DANTE: a New Resource for Research at the Syntax-Semantics Interface

Diana McCarthy, Lexical Computing Ltd

Interdisciplinary Verb Workshop, 5th November 2010

Overview

Background

- Lexical Information for Computational Linguistics
- Automatic Lexical Acquistion
- Existing Resources

Dante

- Lexical Information in Dante
- Demo
- A New Resource for Syntax-Semantics Research

Lexical Information: Subcategorisation

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

Lexical Information: Subcategorisation

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

Lexical Information: Subcategorisation

She	*loaded*	*the bag*	*with chicken*
NP	V	NP	PP_with
He	*loaded*	*chicken*	*into the bag*
NP	V	NP	PP_into

Lexical Information: Selectional Preferences

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

Lexical Information: Selectional Preferences

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

Lexical Information: Selectional Preferences

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

Lexical Information: 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 . . .

Lexical Information: 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*

Lexical Acquisition

SYNTAX

parsing

subcategorisation
&
argument slots

diathesis alternations

SEMANTICS

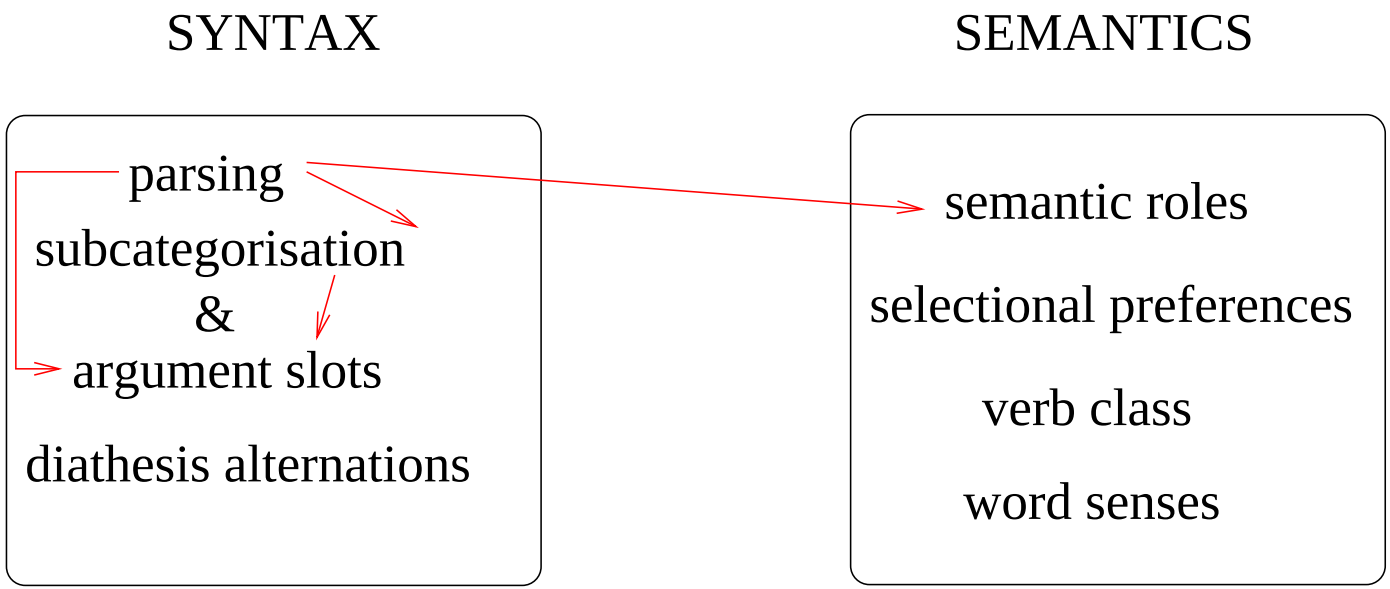
semantic roles

selectional preferences

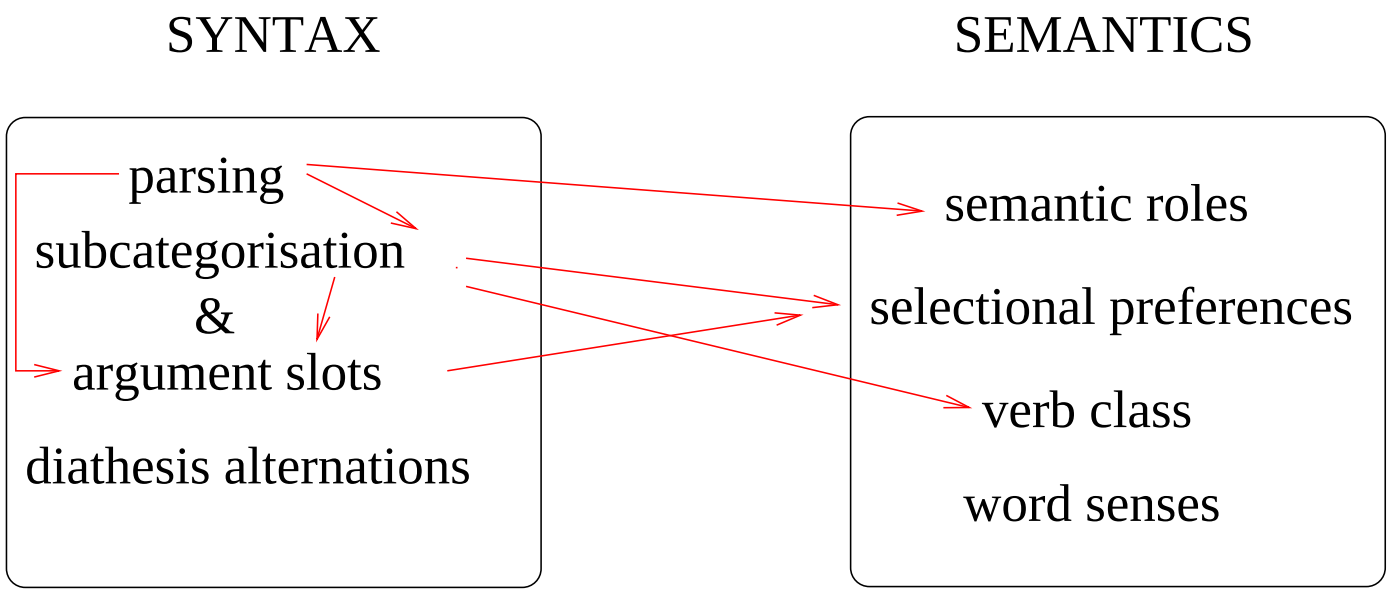
verb class

word senses

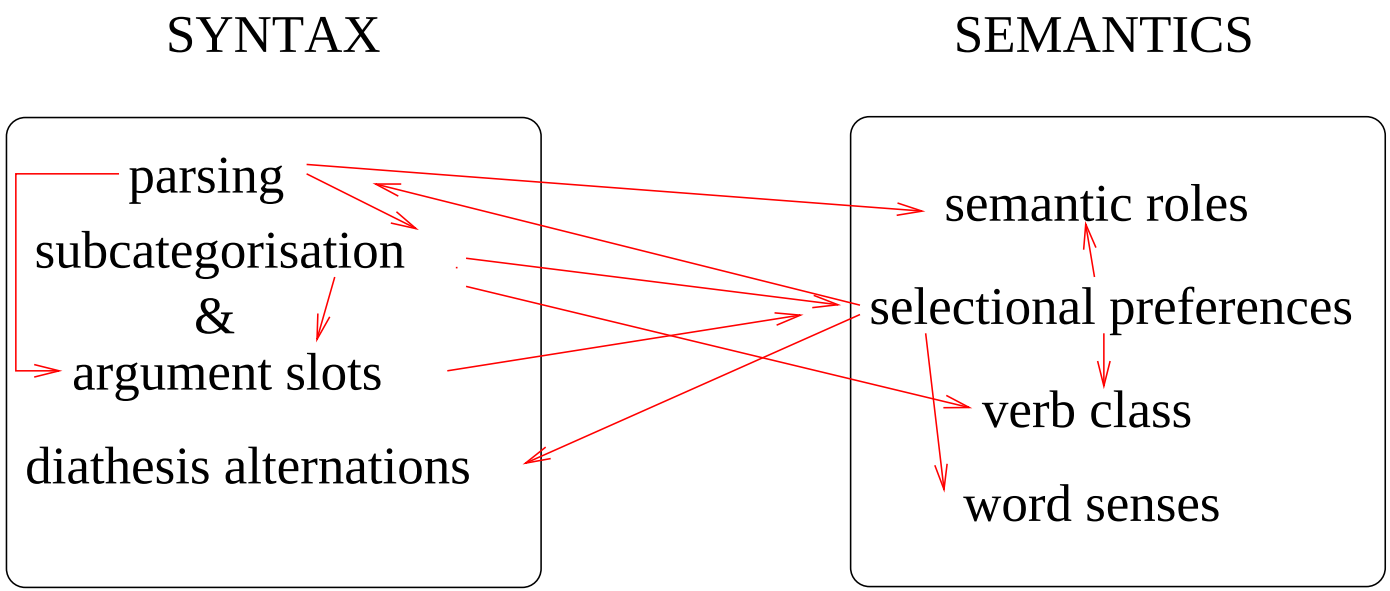
Lexical Acquisition



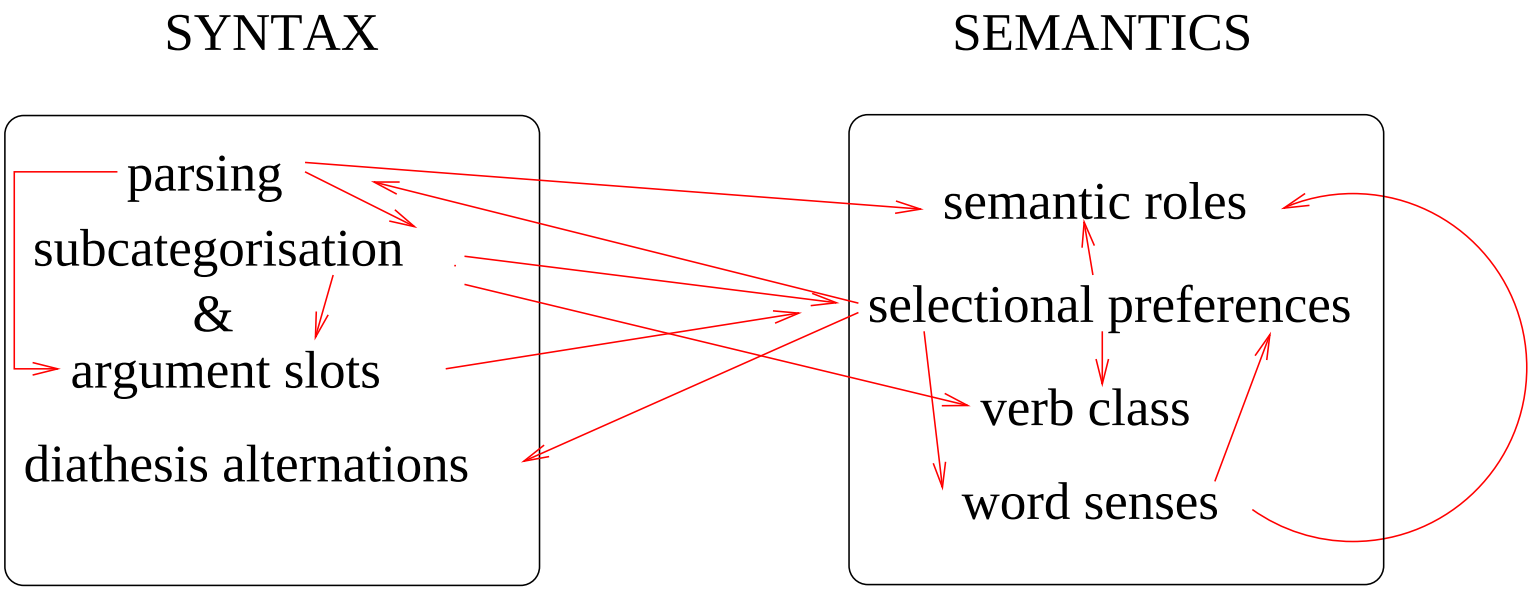
Lexical Acquisition



Lexical Acquisition



Lexical Acquisition



Subcategorisation Acquisition

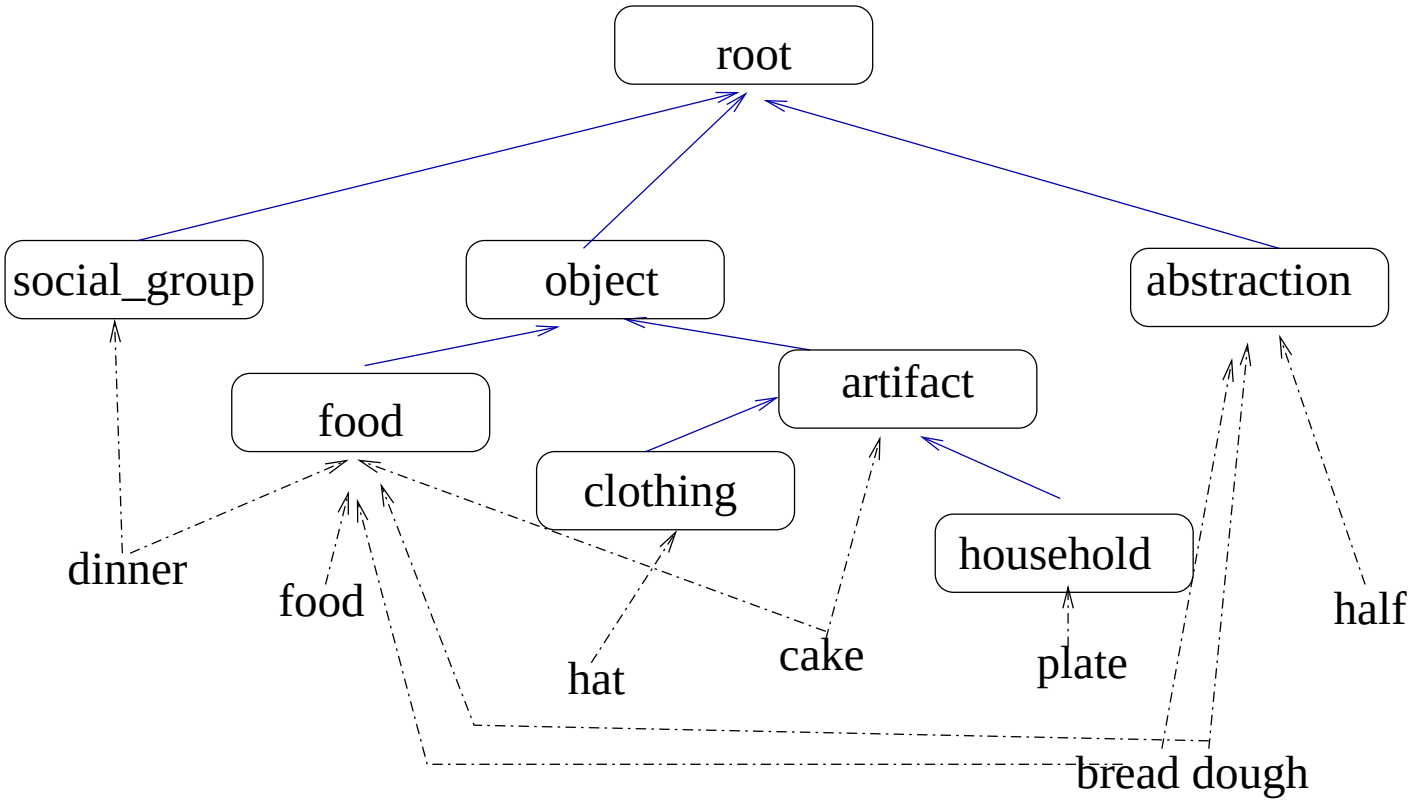
- unambiguous instances [Brent, 1991]
- parsing [Briscoe and Carroll, 1997]
- statistical filtering [Briscoe and Carroll, 1997]
- use of semantic classes for generalising [Korhonen, 2002]

Selectional Preference Acquisition

- use:
  - slots e.g. direct object [Resnik, 1993] or
  - slots in SCF [McCarthy, 2001]
- generalise argument heads with
  - WordNet [Resnik, 1993, Li and Abe, 1998]
  - distributional similarity [Erk, 2007, McCarthy et al., 2007]

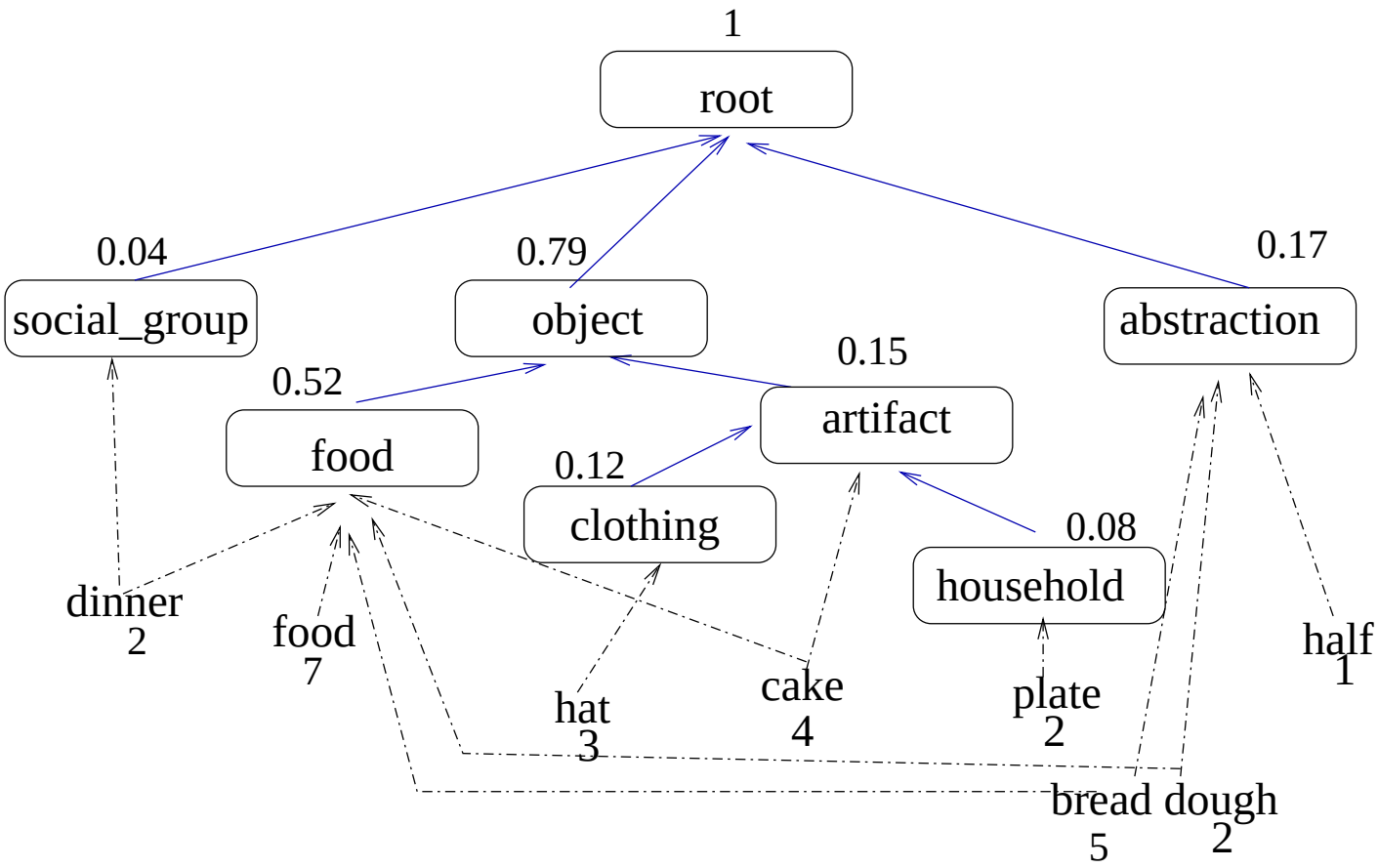
WordNet Based Models: example *eat*

food 7, bread 5, cake 4, hat 3, dinner 2, dough 2, plate 2, half 1



WordNet Based Models: example *eat*

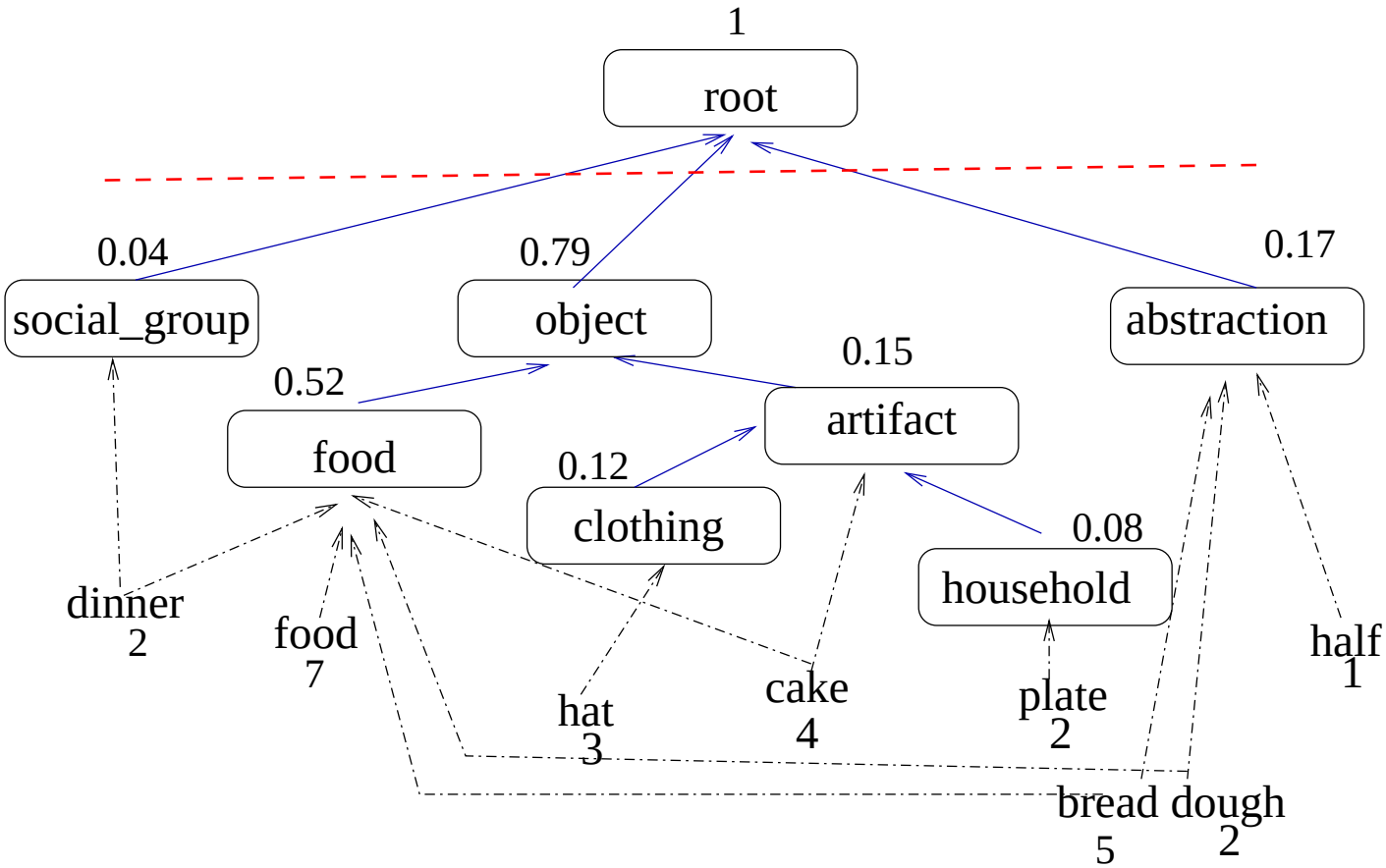
food 7, bread 5, cake 4, hat 3, dinner 2, dough 2, plate 2, half 1



noise from polysemous words, multiwords and other sources

WordNet Based Models: example *eat*

food 7, bread 5, cake 4, hat 3, dinner 2, dough 2, plate 2, half 1



Use frequency to find classes for representing preference and calculate probability distribution over these classes

Distributional Models

Find distributionally similar words:

bread: loaf 0.195, cheese 0.179, cake 0.169, potato 0.158, butter 0.155, meat 0.153, toast 0.148, flour 0.143, bean 0.139, vegetable 0.138

van: truck 0.230, lorry 0.229, car 0.222, vehicle 0.196, bus 0.191, taxi 0.172, train 0.160, tractor 0.150, boat 0.148, cab 0.147

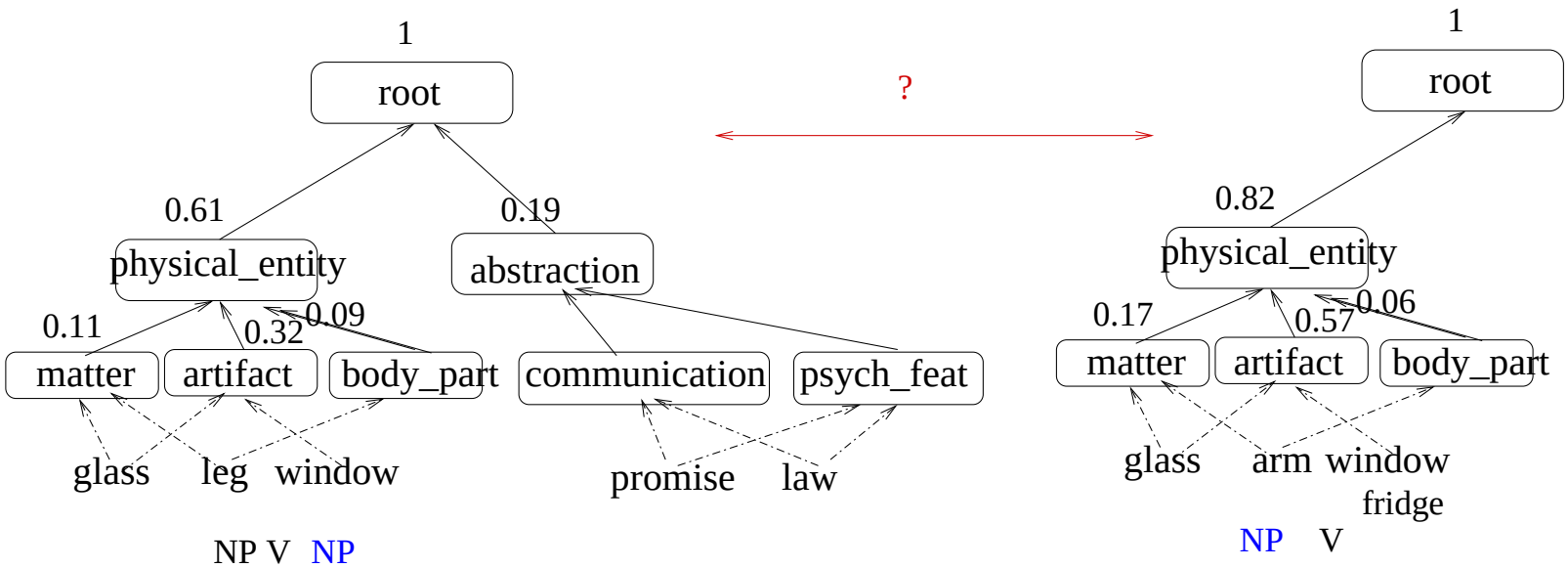
use these directly [Erk, 2007]

or build prototypical classes [McCarthy et al., 2007]

example: object slot of *park*

class ($p(c)$)	disambiguated objects (freq)
van (0.86)	car (174) van (11) vehicle (8) ...
backside (0.02)	backside (2) bum (1) butt (1) ...

Diathesis Alternation Detection: example *break*



Verb Class Acquistion

- decision trees using syntactic and semantic features [Merlo and Stevenson, 2001]
- clustering SCF [Schulte im Walde, 2006]
- clustering SCF and selectional preferences [Sun and Korhonen, 2009]

Verb Resources for Computational Linguistics

- WordNet [Fellbaum, 1998] 11529 verbs with verb senses and semantic relations
- Levin's classification: [Levin, 1993] 3100 verbs into 193 classes
- VerbNet [Kipper-Schuler, 2005] 3769 lemmas with 5257 senses, WordNet classes + syntactic info, roles and selectional restrictions
- ProbBank [Palmer et al., 2005] 1M WSJ corpus predicate-argument and verb specific role information
- Valex [Korhonen et al., 2006] automatically produced SCF lexicon of 6397 verbs
- FrameNet [Ruppenhofer et al., 2010] classifies verbs into approx 800 frames. Currently 135K examples (BNC and American newswire)

DANTE: Database of Analysed Texts of English

- 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

Outline

Background
- Lexical Information for Computational Linguistics
- Automatic Lexical Acquistion
- Existing Resources

Dante
- Lexical Information in Dante
- Demo
- A New Resource for Syntax-Semantics Research

DANTE (Database of ANalysed Texts of English)

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 .*

DANTE: for Research at the Syntax-Semantics Interface

Advantages:

- coverage
  - 6,300 head word verbs
  - 3000 phrasals
- 300,000 corpus examples for verbs (630,000 all PoS)
- varied corpus (1.7 billion words)
- syntax and semantics, attested in corpus and manually validated

DANTE: for Research at the Syntax-Semantics Interface

Potential:

- subcategorisation acquisition
- selectional preference acquisition
- word sense disambiguation and word sense induction
- diathesis alternation detection
- merging
  resources [Merlo and van der Plas, 2009, Atkins, 2010]

DANTE for Subcategorisation Frame Acquisition

As a gold standard

- examples as input data
- further examples of these DANTE SCF in sketch engine
- did system find the **same** frames in this data?
- did the system find erroneous frames?
- combine with verb sense

DANTE for Selectional Preference Acquisition

- reduce noise from parser stage,
- collect argument heads from examples
- collect further examples from DANTE corpus in sketch engine
- combine with SCF
- combine with verb sense (input data or detection)

DANTE for Word Sense Experiments

- word sense disambiguation
  inital experiments using:
  - collocates : match with context
  - SCF match with context: particularly promising for verbs
  - definitions : overlap with definitions of words in context
  - domain: overlap with domain of words in context
- word sense induction:
  - DANTE senses from contexts of examples
  - or classes based on SCF

DANTE for Diathesis Alternation Detection

- use SCF from same sense as input
- data at candidate slots from examples in sketch engine
- similarity of argument heads using distributional thesaurus
- use for verb class induction

Merge DANTE with Other Resources

- SCF (FrameNet, VerbNet)
- use distributional thesaurus to link collocates in different resources [Atkins, 2010]
- similarity of definitions, and collocates (WordNet)
- apply DANTE grammar to corpus (PropBank, FrameNet)
- disambiguate corpus by DANTE sense

Summary

- Verbal Lexical Acquisition for Comp. Ling.
  - SCF
  - selectional preferences
  - diathesis alternations
  - verb class

Summary

- Verbal Lexical Acquisition for Comp. Ling.
  - SCF
  - selectional preferences
  - diathesis alternations
  - verb class
- resources: VerbNet FrameNet PropBank ...
- DANTE: wide variety syntax and semantics phenomena associated with examples from 1.7 billion word corpus

Finally . . .

Finally . . .

COMPOUND

thank you

1 interj *[spok]* used to acknowledge sth sb has done for you/given you/said to you

- *' Are you cold ? ' ' No . ' ' Yes , you are , ' and he wrapped a warm rug around my legs and slung a coat behind my shoulders . '* **Thank you** *, ' I said .*
- *Normally we do n't like the training programme interrupted , but in view of the circumstances you may take a week 's compassionate leave . " "* **Thank you** *very much , sir . "*
- *"* **Thank you** *, thank you , thank you , " she said . " I 'll miss the game but I 'm ready for my new life . "*
- *What do you say?* **Thank you** *, Granny.*
- *.' I love your dress; it's fantastic."* **Thank you** *,' Aoife said.*
- *Is your back better? " "Improved,* **thank you** *. "*

CHUNK say thank you

- *It will be an opportunity for the Irish people to say* **thank you**.
- *They always said `* **Thank you** *' and smiled at me.*

CHUNK thank you for sth

- **Thank you** *for the grapes - presumably your own and the jam.*
- *David, a chara.* **Thank you***for your letter, which I received last Wednesday, and for the article which you enclosed.*

And Thanks to Colleagues on DANTE

- Sue Atkins
- Adam Kilgarriff
- Cathal Convery
- Michael Rundell
- Diana Rawlinson
- Valerie Grundy


Atkins, S. (2010).
The DANTE database: Its contribution to English lexical research, and in particular to complementing the FrameNet data.
In de Schryver, G.-M., editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Menha.


Brent, M. R. (1991).
Automatic acquisition of subcategorization frames from untagged text.
In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 209–214.


Briscoe, T. and Carroll, J. (1997).
Automatic extraction of subcategorization from corpora.
In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pages 356–363.


Erk, K. (2007).
A simple, similarity-based model for selectional preferences.
In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic. Association for Computational Linguistics.


Fellbaum, C., editor (1998).
WordNet, An Electronic Lexical Database.
The MIT Press, Cambridge, MA.


Kipper-Schuler, K. (2005).
VerbNet: A broad-coverage, comprehensive verb lexicon.
PhD thesis, Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.


Korhonen, A. (2002).
Subcategorization Acquisition.
PhD thesis, University of Cambridge.


Korhonen, A., Krymolowski, Y., and Briscoe, T. (2006).
A large subcategorization lexicon for natural language processing applications.
In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy.


Levin, B. (1993).
English Verb Classes and Alternations: a Preliminary Investigation.
University of Chicago Press, Chicago and London.


Li, H. and Abe, N. (1998).
Generalizing case frames using a thesaurus and the MDL principle.
Computational Linguistics, 24(2):217–244.


McCarthy, D. (2001).

Lexical Acquisition at the Syntax-Semantics Interface: diathesis alternations, subcategorization frames and selectional preferences.
PhD thesis, University of Sussex.


McCarthy, D., Venkatapathy, S., and Joshi, A. (2007).
Detecting compositionality of verb-object combinations using selectional preferences.
In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379.


Merlo, P. and Stevenson, S. (2001).
Automatic verb classification based on statistical distribution of argument structure.
Computational Linguistics, 27(3):373–408.


Merlo, P. and van der Plas, L. (2009).

Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both?
In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore. Association for Computational Linguistics.


Palmer, M., Gildea, D., and Kingsbury, P. (2005).
The proposition bank: A corpus annotated with semantic roles.
Computational Linguistics, 31(1):71–106.


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/.


Schulte im Walde, S. (2006).
Experiments on the automatic induction of german semantic verb classes.
Computational Linguistics, 32(2):159–194.


Sun, L. and Korhonen, A. (2009).
Improving verb clustering with automatically acquired selectional preferences.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Singapore. Association for Computational Linguistics.