.

MERKEL, M., and M. ANDERSSON. 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proc. of Recherche d'Informations Assistee par Ordinateur 2000 (RIAO'2000)*.

——, B. Nilsson, and L. Ahrenberg. 1994. A phrase-retrieval system based on recurrence. In *Proc. of the 2nd Annual Workshop on Very Large Corpora*, 99–108.

Meyer, Ralf. 1993. *Compound Comprehension in Isolation and in Context*. Tübingen: Max Niemeyer Verlag.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography* 3.235–44.

Mitchell, T.F. 1971. Linguistic 'goings on': Collocations and other lexical matters arising on the syntactic record. *Archivum Linguisticum* 2.35–69.

Miyagawa, Shigeru. 1989. Light verbs and the ergative hypothesis. *Linguistic Inquiry* 20.659–68.

Moon, Rosamund (ed.) 1991. *COBUILD Dictionary of Phrasal Verbs*. Collins.

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

Morgan, Pamela S. 1997. Figuring out *Figure out*: Metaphor and the semantics of the English verb-particle construction. *Cognitive Linguistics* 8.327–57.

Mortimer, C. 1972. *Phrasal verbs in conversation*. Longman.

Müller, Stefan. 2001. German particle verbs and the predicate complex. In *Grammatical Interfaces in HPSG (Studies in Constraint-Based Lexicalism)*, ed. by Ronnie Cann, Claire Grover, and Philip Miller. Stanford, USA: CSLI Publications.

Nagata, Masaaki, Teruka Saito, and Kenji Suzuki. 2001. Using the Web as a bilingual dictionary. In *Proc. of the ACL/EACL 2001 Workshop on Data-Driven Methods in Machine Translation*, 95–102, Toulouse, France.

Neeleman, Ad, 1994. *Complex Predicates*. Utrecht University dissertation.

Nicolas, Tim. 1995. Semantics of idiom modification. In (Everaert *et al.* 1995a), chapter 9.

NIIMI, KAZUAKI, YOUICHI YAMAURA, and TOKUKO UTSUNO. 1987. *Compound Verbs*. Tokyo, Japan: Aratake Shuppan. In Japanese.

NUNBERG, GEOFFREY, IVAN A. SAG, and TOM WASOW. 1994. Idioms. *Language* 70.491–538.

ODIJK, JAN. 2004. Reusable lexical representations for idioms. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 903–6, Lisbon, Portugal.

O'DOWD, ELIZABETH M. 1998. *Prepositions and Particles in English*. Oxford University Press.

OHMORI, KUMIKO, and MASANOBU HIGASHIDA. 1999. Extracting bilingual collocations from non-aligned parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, 88–97, Chester, UK.

PEABODY, K.W., 1981. Constraints on the productivity of verb-particle combinations. Master's thesis, Ohio State University.

PEARCE, DARREN. 2001a. Synonymy in collocation extraction. In *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, USA.

——. 2001b. Using conceptual similarity for collocation extraction. In *Proc. of the 4th UK Special Interest Group for Computational Linguistics (CLUK4)*.

——. 2002. A comparative evaluation of collocation extraction techniques. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands.

PIAO, SCOTT S.L., PAUL RAYSON, DAWN ARCHER, ANDREW WILSON, and TONY MCENERY. 2003. Extracting multiword expressions with a semantic tagger. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 49–56, Sapporo, Japan.

PILZ, KLAUS DIETER. 1981. *Phraseologie*. Stuttgart, Germany: Sammlung Metzler.

POLLARD, CARL, and IVAN A. SAG. 1994. *Head-driven Phrase Structure Grammar*. Chicago, USA: The University of Chicago Press.

PULMAN, STEPHEN G. 1993. The recognition and interpretation of idioms. In (Cacciari and Tabossi 1993a), chapter 11.

PUSTEJOVSKY, JAMES. 1995. *The Generative Lexicon*. Cambridge, USA: MIT Press.

QUIRK, RANDOLPH, SIDNEY GREENBAUM, GEOFFREY LEECH, and JAN SVARTVIK. 1985. *A Comprehensive Grammar of the English Language*. London, UK: Longman.

RACKOW, ULRIKE, IDO DAGAN, and ULRIKE SCHWALL. 1992. Automatic translation of noun compounds. In *Proc. of the 14th International Conference on Computational Linguistics (COLING '92)*, 1249–53, Nantes, Frances.

RAPP, REINHARD. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. of the 37th Annual Meeting of the ACL*, 1–17, College Park, USA.

RENOUF, A., and J.M. SINCLAIR. 1991. Collocational frameworks in English. In *English Corpus Linguistics*, ed. by K. Aijmer and B. Altenberg, 128–43. London: Longman.

RESNIK, PHILIP, 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. University of Pennsylvania dissertation.

RIEHEMANN, SUSANNE, 1997. Idiomatic constructions in HPSG. Presented at the 4th International Conference on HPSG.

——, 2001. *A Constructional Approach to Idioms and Word Formation*. Stanford, USA dissertation.

RIES, K., F.D. BUØ, and A. WAIBEL. 1996. Class phrase models for language modelling. In *Proc. of the 4th International Conference on Spoken Language Processing (ICSLP'96)*.

ROSARIO, BARBARA, and MARTI HEARST. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, USA.

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

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002a. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, ed. by Alexander Gelbuk, 1–15. Hiedelberg/Berlin: Springer-Verlag.

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

——, Thomas Wasow, and Emily Bender. 2003. *Syntactic Theory: A Formal Introduction, 2nd Edition*. Stanford, USA: CSLI Publications. Chapter 8: The Structure of the Lexicon.

Salton, Gerald, and Maria Smith. 1990. On the application of syntactic methodologies in automatic text analysis. In *Proc. of 13th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'90)*, 137–50, Cambridge, USA.

Sawyer, Joan H., 1999. *Verb Adverb and Verb Particle Constructions: Their Syntax and Acquisition*. Boston University dissertation.

Schenk, André. 1995. The syntactic behavior of idioms. In (Everaert *et al.* 1995a), chapter 10.

Schenk, Anreé. 1986. Idioms in the Rosetta machine translation system. In *Proc. of the 11th International Conference on Computational Linguistics (COLING '86)*, Bonn, Germany.

Schone, Patrick, and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, 100–108.

Schreuder, Robert. 1990. Lexical processing of verbs with separable particles. *Yearbook of Morphology* 3.65–79.

Sekine, Satoshi, Jeremy J. Carroll, Sophia Ananiadou, and Jun'ichi Tsujii. 1992. Automatic learning for semantic collocation. In *Proc. of the 3rd Conference on Applied Natural Language Processing (ANLP)*.

SELKIRK, ELISABETH O. 1982. *The Syntax of Words*. Cambridge, USA: MIT Press.

SELVER, P. 1957. *English Phraseology, a Dictionary Containing more than 5,000 Idiomatic and Colloquial Words and Expressions*. London : J. Brodie.

SHIMAMURA, REIKO. 1998. Lexicalization of syntactic phrases: The case of genitive compounds like *Woman's Magazine*. In *Proc. of the Kanda University of International Studies Graduate School of Language Sciences Centre of Excellence in Linguistics (COE) International Workshop*.

——. 2001. The A-N expression within the compound and the phrase/word distinction. In *Proc. of the Linguistics Society of America (LSA) Annual Meeting*.

SHIMOHATA, SAYORI, TOSHIYUKI SUGIO, and JUNJI NAGATA. 1997. Retrieving collocations by co-occurrences and word order constraints. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, 476–81, Madrid, Spain.

SHIRAI, SATOSHI, YOSHIFUMI OOYAMA, SHINOBU TAKECHI, KEIKO WAKEBE, and HIROSHI AIZAWA. 1998. Compiling Japanese and English corpus for compound verbs of Japanese origin. In *Proc. of the 57th Annual Meeting of IPSJ*, 267–8, Nagoya, Japan. In Japanese.

SIDE, RICHARD. 1990. Phrasal verbs: sorting them out. *ELT Journal* 44.144–52.

SINCLAIR, JOHN M. 1987. Collocation: A progress report. In *Language Topics: Essays in Honour of Michael Halliday*, ed. by R. Steele and T. Threadgold, volume II, 319–31. John Benjamins.

SMADJA, FRANK. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19.143–77.

——, KATHLEEN R. MCKEOWN, and VASILEIOS HATZIVASSILOGLOU. 1996a. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22.1–38.

——, ——, and ——. 1996b. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22.1–38.

SMADJA, FRANK A., and KATHLEEN R. MCKEOWN. 1990. Automatically extracting and representing collocations for language generation. In *Proc. of the 28th Annual Meeting of the ACL*, 252–9.

SMITH, JEFFREY D. 1999. English number names in hpsg. In *Lexical and Constructional Aspects of Linguistic Explanation*, ed. by Andreas Kathol, Jean-Pierre Koenig, Gert Webelhuth, and eds., 145–160. Stanford, USA: CSLI Publications.

SOEHN, JAN-PHILIPP, and MANFRED SAILER. 2003. At first blush on tenterhooks. About selectional restrictions imposed by nonheads. In *Proceedings of Formal Grammar 2003*, ed. by Gerhard Jäger, Paola Monachesi, Gerald Penn, and Shuly Winter, 149–161.

SPROAT, RICHARD W. 1994. English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language* 8.79–94.

——, and MARK Y. LIBERMAN. 1987. Toward treating English nominals correctly. In *Proc. of the 25th Annual Meeting of the ACL*, Stanford, USA.

STOCK, OLIVERIO, JON SLACK, and ANDREW ORTONY. 1993. Building castles in the air: Some conceptual and theoretical issues in idiom comprehension. In (Cacciari and Tabossi 1993a), chapter 10.

STOCK, OLIVIERO. 1987. Getting idioms into a lexicon based parser's head. In *Proc. of the 25th Annual Meeting of the ACL*, 52–8, Stanford, USA.

STOWELL, TIM, 1981. *Origins of Phrase Structure*. MIT dissertation.

STVAN, LAUREL SMITH, 1998. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. Northwestern University dissertation.

SWEETSER, EVE. 1990. *From etymology to pragmatics*. Cambridge, UK: Cambridge University Press.

TANAKA, TAKAAKI. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, 981–7, Taipei, Taiwan.

——, and TIMOTHY BALDWIN. 2003a. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 17–24, Sapporo, Japan.

——, and ——. 2003b. Translation selection for Japanese-English noun-noun compounds. In *Proc. of the Ninth Machine Translation Summit (MT Summit IX)*, 89–96, New Orleans, USA.

——, and YOSHIHIRO MATSUO. 1999. Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, 109–19, Chester, UK.

TER STAL, W., and P. VAN DER VET. 1994. Two-level semantic analysis of compounds: A case study in linguistic engineering. In *Papers from the 4th CLIN Meeting*.

TOUTANOVA, KRISTINA, and CHRISTOPER D. MANNING. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong, China.

TSCHICHOLD, CORNELIA. 1995. English multi-word lexemes in a lexical database. In *Proc. of the ESSLI Workshop on the Computational Lexicon*, ed. by M. Felisa Verdejo.

——, 1998. *Multi-word Units in Natural Language Processing*. University of Basel dissertation.

TSUJI, KEITA, and KYO KAGEURA. 2001. Extracting morpheme pairs from bilingual terminological corpora. *Terminology* 7.101–14.

ČERMÁK, F. 1988. On the substance of idioms. *Folia Linguistica* 22.413–38.

—— 1994. Idiomatics. In *Prague School of Structural and Functional Linguistics*, ed. by P.A. Luelsdorff, 185–95. Amsterdam and Philadelphia: John Benjamins.

UCHIYAMA, KIYOKO, and TIMOTHY BALDWIN. 2004. A machine learning method for disambiguating Japanese verb compounds. In *Proc. of the 10th Annual Meeting of the Association for Natural Language Processing (Japan)*, Tokyo, Japan. (in Japanese).

——, ——, and SHUN ISHIZAKI. to appear. Disambiguating japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions* .

VAN DE VOORT, MARLIES E.C., and WIETSKE VONK. 1995. You don't die immediately when you kick an empty bucket: A processing view on semantic and syntactic characteristics of idioms. In (Everaert *et al.* 1995a), chapter 12.

VAN DER LINDEN, ERIK-JAN. 1992. Incremental processing and the hierarchical lexicon. *Computational Linguistics* 18.219–38.

VAN GESTEL, FRANK. 1995. En bloc insertion. In (Everaert *et al.* 1995a), chapter 4.

VAN LANCKER, DIANA, and GERALD J. CARTER. 1981. Idiomatic versus literal interpretations of ditropically ambiguous sentences. *Journal of Speech and Hearing Research* 24.64–9.

——, and DALE TERBEEK. 1981. Disambiguation of ditropic sentences: Acoustic and phonetic clues. *Journal of Speech and Hearing Research* 24.330–5.

VANDERWENDE, LUCY. 1994. Algorithm for automatic interpretation of noun sequences. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan.

VELARDI, PAOLA, MARIA TERESA PAZIENZA, and MICHELA FASOLO. 1991. How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition. *Computational Linguistics* 17.153–70.

VILLAVICENCIO, ALINE. 2003a. Verb-particle constructions and lexical resources. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 57–64, Sapporo, Japan.

——. 2003b. Verb-particle constructions in the world wide web. In *Proc. of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France.

——, and ANN COPESTAKE, 2002a. Dictionaries and the regularities in phrasal verbs. *LinGO Working Paper No. 2002-03*.

——, and ANN COPESTAKE, 2002b. The nature of idioms. *LinGO Working Paper No. 2002-04*.

——, and ANN COPESTAKE, 2002c. Phrasal verbs and the LinGO-ERG. *LinGO Working Paper No. 2002-01*.

——, and ANN COPESTAKE, 2002d. The treatment of multiword expressions in the LKB system. *LinGO Working Paper No. 2002-05*.

——, and ANN COPESTAKE. 2002e. Verb-particle constructions in a computational grammar of English. In *Proc. of the 9th International Conference on Head-Driven Phrase Structure Grammar (HPSG-2002)*, Seoul, South Korea.

WACHOLDER, NINA, and PENG SONG. 2003. Toward a task-based gold standard for evaluation of np chunks and technical terms. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, 189–96, Edmonton, Canada.

WASOW, THOMAS. 2002. *Postverbal Behavior*. Stanford, USA: CSLI Publications.

WEHRLI, ERIC. 1998. Translating idioms. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics: COLING/ACL-98*, 1388–92, Montreal, Canada.

WEINREICH, URIEL. 1963. Lexicology. In *Current Trends in Linguistics*, ed. by T. Sebeok, volume I, 60–93. The Hague: Mouton.

WIDDOWS, DOMINIC, and BEATE DOROW. 2002. A graph model for unsupervised lexical acquisition. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, 1093–9, Taipei, Taiwan.

WOODS, WILLIAM. 1972. *The Lunar sciences natural language information system*. Final Report, Bolt, Beranek and Newman, Cambridge, MA.

WU, DEKAI. 1993. Approximating maximum-entropy ratings for evidential parsing and semantic interpretation. In *Proc. of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1290–6.

WUNDERLICH, DIETER. 1983. On the compositionality of German prefix verbs. In *Meaning, Use and the Interpretation of Language*, ed. by R. Bäuerle, C. Schwarze, and A. von Stechow, 452–65. De Gruyter.

WURMBRAND, SUSI, 2000. The structure(s) of particle verbs. Talk given at the DGfS 2000 (AG3: Semantisch komplexe Verben und ihre Argumentstruktur), Marburg, Germany.

Y, G., and M. JOHNSON. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

YAMAMOTO, MIKIO, and KENNETH W. CHURCH. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics* 27.1–30.

ZELLER, JOCHEN, 1997. Particle verbs and a theory of late lexical insertion. University of Frankfurt.

——, 1999. *Particle verbs, local domains, and a theory of lexical licensing*. University of Frankfurt dissertation.

——. 2001. *Particle Verbs and Local Domains*. Amsterdam: John Benjamins.

ZHANG, LEI, JIANFENG GAO, and MING ZHOU. 2000. Extraction of Chinese compound words – an experimental study on a very large corpus. In *Proc. of the 2nd Chinese Language Processing Workshop, ACL 2000*.

ZWICKY, ARNOLD. 1985. Clitics and particles. *Language* 61.283–305.