

Overview: Computational Lexical Semantics and the Week Ahead

Diana McCarthy

University of Melbourne, July 2011

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Motivation

- we interpret and use language for communication
- words have meaning
- if we want to machines to manipulate language as we do they need to be able to distinguish meanings and use words appropriately

Drawbacks

- semantic phenomena covert
- what is the appropriate representation?
- more variation compared to syntax and morphology
- less straightforward to evaluate, unless we focus on easy distinctions

The Emergence of Computational Lexical Semantics

- importance of the lexicon, 80's onwards [Gazdar, 1996]
- default inheritance (expressing generalisations e.g. DATR
<http://www.informatics.susx.ac.uk/research/groups/nlp/datr/>)
- word sense disambiguation [Weaver, 1949]
- importance of lexical semantics (growing fast)
 - word meanings
 - semantic relationships

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Representing Word Meaning

- with other words
 - from manually produced resources e.g. dictionaries, thesauruses
 - automatic extraction (from corpora)
- potentially other media [Feng and Lapata, 2010]

Manually produced inventories: e.g. WordNet

- man made on-line thesaurus
- organised by POS
- synonym sets - senses rather than word form
- relations between these sets e.g. hyponymy meronymy.

coach (noun) has 5 senses in WordNet:-

- ① (20) coach, manager, handler – ((sports) someone in charge of training an athlete or a team)
- ② coach, private instructor, tutor – (a person who gives private instruction (as in singing, acting, etc.))
- ③ passenger car, coach, carriage – (a railcar where passengers ride)
- ④ coach, four-in-hand, coach-and-four – (a carriage pulled by four horses with one driver)
- ⑤ bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger vehicle – (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work")

Lumper or Splitter?

- ① (20) coach, manager, handler – ((sports) someone in charge of training an athlete or a team)
- ② coach, private instructor, tutor – (a person who gives private instruction (as in singing, acting, etc.))
- ③ passenger car, coach, carriage – (a railcar where passengers ride)
- ④ coach, four-in-hand, coach-and-four – (a carriage pulled by four horses with one driver)
- ⑤ bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger vehicle – (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work")

Lumper or Splitter? WordNet *amount* (noun)

- (40) sum, sum of money, amount, amount of money – (a quantity of money; "he borrowed a large sum"; "the amount he had in cash was insufficient")
- (39) amount – (the relative magnitude of something with reference to a criterion; "an adequate amount of food for four people")
- (20) measure, quantity, amount – (how much there is or how many there are of something that you can quantify)
- (6) sum, amount, total – (a quantity obtained by the addition of a group of numbers)

OntoNotes [Hovy et al., 2006]

- multi-year large scale semantic annotation project
- consortium (BBN Technologies, the University of Colorado, the University of Pennsylvania, the University of Southern Californias Information Sciences Institute)
- various levels of annotation (syntax, propositions, word sense, names, coference)
- 90% agreement from annotators (Inter-tagger agreement: average pairwise agreement on same sense for an item)
- English, Chinese, Arabic

Ontonotes (2.0): *amount* noun

```
<sense n="1" type="" name="A quantity of something" group="1">  
  <commentary>AMOUNT[+quantity]
```

Note: the quantity may be referred to precisely or approximately.

Note: usually occurs with mass nouns, but usage with count nouns is increasing.</commentary>

```
  <examples>
```

We have an adequate amount of food for four people.

Writing my thesis involved a certain amount of procrastination.

```
<wn version="2.1">1 2 4</wn>
```

```
<sense n="3" type="" name="A sum of money" group="1">
```

```
  <commentary> AMOUNT[+quantity][+sum][+money]
```

Note: always refers to a quantity of money.

Note: a narrow, specialized use of Sense 1</commentary>

```
  <examples>
```

He borrowed a large amount when he started that business.

The amount he had in his wallet was insufficient.

...

```
<wn version="2.1">3</wn>
```

Roget: *amount* noun

- Definition: quantity
 - Synonyms: aplenty, bags, bulk, bundle, chunk, expanse, extent, flock, gob, heap, hunk, jillion, load, lot, magnitude, mass, measure, mess*, mint, mucho, number, oodles*, pack, passel, peck, pile, scads, score, slat, slew, supply, ton, volume, whopper
 - Notes: use ' amount ' with things that cannot be counted but ' number ' with things that can be counted number is regularly used with count nouns, while amount is mainly used with mass nouns: number of mistakes, amount of money
- Definition: total
 - Synonyms: addition, aggregate, all, bad news, body, budget, cost, damage*, entirety, expense, extent, list, lot, net, outlay, output, price tag, product, quantum, score, set-back, sum, tab, tidy sum, whole

* = informal/non-formal usage

Homonymy vs polysemy

- homonyms: same spelling, pronunciation but different meanings. Two different 'words'
 - *bank*: a financial institution
 - *bank*: slope at the side of a river
 - different words on basis of etymology: historical origin (but not so straightforward)
- polysemes different meanings, same origins *mouth* (river or animal)
- systematic polysemes - regular difference in meaning e.g. meat - animal (*chicken, duck, goose*)
- homographs: words that are written the same but pronounced differently e.g. *lead*
- homophones: words that are pronounced the same, but written differently e.g. *two, to, too read, reed*

Homonymy vs Polysemy

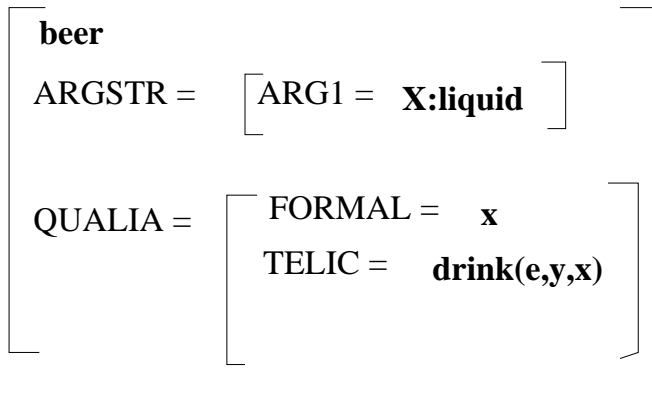
but ...

- We don't always know the etymology
- some meanings are more related than others
- how do we decide the degree of relatedness?

Generative Lexicon [Pustejovsky, 1995]

- related senses generated from rules capturing regularities
- lexical typing structure, argument structure, event structure and qualia structure
- senses expressed by qualia roles (semantic features)
 - formal (type or relation)
 - constitutive (relation between object and its parts)
 - telic (purpose or function)
 - agentive (origins)

Generative Lexicon [Pustejovsky, 1995]



Generative Lexicon [Pustejovsky, 1995]

book

ARGSTR = $\left[\begin{array}{l} \text{ARG1} = \mathbf{x:\text{information}} \\ \text{ARG2} = \mathbf{y:\text{phys_obj}} \end{array} \right]$

QUALIA = $\left[\begin{array}{l} \mathbf{\text{information} . \text{phys_obj_lcp}} \\ \text{FORMAL} = \mathbf{\text{hold}(y,x)} \\ \text{TELIC} = \mathbf{\text{read}(e,w,x . y)} \\ \text{AGENT} = \mathbf{\text{write}(e',v,x . y)} \end{array} \right]$

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - **Distributional Similarity**
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Distributional approaches to Word Meaning

plant:

- ① *see if you can make the **plant** grow to its full and healthy height*
- ② *A hydro power **plant** can be operated using either a diverted water stream system*
- ③ *Job profile of a water/ wastewater treatment **plant** worker*
- ④ *We know from a very early age that **plants** obtain water through their roots*

Distributional approaches to Word Meaning

plant:

- ① *see if you can make the **plant** grow to its full and healthy height*
- ② *A hydro power **plant** can be operated using either a diverted **water** stream system*
- ③ *Job profile of a **water**/ wastewater treatment **plant** worker*
- ④ *We know from a very early age that **plants** obtain **water** through their roots*

water	grow	root	job	hydro	power ...
3	1	1	1	1	1

Proximity Relations

context	frequency		
	<i>plant</i>	<i>tree</i>	<i>factory</i>
<i>worker</i>	55	8	45
<i>healthy</i>	32	21	3
<i>water</i>	34	18	10
<i>root</i>	8	6	0
<i>operate</i>	4	1	23
<i>power</i>	3	1	49

Dependency Relations

context		frequency		
		<i>plant</i>	<i>tree</i>	<i>factory</i>
<i>grow</i>	verb object	52	60	10
<i>weed</i>	verb object	31	23	2
<i>water</i>	verb object	23	15	4
<i>dead</i>	adj modifier	10	12	0
<i>operate</i>	verb subject	16	2	22
<i>demolish</i>	verb object	11	5	15

Distributional similarity: nearest neighbours

Thesaurus (nearest neighbour) output

Word: <closest word> <score> <2nd closest > <score>...

Distributional similarity: nearest neighbours

Thesaurus (nearest neighbour) output

Word: <closest word> <score> <2nd closest > <score>...

plant: tree 0.178 flower 0.163 factory 0.152 bush 0.131

coach: train 0.171 bus 0.166 player 0.149 captain 0.131 car 0.131

match: game 0.171 tournament 0.166 matchstick 0.149 cigarette
0.131 competition 0.131

Distributional similarity: nearest neighbours

Thesaurus (nearest neighbour) output

Word: <closest word> <score> <2nd closest> <score>...

plant: tree 0.178 flower 0.163 factory 0.152 bush 0.131

coach: train 0.171 bus 0.166 player 0.149 captain 0.131 car 0.131

match: game 0.171 tournament 0.166 matchstick 0.149 cigarette
0.131 competition 0.131

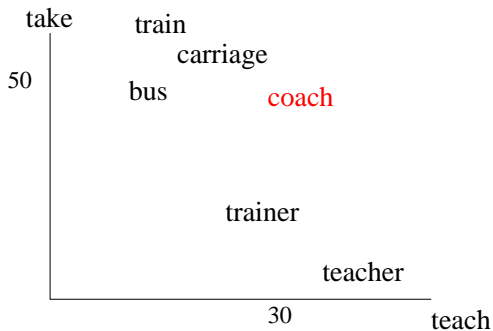
Grouping similar words [Pantel and Lin, 2002]

Distributional Similarity: Vector Space Models

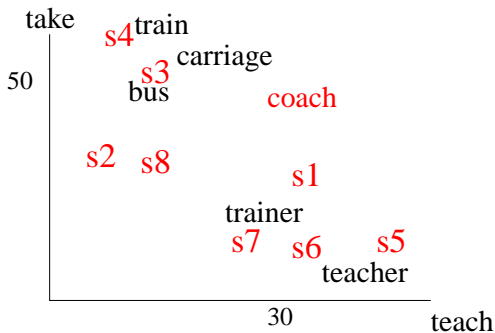
Frequency data input to a vector space model:

context	target words		
	<i>coach</i>	<i>bus</i>	<i>trainer</i>
<i>take</i>	50	60	10
<i>teach</i>	30	2	25
<i>ticket</i>	8	5	0
<i>match</i>	15	2	6

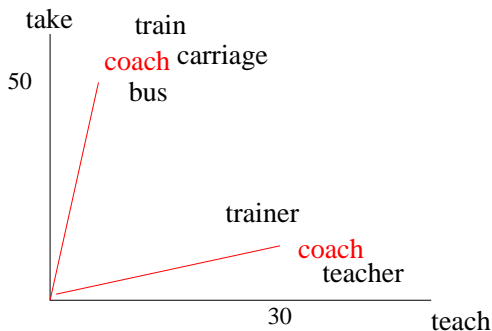
Vector Based Approaches



Vector Based Approaches



Vector Based Approaches



Prototypes and Exemplars

- Prototype single vector per category (centroid)
- cluster instances [Schütze, 1998] and then take centroid
- Multiprototype: [Reisinger and Mooney, 2010]
- can remove irrelevant vectors given the context [Thater et al., 2010]
- Exemplars: keep all vectors and compare to input (kNN, $q\%$) [Erk and Padó, 2010]

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Word Sense Disambiguation

- process of assigning words to word senses
- class of semantic tagging, but note tagging can include broader semantic classes that apply to groups of words
- which meanings?
- which words? (all or just a fixed set)

Word Sense Disambiguation

- how easy?

*[Bar-Hillel, 1960] No existing or imaginable program will enable an electronic computer to determine that the word pen is used in its **enclosure** sense in the passage below, because of the need to model, in general, all world knowledge like, for example, the relative sizes of objects:*

“Little John was looking for his toy box. Finally he found it. The box was in the pen. John was happy.”

WSD and Semantic tagging: why bother?

- NLP applications e.g.
 - question answering
 - machine translation
 - information retrieval
 - summarisation
- enabling other tasks e.g.
 - anaphora resolution
 - lexical acquisition (preferences)
 - parsing / semantic role labelling
- Lexicography, linking dictionary with corpus, synonym extraction

Word sense disambiguation (WSD)

Given a word in context, find the **best-fitting** “sense”

*Residents say militants in a station wagon pulled up, doused the building in gasoline, and struck a **match**.*

Word sense disambiguation (WSD)

Given a word in context, find the **best-fitting** “sense”

*Residents say militants in a station wagon pulled up, doused the building in gasoline, and struck a **match**.*



Word sense disambiguation (WSD)

Given a word in context, find the **best-fitting** “sense”

*Residents say militants in a station wagon pulled up, doused the building in gasoline, and struck a **match**.*



match#n#1

WSD Evaluation

- against manually tagged resources e.g. SemCor [Miller et al., 1993]
 - SemCor largest manually tagged resource
 - English, WordNet 1.6 (later versions simply remapped taggings)
 - ① 230,000 words of Brown Corpus [Francis and Kučera, 1979]:
 - ② also Red Badge of Courage
 - cntlist (both sources, used to order WordNet senses) vs release SemCor files (Brown only)
 - no inter-tagger agreement figures, but remarkable resource for its size and availability!

10743 be%2:42:03:: 1
7154 person%1:03:00:: 1
3020 be%2:42:06:: 2
2592 say%2:32:00:: 1
2333 group%1:03:00:: 1
1865 stock%1:21:00:: 1
1838 not%4:02:00:: 1
1449 man%1:18:00:: 1
1330 use%2:34:01:: 1
1268 business%1:14:00:: 1
1240 want%2:37:00:: 1
1202 have%2:40:00:: 1
1194 small%3:00:00:: 1
1158 big%3:00:01:: 1
1118 call%2:32:02:: 1
1094 walk%2:38:00:: 1
1090 hold%2:36:00:: 1
1049 house%1:06:00:: 1
1007 n't%4:02:00:: 1
992 location%1:03:00:: 1

SemCor example

```
<p pnum=1>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexsns=1:0
n_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 lexsns=2:32:00::>said<
<wf cmd=done pos=NN lemma=friday wnsn=1 lexsns=1:28:00::>Fri
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1 lexsns=1:09:0
>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1 lexsns=1:15:00::>At
<wf cmd=ignore pos=POS>'s</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2 lexsns=5:00:00:past
<wf cmd=done pos=NN lemma=primary_election wnsn=1 lexsns=1:0
on</wf>
<wf cmd=done pos=VB lemma=produce wnsn=4 lexsns=2:39:01::>pr
<punc>'</punc>
```

WSD Evaluation Measures

$$\text{Coverage} = \frac{\# \text{answers provided}}{\text{total } \# \text{items requiring answer}}$$

$$\text{Precision} = \frac{\# \text{correct answers provided}}{\# \text{answers provided}}$$

$$\text{Recall} = \frac{\# \text{correct answers provided}}{\text{total } \# \text{items requiring answer}}$$

$$F_{\alpha} = \frac{(1 + \alpha)PR}{\alpha P + R}$$

$$F_1 = \frac{2PR}{P + R}$$

harmonic mean of precision and recall, or balanced F score

Aside on Evaluation Measures

$$F_1 = \frac{2PR}{P + R}$$

- In WSD Precision denominator is subset of Recall.
- Lexical acquisition meanwhile uses terms for trade off (finding things, and are they correct) and not interested in true negatives (in abundance).

$$\textit{precision} = \frac{\textit{True positives}}{\textit{true positives} + \textit{false positives}}$$

$$\textit{recall} = \frac{\textit{true positives}}{\textit{true positives} + \textit{false negatives}}$$

For some semantic tasks also see

Correlation Spearman's ρ

$$\rho(X, Y) = \frac{\text{covariance}(X, Y)}{\sigma_X \sigma_Y}$$

- Pearson's coefficient
- correlation between two random variable (X and Y)
- Spearman's (non parametric) uses ranks rather than absolute values
- ρ tends to yield smaller coefficients compared to parametric counterparts [Mitchell and Lapata, 2008]

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Semantic Relations

- in addition to lexical information we often want information on the relationships between two lexical items
- semantic relations e.g. synonymy
- syntactic relations e.g. subcategorization *eat*, *direct object*
- semantics and syntax e.g. selectional preferences

Semantic Relations

- gem jewel

Semantic Relations

- gem jewel **synonyms**

Semantic Relations

- gem jewel **synonyms**
- dog animal

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**
- stroll, walk

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**
- stroll, walk **troponym**

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**
- stroll, walk **troponym**
- cough, make a noise

Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**
- stroll, walk **troponym**
- cough, make a noise **entailment**

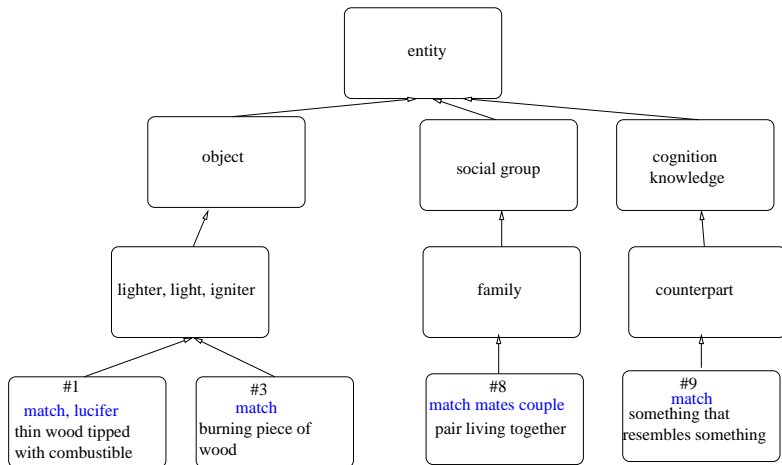
Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**
- stroll, walk **troponym**
- cough, make a noise **entailment**
- hot cold

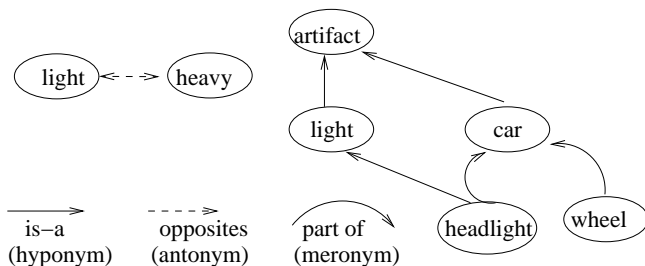
Semantic Relations

- gem jewel **synonyms**
- dog animal **hyponym**
- animal cat **hypernym** (or **hyperonym** [Sampson, 2000])
- car bus **co-hyponyms**
- hand body **meronym**
- tree bark **holonym**
- stroll, walk **troponym**
- cough, make a noise **entailment**
- hot cold **antonym**

WordNet Provides Semantic Relationships



WordNet Provides Semantic Relationships



Acquiring Semantic Relations

hypernyms [Hearst, 1992] output of a parser and then bootstrap patterns

e.g. such NP as {NP ,} {(or | and)} NP

... works by such authors as Herrick, Goldsmith, and Shakespeare

\Rightarrow *hypo(author, Herrick), hypo(author, Goldsmith), hypo(author, Shakespeare)* NP {, NP} * {,} or other NP

Bruises, wounds, broken bones or other injuries ...

\Rightarrow *hypo(bruise, injury), hypo(wound, injury), hypo(broken bone, injury)*

extended [Snow et al., 2004, Snow et al., 2006]

Synonyms and Antonyms

- distributional similarity [Padó and Lapata, 2007, McCarthy et al., 2010]
- interference from different relations [Weeds et al., 2004, Geffet and Dagan, 2004]
- ruling out antonyms [Lin et al., 2003]
patterns from X to Y, either X or Y, \Rightarrow
- antonym discovery (uses a thesaurus) [Mohammad et al., 2008]

WordNet Similarity

- Leacock and Chodrow [Leacock and Chodorow, 1998] path based, scaled by depth of hierarchy
- lesk [Lesk, 1986]: gloss overlap, uses semantic relations

$$lesk(s_1, s_2) = |\{w_1 \in definition(s_1)\} \cap \{w_2 \in definition(s_2)\}| \quad (1)$$

- Information Content e.g. jcn [Jiang and Conrath, 1997]: uses frequency counts from corpus

$$IC(s) = -\log(p(s)) \quad (2)$$

Probability of a concept (s), high information content for very specific terms

Jiang and Conrath specify a distance measure:

$$D_{jcn}(s_1, s_2) = IC(s_1) + IC(s_2) - 2 \times IC(s_3) \quad (3)$$

Use of WSD and relationships in Lexical chains

It was an important moment for Jake, all his friends and family were watching him. There was only a minute of the game left and neither team had scored yet. The crowd watched in silence as Jake took the penalty shot.

Use of WSD and relationships in Lexical chains

*It was an important moment for Jake, all his friends and family were watching him. There was only a minute of the **game** left and neither **team** had **scored** yet. The **crowd** watched in silence as Jake took the **penalty shot**.*

textual cohesion (linguistics) [Halliday and Hasan, 1976]
structure of texts [Morris and Hirst, 1991]

Associating Lexical Inventories with Corpus Data

- Selectional Preferences [Resnik, 1993]
 - argument head data e.g. direct objects of *eat*
 - propagate frequencies in noun hierarchy
- feature vectors at senses [Pantel, 2005]
 - propagate features shared by hyponyms
 - second phase (disambiguate) remove features at leaf that are in other senses parents

Domain and Topic Information

- topic signatures [Agirre et al., 2001] attributed to senses, retrieved from documents, pertinent to this sense but not others of the same word

topic signature

star/celebrity	gossip marriage divorce screen actor football
star/celestial	planet galaxy space telescope science journal

- domain
models [Magnini and Cavaglià, 2000, Magnini et al., 2002] attributed to senses
- sentiment (pragmatics, but most people focusing on semantic prosodies of words) [Wiebe and Mihalcea, 2006]

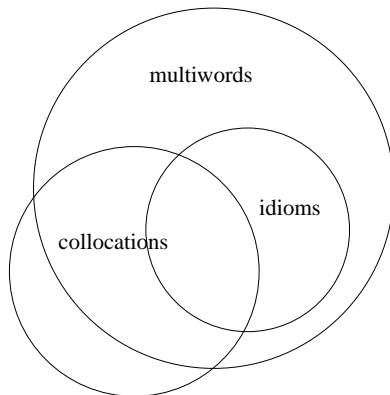
Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - **Multiword Expressions**
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Multiword Expression NLP Publications

- A Pain in the Neck for NLP [Sag et al., 2002]
- workshops:
 - Collocations/ Multiwords (ACL) 2001, 2003, 2004, 2007, 2009
 - Collocations (Vienna) 2002
 - Collocations and Idioms (Berlin) 2003, 2006,
 - Multiwords (LREC) 2008
 - Multiwords Coling 2010
 - Multiwords ACL 2011
- Multiword Special Issues:
 - Having a crack at a hard nut [Villavicencio et al., 2005]
 - Hard going or plain sailing? [Rayson et al., 2005]

Terminology: Multiwords, Idioms and Collocations



Multiword Expression: A Working Definition

*A multiword expression is a combination of two or more words whose semantic, syntactic etc... properties cannot fully be predicted from those of its components, and which therefore has to be listed in a lexicon.
[Boleda and Evert, ESLLI 2009]*

Motivation for finding MWEs

1 NLP

- semantic interpretation
e.g. *throw me a bone*
- associated syntactic behaviour
e.g. *blow up the houses of parliament*
- lexical acquisition e.g. *eat my hat*
- associated behaviour important for generation

2 lexicography

3 corpus linguistics

Approaches for Detecting MWEs

- statistical: e.g. pointwise mutual information

$$PMI = \log \frac{p(chew, fat)}{p(chew)p(fat)}$$

- translations in parallel text:

chew the fat \leftrightarrow *conversar*

- dictionaries:

listings, semantic codes and relationships

- lexical variation *couch potato*

sofa potato, couch onion

- syntactic variation:

take heart

- distributional similarity: *hot* and *dog* vs *hot dog*

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

Subcategorisation

She loaded the bag with chicken
NP V NP PP

Subcategorisation

She loaded the bag with chicken
NP V NP PP_with

Subcategorisation

She loaded the bag with chicken
NP V NP PP_with

He loaded chicken into the bag
NP V NP PP_into

Selectional Preferences

<i>She</i>	<i>loaded</i>	<i>the bag</i>	<i>with chicken</i>
NP	V	NP	PP

Selectional Preferences

She *loaded* *the bag* *with chicken*
NP V NP PP

load *with ?*

Selectional Preferences

She *loaded* *the bag* *with chicken*
NP V NP PP

load *with ?*

load NP *with ?*

Selectional Preferences

She *loaded* *the bag* *with chicken*
NP V NP PP

load *with ?*

load NP *with ?*

explosive ammunition scrap fish supplies brick fat food water ...

Semantic Role Labelling

<i>She</i>	<i>loaded</i>	<i>the bag</i>	<i>with chicken</i>
NP	V	NP	PP

Semantic Role Labelling

<i>She</i>	<i>loaded</i>	<i>the bag</i>	<i>with chicken</i>
NP	V	NP	PP

FrameNet style labels [Ruppenhofer et al., 2010]

agent	predicate	object / goal	theme
-------	-----------	---------------	-------

Propbank style labels [Palmer et al., 2005]

Arg0	predicate	Arg2	Arg1
------	-----------	------	------

SRL identify the arguments of a given verb and assign them semantic labels describing the roles they fulfil

Diathesis Alternations

She loaded the bag with chicken
She loaded chicken into the bag

Lexical Information: Verb Class

Pour Verbs: *dribble, drop, pour, slop, slosh, spew, spill, spurt*

Causative Alternation:

I pour water into the pot \leftrightarrow *Water poured into the pot*

*Locative Alternation:

I pour water into the pot \leftrightarrow **I poured the pot with water*

*Conative Alternation:

I pour water into the pot \leftrightarrow **I poured at water into the pot*

Outline

- 1 Computational Lexical Semantics
 - Word Meaning Representation
 - Distributional Similarity
 - Word Sense Disambiguation
 - Semantic Relations
 - Multiword Expressions
 - Predicate Argument Structure: the syntax-semantics interface
- 2 My Background and Research Interests
 - Academic Interests
 - Commercial Interests /Demos
 - Sketch Engine, and related tools
 - Dante
 - Other Related Projects

My background: academia

- syntax-semantics interface
 - subcategorization frames
 - selectional preferences
 - diathesis alternations
- word sense disambiguation
 - selectional preferences
 - prior sense distributions
 - evaluation
- distributional similarity
- multiwords and compositionality
- lexical substitution
- lexical semantics for reading comprehension

My background: recent commercial

- Using Computational Linguistics for Corpus Linguistics
- Corpora for computational linguistics
- Corpus linguistics e.g. CLAEVIPS
- Corpus linguistics for lexicography e.g. Dante
- Learner English

Sketch Engine Demos: Preliminaries

- corpus (plural: corpora):
- concordancer (output: concordance):
- collocation:
- word sketch:
- distributional thesaurus:
- Web Boot Cat:

Sketch Engine Demos: Preliminaries

- **corpus (plural: corpora)**: a large set of texts for studying language as it is used in real life
- **concordancer (output: concordance)**:
- **collocation**:
- **word sketch**:
- **distributional thesaurus**:
- **Web Boot Cat**:

Sketch Engine Demos: Preliminaries

- **corpus (plural: corpora)**: a large set of texts for studying language as it is used in real life
- **concordancer (output: concordance)**: A program which displays all occurrences from the corpus for a given query
- **collocation**:
- **word sketch**:
- **distributional thesaurus**:
- **Web Boot Cat**:

Sketch Engine Demos: Preliminaries

- **corpus (plural: corpora)**: a large set of texts for studying language as it is used in real life
- **concordancer (output: concordance)**: A program which displays all occurrences from the corpus for a given query
- **collocation**: a sequence of words that co-occur more often than would be expected by chance.
- **word sketch**:
- **distributional thesaurus**:
- **Web Boot Cat**:

Sketch Engine Demos: Preliminaries

- **corpus (plural: corpora)**: a large set of texts for studying language as it is used in real life
- **concordancer (output: concordance)**: A program which displays all occurrences from the corpus for a given query
- **collocation**: a sequence of words that co-occur more often than would be expected by chance.
- **word sketch**: a corpus-based summary of a word's grammatical and collocational behaviour.
- **distributional thesaurus**:
- **Web Boot Cat**:

Sketch Engine Demos: Preliminaries

- **corpus** (plural: **corpora**): a large set of texts for studying language as it is used in real life
- **concordancer** (output: **concordance**): A program which displays all occurrences from the corpus for a given query
- **collocation**: a sequence of words that co-occur more often than would be expected by chance.
- **word sketch**: a corpus-based summary of a word's grammatical and collocational behaviour.
- **distributional thesaurus**: an automatically produced 'thesaurus' which finds words that tend to occur in similar contexts as the target word.
- **Web Boot Cat**:

Sketch Engine Demos: Preliminaries

- **corpus** (plural: **corpora**): a large set of texts for studying language as it is used in real life
- **concordancer** (output: **concordance**): A program which displays all occurrences from the corpus for a given query
- **collocation**: a sequence of words that co-occur more often than would be expected by chance.
- **word sketch**: a corpus-based summary of a word's grammatical and collocational behaviour.
- **distributional thesaurus**: an automatically produced 'thesaurus' which finds words that tend to occur in similar contexts as the target word.
- **Web Boot Cat**: a web-based tool for building corpora instantly from publicly accessible documents on the web

Purposes

- teaching language (schools, second language learners)
- linguistics research and teaching
- lexicography
- computational linguistics
- translation

CLAEVIPS: A Corpus Linguistics Analysis of Ecosystems Vocabulary in the Public Sphere

- commissioned by the UK National Ecosystem Assessment (NEA)
- 100 words and phrases concerning the ecosystem
- 4 corpora:
 - UKWaC [Ferraresi et al., 2008]
 - 3 specialised corpora

CLAEVIPS: Corpora

- ukWaC [Ferraresi et al., 2008] 1.5 billion word corpus from internet domains ending '.uk'
- three specialised corpora harvest from the web. Web pages contain at least:
 - three types from a set of seed words, and
 - at least three occurrences of a subset of whitelist words
- the three corpora (each approx 1.5 million words)
 - 1 academic (ac.uk)
 - 2 government (.gov.uk)
 - 3 public (news, NGO, blogs)

CLAEVIPS: Methodology

- examine salient collocates using ‘word sketch’ (words), and contrasted in the 3 specialised corpora
- examine 100 random citations from UKWaC:
 - subjective/objective
 - positive / negative / neutral
 - other ...
- (phrases) find collocates in above citations and contrast to 50 random from specialised corpora
- some words selected for additional examination using thesaurus and sketch diff

CLAEVIPS: (some) Findings [Wild et al., 2011]

- words not widely understood e.g. *biotype*, *natural capital*
- differences in specialised corpora e.g. public interest in rainforest and global warming
- promotional use of nature in advertising ‘eco’
- nature as a commodity (esp government corpus)
- in ukWaC and public corpus: evidence of scepticism regarding empty use of words *sustainable* and claims on climate change
- relationship between humans and nature
- fear of open spaces
- avoid reference to agency with words such as *pollute*, see also [Schleppegrell, 1997]

Dante: Database of Analysed Texts of English [Atkins et al., 2010]

- commissioned by Foras na Gaeilge for production of New English Irish Dictionary
- lexical resource as monolingual analysis of English
- corpus based. Lexicographers produced using Word Sketches from a corpus of 1.7 billion words (UKWaC, American newspaper, Irish English data)
- concordance sorted according to the 'GDEX' program
- containing entries for:
 - 42,000 headwords (6,300+ verbs)
 - 27,000 idioms and phrases
 - 20,500 compounds
 - just under 3,000 phrasal verbs

Dante: Contents

- meanings with definitions
- over 622,000 examples from the corpus,
- argument structure (valency) e.g. NP-Vinf *let him go* (42 frames for verbs, further specified by preposition)
- attitude e.g. *meddle* (pejorative)
- regional e.g. *nick* (British) as in *you're nicked*
- style e.g. *oxidise* (technical) *perambulate* (humorous)
- register e.g. *ameliorate* (formal) *go ape* (informal)
- subject e.g. *multiply* (maths)
- time e.g. *punch* (cattle: dated) or *quoth* (obsolete)
- inherent grammar e.g. reciprocal
John marries Mary \leftrightarrow *Mary and John marry*
- support verbs e.g. *make an appeal*

see webdante.com

blend: (PoS: v)

meaning: combine

SCF: NP

corpus pattern: with plural noun as object

example: *I have very little idea of how to **blend** colour.*

corpus pattern: blend sth and sth

example: *High Points : The attempt to **blend** melodrama comedy and horror is a worthy if failed effort.*

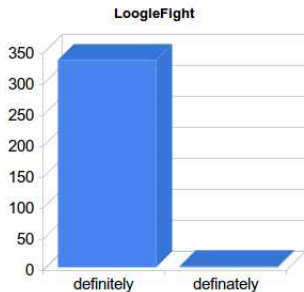
SCF: NP_PP_X with

example: *Kazakhstan was interested in **blending** palm oil with its own cotton seed and sunflower seed oils for industrial application , officials said.*

...

SCF: NP_PP_X into

example: *I **blend** different colours into the background of my paintings to evoke sections of light .*



The winner is **definitely**
 (Click on columns to see concordance)

LoogleFight takes two terms as input and finds their frequencies in the [ACL Reference Corpus](#). We hope it is a useful tool for non-native speakers of English (and possibly also native speakers) writing NLP research papers. It accesses the ACL Anthology compiled in the [Sketch Engine](#), to which free access is provided [here](#). You can input words, phrases or [CQL expressions](#)

ForBetterEnglish.com

The GDEX Demo Dictionary

This is an experimental automatic collocations dictionary, based on the Sketch Engine technology. Met
Kilgarriff et al 2008: GDEX: Automatically finding good dictionary examples in a corpus. Proc EURALE

Enter a word (in English) here to see its collocations, each with an example sentences from the corpus.

Find

involve (v)

object	everyone :	Few words can describe how delighted and proud everyone involved in the
	risk :	Scuba Diving is a sport that involves some risk to life .
	people :	For many people involved in politics the defining characteristic of the drive state intervention in reducing inequality .
	staff :	Involve primary care staff in the delivery of the programme or ensure that the same messages .
	anyone :	Useful for anyone involved , or planning to be involved , in humanitarian as
	party :	For example , ACAS mediation may involve the third party neutral issuing a
pp_at	stage :	All Welsh media channels (TV , radio and press) would need to be involved
ing_comp	widen :	Aimhigher 's remit involves widening participation in higher education .
	inject :	Awful experiments , involving injecting BSE material into the brains of living

This field sales role will involve selling both online and directory advertising

"TEDDCLoG"

Taiwan English Data-Driven Cloze Generator

Lemma	<input type="text" value="involve"/>
POS	<input type="text" value="verb"/>
Corpus	<input type="radio"/> BNC <input checked="" type="radio"/> ukWaC
Statistic used	<input checked="" type="radio"/> Saliency <input type="radio"/> Frequency
Search priority	<input checked="" type="radio"/> Collocate <input type="radio"/> Synonym
No. of Items	<input type="text" value="2"/>
No. of KOCs	<input type="text" value="3"/>
No. of Distractors	<input type="text" value="4"/>
Answers	<input checked="" type="radio"/> Blank <input type="radio"/> Underline
<input type="button" value="Submit"/>	
* For further details link	

“TEDDCLoG”

Taiwan English Data-Driven Cloze Generator

KOC 1': 'actively'

- 1) The modern father generally wants to be more actively _____ at home.
- 2) Young children learn most effectively when they are actively _____ in first hand experiences.

4 suggested distractors for KOC 'actively' :

(a) suggest (b) enable (c) allow (d) regard

KOC 2': 'heavily'

- 1) Are you heavily _____ in the visual side as well?
- 2) Those same mood swings and the need to become heavily _____ in crime also severely damage family and other relationships.

4 suggested distractors for KOC 'heavily' :

(a) mean (b) require (c) suggest (d) need

KOC 3': 'directly'

- 1) There are only a few disputes every year that directly _____ band parades.

The week ahead

- Tuesday: Word Sense disambiguation
- Wednesday: Lexical substitution, monolingual and crosslingual, motivations results and analyses
- Thursday: Alternative graded judgments of word meaning in context
- Friday: Predicate Argument Structure (very brief overview biased to my work) + discussion and project ideas for the future



Agirre, E., Ansa, O., Martinez, D., and E., H. (2001).

Enriching wordnet concepts with topic signatures.

In *Proceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*. In conjunction with NAACL."

<http://ixa.si.ehu.es/ixa/Argitalpenak/Artikuluak/1018540147/publiko>



Atkins, S., Rundell, M., and Kilgarriff, A. (2010).

Database of ANalysed Texts of English (DANTE).

In *Proceedings of Euralex*.



Bar-Hillel, Y. (1960).

The present state of automatic translation of languages.

In Alt, F. et al., editors, *Advances in Computers*. New York Academic Press.



Erk, K. and Padó, S. (2010).

Exemplar-based models for word meaning in context.

In *Proceedings of the ACL 2010 Conference Short Papers*,
ACLShort '10, pages 92–97, Stroudsburg, PA, USA.
Association for Computational Linguistics.



Feng, Y. and Lapata, M. (2010).

Visual information in semantic representation.

In *Human Language Technologies: The 2010 Annual
Conference of the North American Chapter of the Association
for Computational Linguistics*, pages 91–99, Los Angeles,
California. Association for Computational Linguistics.



Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S.
(2008).

Introducing and evaluating ukwac, a very large web-derived
corpus of english.

In *Proceedings of the Sixth International Conference on
Language Resources and Evaluation (LREC 2008)*, Marrakech,
Morocco.



Francis, W. N. and Kučera, H. (1979).

Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers.

Department of Linguistics, Brown University, Rhode Island.
Revised and amplified ed.



Gazdar, G. (1996).

Paradigm merger in natural language processing.

In Milner, R. and Wand, I., editors, *Computing Tomorrow: Future Research Directions in Computer Science*, pages 88–109. Cambridge University Press, Cambridge, UK.



Geffet, M. and Dagan, I. (2004).

Feature vector quality and distributional similarity.

In *Proceedings of Coling 2004*, pages 247–253, Geneva, Switzerland. COLING.



Halliday, M. and Hasan, R. (1976).

Cohesion In English.

Longman.



Hearst, M. (1992).

Automatic acquisition of hyponyms from large text corpora.
In Proceedings of the 14th International Conference of Computational Linguistics. COLING-92.



Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006).

Ontonotes: The 90% solution.

In Proceedings of the HLT-NAACL 2006 workshop on Learning word meaning from non-linguistic data, New York City, USA.
Association for Computational Linguistics.



Jiang, J. and Conrath, D. (1997).

Semantic similarity based on corpus statistics and lexical taxonomy.

In *International Conference on Research in Computational Linguistics*, Taiwan.



Leacock, C. and Chodorow, M. (1998).

Combining local context and WordNet similarity for word sense disambiguation.

In Fellbaum, C., editor, *WordNet: an Electronic Lexical Database*, pages 268–283. MIT Press.



Lesk, M. (1986).

Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from and ice cream cone.

In *Proceedings of the ACM SIGDOC Conference*, pages 24–26, Toronto, Canada.



Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003).

Identifying synonyms among distributionally similar words.

In *Proceedings of IJCAI-03*, pages 1492–1493.



Magnini, B. and Cavaglià, G. (2000).
Integrating subject field codes into WordNet.
In Proceedings of LREC-2000, Athens, Greece.



Magnini, B., Strapparava, C., Pezzulo, G., and GlioZZo, A.
(2002).
The role of domain information in word sense disambiguation.
Natural Language Engineering, 8(4):359–373.



McCarthy, D., Keller, B., and Navigli, R. (2010).
Getting synonym candidates from raw data in the english
lexical substitution task.
In Proceedings of the 14th Euralex International Congress,
Leeuwarden, The Netherlands.



Miller, G. A., Leacock, C., Teng, R., and Bunker, R. T.
(1993).
A semantic concordance.

In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman.



Mitchell, J. and Lapata, M. (2008).

Vector-based models of semantic composition.

In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.



Mohammad, S., Dorr, B., and Hirst, G. (2008).

Computing word-pair antonymy.

In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 982–991, Honolulu, Hawaii. Association for Computational Linguistics.



Morris, J. and Hirst, G. (1991).

Lexical cohesion, the thesaurus, and the structure of text.
Computational Linguistics, 17(1):211–2632.



Padó, S. and Lapata, M. (2007).

Dependency-based construction of semantic space models.
Computational Linguistics, 33(2):161–199.



Palmer, M., Gildea, D., and Kingsbury, P. (2005).

The proposition bank: A corpus annotated with semantic roles.

Computational Linguistics, 31(1):71–106.



Pantel, P. (2005).

Inducing ontological co-occurrence vectors.

In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 125–132, Ann Arbor, Michigan. Association for Computational Linguistics.



Pantel, P. and Lin, D. (2002).

Discovering word senses from text.

In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.



Pustejovsky, J. (1995).
The Generative Lexicon.
MIT Press, Cambridge.



Rayson, P., Piao, S., Sharoff, S., and Stefan Evert, B. n. V. M.
(2005).
Multiword expressions: hard going or plain sailing?
Language Resources and Evaluation, 44(1-2):1–5.



Reisinger, J. and Mooney, R. J. (2010).
Multi-prototype vector-space models of word meaning.
In *Proceedings of the 11th Annual Conference of the North
American Chapter of the Association for Computational
Linguistics (NAACL-2010)*, pages 109–117.



Resnik, P. (1993).
*Selection and Information: A Class-Based Approach to Lexical
Relationships*.

PhD thesis, University of Pennsylvania.



Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010).

FrameNet II: Extended theory and practice.

Technical report, International Computer Science Institute, Berkeley.

<http://framenet.icsi.berkeley.edu/>.



Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).

Multiword expressions: A pain in the neck for NLP.





In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, pages 1–15, Mexico City, Mexico.



Sampson, G. (2000).

Book review: "wordnet", ed. christiane fellbaum.

International Journal of Lexicography, 13:54–59.

-  Schleppegrell, M. (1997).
Linguistics and education.
Agency in Environmental Education, 9:49–67.
-  Schütze, H. (1998).
Automatic word sense discrimination.
Computational Linguistics, 24(1):97–123.
-  Snow, R., Jurafsky, D., and Ng, A. Y. (2004).
Learning syntactic patterns for automatic hypernym discovery.
In *Advances in Neural Information Processing Systems*,
volume 17.
-  Snow, R., Jurafsky, D., and Ng, A. Y. (2006).
Semantic taxonomy induction from heterogenous evidence.
In *Proceedings of the 21st International Conference on
Computational Linguistics and 44th Annual Meeting of the
Association for Computational Linguistics*, pages 801–808,
Sydney, Australia. Association for Computational Linguistics.



Thater, S., Fürstenau, H., and Pinkal, M. (2010).

Contextualizing semantic representations using syntactically enriched vector models.

In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 948–957, Uppsala, Sweden. Association for Computational Linguistics.



Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D. (2005).

Introduction to the special issue on multiword expressions: having a crack at a hard nut.

Computer Speech and Language, 19(4):365–377.



Weaver, W. (1949).

Translation.

In Locke, W. N. and Booth, A. D., editors, *Machine translation of languages: fourteen essays (written in 1949,*

published in 1955), pages 15–23. Technology Press of the MIT and John Wiley and Sons, Inc., New York, USA.



Weeds, J., Weir, D., and McCarthy, D. (2004).

Characterising measures of lexical distributional similarity.

In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 1015–1021, Geneva, Switzerland.



Wiebe, J. and Mihalcea, R. (2006).

Word sense and subjectivity.

In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia. Association for Computational Linguistics.



Wild, K., McCarthy, D., Church, A., and Burgess, J. (2011).

A corpus linguistics analysis of ecosystems vocabulary in the public sphere.

In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics*, Birmingham, UK.