Distributional semantics for linguists

Distributional semantics

Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Distributional semantics

Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Distributional semantics

Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Some history

- Early discussion: Osgood (1952), Zelig Harris (1954).
- Firth (1957): 'You shall know a word by the company it keeps'.
- 'distributional semantics' by 1960s: e.g., Garvin (1962).
- Spärck Jones (1964): PhD thesis 'Synonymy and Semantic Classification' (dictionaries for context).
- First experiments on sentential contexts: Harper (1965) inspired by Harris; Spärck Jones (1967).
- Grefenstette (1994), Schütze (1998); Landauer and Dumais (1997) 'Latent Semantic Analysis' (LSA).
- Huge proliferation of papers in computational linguistics (CL) once corpora (and large scale parsing) become available.

Vector representations and clustering

Words represented as vectors of features:

	feature_1	feature_2	...	feature_n
word_1	$f_{1,1}$	$f_{2,1}$		$f_{n,1}$
word_2	$f_{1,2}$	$f_{2,2}$		$f_{n,2}$
...				
word_m	$f_{1,m}$	$f_{2,m}$		$f_{n,m}$

Features: co-occur with word_n in some window, co-occur with word_n as a syntactic dependent, occur in paragraph_n, occur in document_n ...

First computational application: Spärck Jones (1964)

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with ‘similar’ vectors.

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with 'similar' vectors.

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with ‘similar’ vectors.

Early clustering

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.

Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.

Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Early clustering

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.
Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.

Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Early clustering

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.
Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.
Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Spärck Jones (1967)

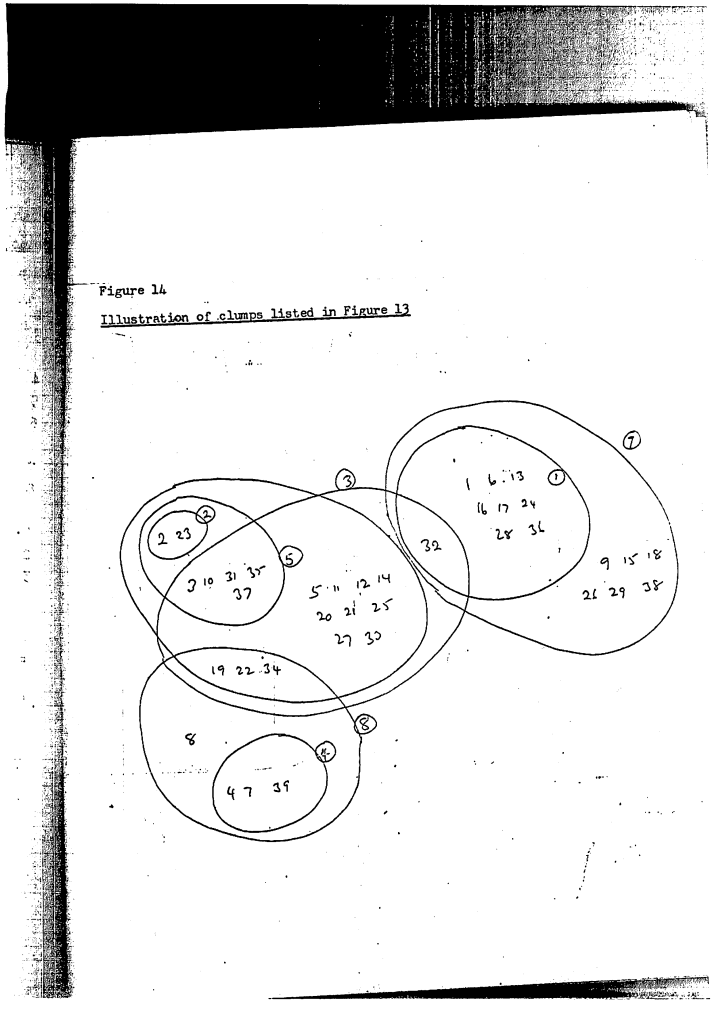
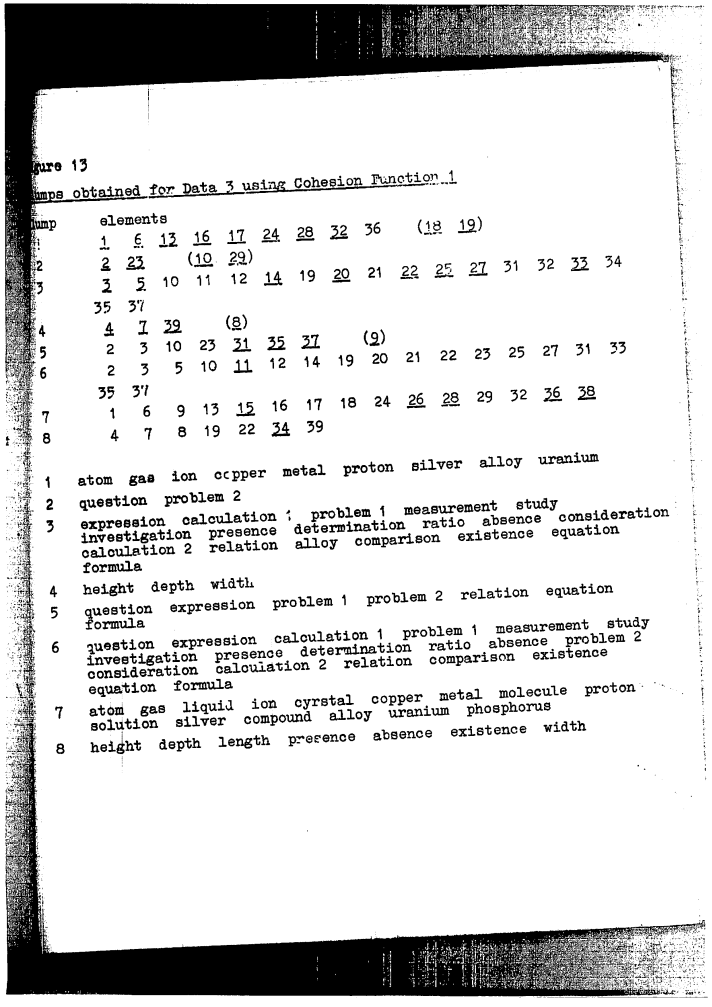


Figure 14

Illustration of clumps listed in Figure 13



ure 13

mps obtained for Data 3 using Cohesion Function 1

lump	elements
1	1 6 13 16 17 24 28 32 36 (18 19)
2	2 23 (10 29)
3	3 5 10 11 12 14 19 20 21 22 25 27 31 32 33 34 35 37
4	4 7 39 (8)
5	2 3 10 23 31 35 37 (9)
6	2 3 5 10 11 12 14 19 20 21 22 23 25 27 31 33 35 37
7	1 6 9 13 15 16 17 18 24 26 28 29 32 36 38
8	4 7 8 19 22 34 39

1 atom gas ion copper metal proton silver alloy uranium

2 question problem 2

3 expression calculation 1 problem 1 measurement study investigation presence determination ratio absence consideration calculation 2 relation alloy comparison existence equation formula

4 height depth width

5 question expression problem 1 problem 2 relation equation formula

6 question expression calculation 1 problem 1 measurement study investigation presence determination ratio absence problem 2 consideration calculation 2 relation comparison existence equation formula

7 atom gas liquid ion cyrstal copper metal molecule proton solution silver compound alloy uranium phosphorus

8 height depth length presence absence existence width

CS history and distributional semantics

- Early distributional work not followed up:
  - limitations of computers and available corpora.
  - 1966 ALPAC report led to diminished funding for CL.
  - “It must be recognized that the notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.” (Chomsky 1969)
  - KSJ and others switched to Information Retrieval: KSJ (inspired by classification experiments) and Robertson develop tf*idf measure.
- Early 1990s: influence from IR: large corpora, computer memory, disk space make simple distributional techniques practical.
- Early 2000s: large scale, robust parsing makes more complex notions of context practical.

Characteristic contexts: beer

0.484118::can_n+of_p()	0.323999::and_c+drink_n
0.470041::and_c+wine_n	0.323292::alcoholic_a
0.451887::brand_n+of_p()	0.321707::tear_n+in_p()
0.444771::pron_rel_+drink_v	0.321004::and_c+brewery_n
0.407286::wine_n+and_c	0.31969::and_c+beverage_n
0.403163::duff_a	0.317467::bread_n+and_c
0.392823::and_c+cigarette_n	0.315654::recipe_n+for_p()
0.388944::liter_n+of_p()	0.312405::premium_a
0.38283::sweat_n+and_c	0.306168::rye_a
0.364612::wheat_a	0.30428::have_v+taste_n
0.341821::seasonal_a	0.301791::lite_a
0.3409::in_p()+Hell_n	0.300422::in_p()+glass_n
0.333707::or_c+spirit_n	0.299759::style_n+of_p()
0.325886::for_p()+horse_n	0.297687::stale_a
0.324157::drink_n+and_c	0.297159::be_v+drink_n

Characteristic contexts: ?

0.532551::and_c+Perry_n	0.224517::homemade_a
0.475489::sparkle_v	0.217018::ferment_v
0.462226::beer_n+and_c	0.215903::pron_rel_+drink_v
0.324184::be_v+drink_n	0.215738::and_c+wine_n
0.313665::alcoholic_a	0.212648::in_p()+Denmark_n
0.295653::hard_a	0.199628::fruit_n+and_c
0.272322::brand_n+of_p()	0.183856::eat_v+and_c
0.268747::wine_n+and_c	0.18323::and_c+apple_n
0.264604::for_p()+star_n	0.183142::and_c+grape_n
0.256199::in_p()+branch_n	0.182793::from_p()+Wales_n
0.255403::and_c+beer_n	0.182706::have_v+density_n
0.246708::liter_n+of_p()	0.180874::to_p()+productio
0.243786::and_c+spice_n	0.180084::in_p()+layer_n
0.241399::cloudy_a	0.178431::hazy_a
0.239619::gallon_n+of_p()	0.178213::Tech_n+and_c

Psycholinguistics

- Latent Semantic Analysis (LSA) popular as a technique for investigating lexical semantics.
- Neural basis of word meaning: functional web of neurons associated with a lexeme connects recognizers, semantics and articulators (e.g. Pulvermüller 2002).
- Hebbian learning principle: paraphrased as “Neurons that fire together wire together”.
- Under these assumptions: if two lexemes co-occur frequently this would necessarily lead to strong associations between their functional webs.

Assumptions about lexical semantics

1. Limited (if any) role for semantic primitives (*kill* not CAUSE(x (DIE(y))) or similar).
2. No hard boundary between linguistic knowledge and world knowledge.
3. Acquisition must be considered.
4. Word meaning is fuzzy, speakers negotiate meaning.
5. Senses (other than homonyms) are not discrete.

Why ‘Distributional semantics for linguists’?

- Part of an approach to meaning representation?
- More modestly:
  - Semantic classification for investigation of syntax-semantic interface.
  - Investigative tool for sociolinguists etc.
- Practicalities: free/cheap corpora and ordinary computer hardware are now fully adequate for most experiments.

The general intuition

- **Distributions** are vectors in a multidimensional semantic space, that is, objects with a magnitude (length) and a direction.
- The **semantic space** has dimensions which correspond to possible contexts.
- For our purposes, a distribution can be seen as a point in that space (the vector being defined with respect to the origin of that space).
- *cat* [...dog 0.8, eat 0.7, joke 0.01, mansion 0.2, zebra 0.1...]

The notion of context

- **Context:** if the meaning of a word is given by its context, what does 'context' mean?
  - Word windows (unfiltered): *n* words on either side of the lexical item under consideration (unparsed text).
    Example: n=2 (5 words window):

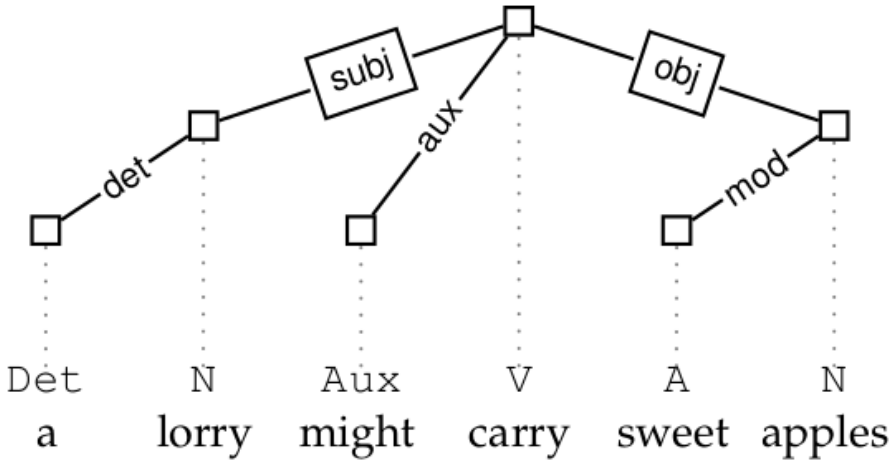
    *... the prime **minister** acknowledged that ...*

  - Word windows (filtered): *n* words on either side of the lexical item under consideration (unparsed text). Some words are not considered part of the context (e.g. function words, some very frequent content words). The stop list for function words is either constructed manually, or the corpus is POS-tagged.
    Example: n=2 (5 words window):

    *... the prime **minister** acknowledged that ...*

The notion of context

- Dependencies: syntactic or semantic. The corpus is converted into a list of directed links between heads and dependents. Context for a lexical item is the dependency structure it belongs to. The length of the dependency path can vary according to the implementation (Padó and Lapata, 2007).




```
a        [Det,det,N]   lorry
lorry    [N,subj,V]    carry
might    [Aux,aux,V]   carry
apples   [N,obj,V]     carry
sweet    [A,mod,N]     apples
```

Parsed vs unparsed data: examples

word (unparsed)	**word (parsed)**
meaning_n	or_c+phrase_n
derive_v	and_c+phrase_n
dictionary_n	syllable_n+of_p
pronounce_v	play_n+on_p
phrase_n	etymology_n+of_p
latin_j	portmanteau_n+of_p
ipa_n	and_c+deed_n
verb_n	meaning_n+of_p
mean_v	from_p+language_n
hebrew_n	pron_rel_+utter_v
usage_n	for_p+word_n
literally_r	in_p+sentence_n

Context weighting

- Binary model: if context *c* co-occurs with word *w*, value of vector $\vec{w}$ for dimension *c* is 1, 0 otherwise.

  ... [a long long long ***example*** *for a distributional semantics] model... (n=4)*

  ... {a 1} {dog 0} {long 1} {sell 0} {semantics 1}...

- Basic frequency model: the value of vector $\vec{w}$ for dimension *c* is the number of times that *c* co-occurs with *w*.

  ... [a long long long ***example*** *for a distributional semantics] model... (n=4)*

  ... {a 2} {dog 0} {long 3} {sell 0} {semantics 1}...

Context weighting

- Characteric model: the weights given to the vector components express how *characteristic* a given context is for *w*. Functions used include:
  - Pointwise Mutual Information (PMI), with or without discounting factor.

$$pmi_{wc} = log(\frac{f_{wc} * f_{total}}{f_w * f_c}) \quad (1)$$

  - Derivatives such as Mitchell and Lapata's (2010) weighting function (PMI without the log).

What semantic space?

- Entire vocabulary.
  - + All information included – even rare, but important contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png|thumb|right|200px|graph_n*)
- Top *n* words with highest frequencies.
  - + More efficient (5000-10000 dimensions). Only 'real' words included.
  - - May miss out on infrequent but relevant contexts.

What semantic space?

- Singular Value Decomposition (LSA – Landauer and Dumais, 1997): the number of dimensions is reduced by exploiting redundancies in the data. A new dimension might correspond to a generalisation over several of the original dimensions (e.g. the dimensions for *car* and *vehicle* are collapsed into one).
  - + Very efficient (200-500 dimensions). Captures generalisations in the data.
  - - SVD matrices are not interpretable.
- Other, more esoteric variants...

Our reference text

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Example:** Produce distributions using a word window, frequency-based model

The semantic space

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- We assume that we only keep content words in the semantic space.
- **Dimensions:**

difference	impossible	thing
get	major	turns
go	possibly	usually
goes	repair	wrong

Frequency counts...

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Counts:**

difference 1	impossible 1	thing 3
get 1	major 1	turns 1
go 3	possibly 2	usually 1
goes 1	repair 1	wrong 4

Conversion into 5-word windows...

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- ∅ ∅ **the** major difference
- ∅ the **major** difference between
- the major **difference** between a
- major difference **between** a thing
- ...

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

The major difference between a thing that [might go wrong and a] thing that cannot [possibly go wrong is that] when a thing that cannot [possibly go [wrong goes wrong] it usually] turns out to be impossible to get at or repair.

- **Distribution (frequencies):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 1	possibly 1	usually 1
goes 2	repair 0	wrong 2

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

The major difference between a thing that [might go wrong and a] thing that cannot [possibly go wrong is that] when a thing that cannot [possibly go [wrong goes wrong] it usually] turns out to be impossible to get at or repair.

- **Distribution (PMIs):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 0.22184875	possibly 0.397940009	usually 0.698970004
goes 1	repair 0	wrong 0.397940009

Corpus description

- Obtained from the entire English Wikipedia.
- Corpus parsed with the English Resource Grammar (Flickinger, 2000) and converted into DMRS form (Copestake, 2009).
- Dependencies considered include:
  - For nouns: head verbs (+ any other argument of the verb), modifying adjectives, head prepositions (+ any other argument of the preposition).
    e.g. cat: chase_v+mouse_n, black_a, of_p+neighbour_n
  - For verbs: arguments (NPs and PPs), adverbial modifiers.
    e.g. eat: cat_n+mouse_n, in_p+kitchen_n, fast_a
  - For adjectives: modified nouns; rest as for nouns (assuming intersective composition).
    e.g. black: cat_n, chase_v+mouse_n

System description

- Semantic space: top 100,000 contexts.
- Weighting: normalised PMI (Bouma 2007).

$$pmi_{wc} = \frac{log(\frac{f_{wc} * f_{total}}{f_w * f_c})}{-log(\frac{f_{wc}}{f_{total}})} \quad (2)$$

An example noun

- *language*:

0.541816::other+than_p()+English_n
0.525895::English_n+as_p()
0.523398::English_n+be_v
0.48977::english_a
0.481964::and_c+literature_n
0.476664::people_n+speak_v
0.468399::French_n+be_v
0.463604::Spanish_n+be_v
0.463591::and_c+dialects_n
0.452107::grammar_n+of_p()
0.445994::foreign_a
0.445071::germanic_a
0.439558::German_n+be_v
0.436135::of_p()+instruction_n
0.435633::speaker_n+of_p()
0.423595::generic_entity_rel_+speak_v
0.42313::pron_rel_+speak_v
0.42294::colon_v+English_n
0.419646::be_v+English_n
0.418535::language_n+be_v
0.4159::and_c+culture_n
0.410987::arabic_a
0.408387::dialects_n+of_p()
0.399266::part_of_rel_+speak_v
0.397::percent_n+speak_v
0.39328::spanish_a
0.39273::welsh_a
0.391575::tonal_a

An example adjective

- *academic*:

0.517031::Decathlon_n
0.512661::excellence_n
0.449711::dishonesty_n
0.445393::rigor_n
0.426142::achievement_n
0.421246::discipline_n
0.397311::vice_president_n+for_p()
0.391978::institution_n
0.38937::credentials_n
0.378062::journal_n
0.373727::journal_n+be_v
0.372052::vocational_a
0.371873::student_n+achieve_v
0.361359::athletic_a
0.356562::reputation_n+for_p()
0.354674::regalia_n
0.353712::program_n
0.351601::freedom_n
0.347751::student_n+with_p()
0.34621::curriculum_n
0.342008::standard_n
0.34151::at_p()+institution_n
0.340271::career_n
0.337857::Career_n
0.329923::dress_n
0.329358::scholarship_n
0.329281::prepare_v+student_n
0.328009::qualification_n

Corpus choice

- As much data as possible?
  - British National Corpus (BNC): 100 m words
  - Wikipedia: 897 m words
  - UKWac: 2 bn words
  - ...
- In general preferable, *but*:
  - More data is not necessarily the data you want.
  - More data is not necessarily realistic from a psycholinguistic point of view. We perhaps encounter 50,000 words a day. BNC = 5 years' text exposure.

Corpus choice

- Distribution for *unicycle*, as obtained from Wikipedia.

0.448051::motorized_a
0.404372::pron_rel_+ride_v
0.238612::for_p()+entertainment_n
0.235763::half_n+be_v
0.235407::unwieldy_a
0.230275::earn_v+point_n
0.216627::pron_rel_+crash_v
0.190785::man_n+on_p()
0.186325::on_p()+stage_n
0.185063::position_n+on_p()
0.168102::slip_v
0.162611::and_c+1_n
0.159627::autonomous_a
0.155822::balance_v
0.133084::tall_a
0.124242::fast_a
0.106976::red_a
0.0714643::come_v
0.0601987::high_a

Polysemy

- Distribution for *pot*, as obtained from Wikipedia.

0.566454::melt_v
0.442374::pron_rel_+smoke_v
0.434682::of_p()+gold_n
0.40773::porous_a
0.401654::of_p()+tea_n
0.39444::player_n+win_v
0.393812::money_n+in_p()
0.376198::of_p()+coffee_n
0.33117::amount_n+in_p()
0.329211::ceramic_a
0.326387::hot_a
0.323321::boil_v
0.313404::bowl_n+and_c
0.306324::ingredient_n+in_p()
0.301916::plant_n+in_p()
0.298764::simmer_v
0.292397::pot_n+and_c
0.284539::bottom_n+of_p()
0.28338::of_p()+flower_n
0.279412::of_p()+water_n
0.278914::food_n+in_p()
0.262501::pron_rel_+heat_v
0.260375::size_n+of_p()
0.25511::pron_rel_+split_v
0.254363::of_p()+money_n
0.2535::of_p()+culture_n
0.249626::player_n+take_v
0.246479::in_p()+hole_n
0.244051::of_p()+soil_n
0.243797::city_n+become_v

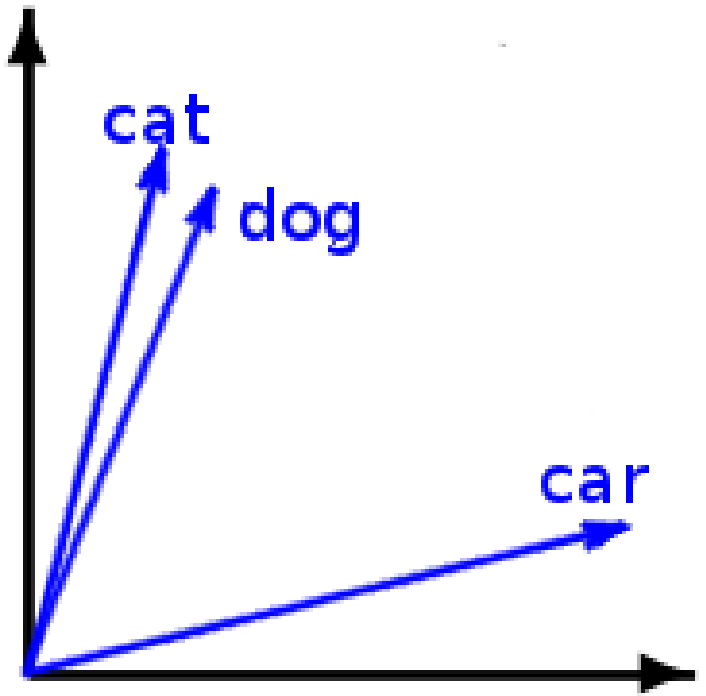
Fixed expressions

- Distribution for *time*, as obtained from Wikipedia.

0.462949::of_p()+death_n
0.448965::same_a
0.446277::1_n+at_p(temp)
0.445338::Nick_n+of_p()
0.423542::spare_a
0.418568::playoffs_n+for_p()
0.416471::of_p()+retirement_n
0.405288::of_p()+release_n
0.397135::pron_rel_+spend_v
0.389886::sand_n+of_p()
0.385954::pron_rel_+waste_v
0.382816::place_n+around_p()
0.37777::of_p()+arrival_n
0.376466::of_p()+completion_n
0.374797::after_p()+time_n
0.374682::of_p()+arrest_n
0.371589::country_n+at_p()
0.370736::age_n+at_p()
0.370626::space_n+and_c
0.370555::in_p()+career_n
0.370464::world_n+at_p()
0.363982::and_c+space_n
0.363241::generic_entity_rel_+mark_v
0.361872::of_p()+introduction_n
0.357929::in_p()+year_n
0.357565::of_p()+appointment_n
0.356229::of_p()+trouble_n
0.355658::of_p()+merger_n
0.354794::on_p()+ice_n
0.353891::practice_n+at_p()
0.351994::of_p()+birth_n
0.351556::full_a
0.348029::of_p()+accident_n
0.34785::state_n+at_p()
0.347753::to_p()+time_n
0.345147::of_p()+election_n
0.345088::area_n+at_p()
0.342571::and_c+money_n
0.342113::time_n+after_p()
0.341877::allotted_a

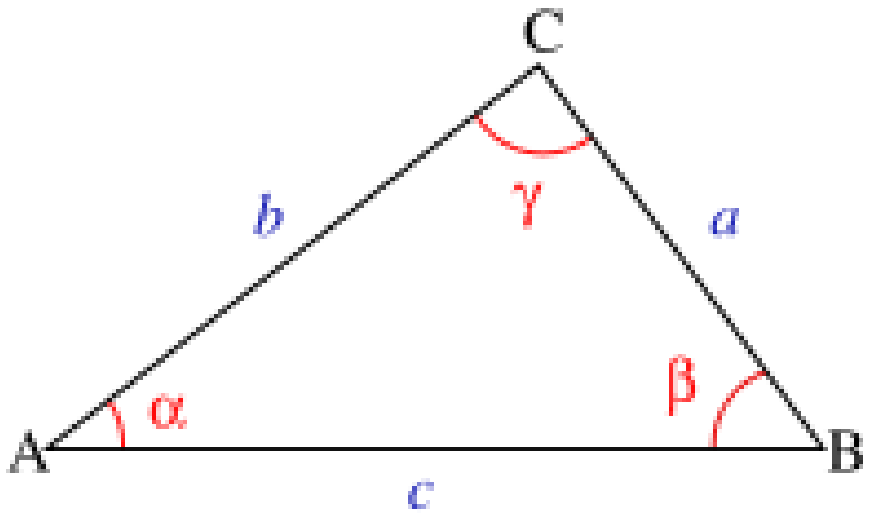
Calculating similarity in a distributional space

- Distributions are vectors, so distance can be calculated.



Some trigonometry

- Law of cosines: $c^2 = a^2 + b^2 - 2ab\cos\gamma$



Measuring similarity

- Cosine:

$$\frac{\sum v1_k * v2_k}{\sqrt{\sum v1_k^2} * \sqrt{\sum v2_k^2}} \quad (1)$$

- The cosine measure calculates the angle between two vectors and is therefore length-independent. This is important, as frequent words have longer vectors than less frequent ones.
- Other measures include Jaccard, Lin... (For an overview: see Weeds, 2004).

Some numbers

- The scale of similarity...
  - house – building 0.428354
  - gem – jewel 0.306866
  - capitalism – communism 0.294677
  - motorcycle – bike 0.29329
  - test – exam 0.269151
  - school – student 0.250291
  - singer – academic 0.168105
  - horse – farm 0.133888
  - man –accident 0.0885102
  - tree – auction 0.0234772
  - cat –county 0.00731196

Example

- Words most similar to *cat*, as chosen from the 5000 most frequent nouns in Wikipedia.

1 cat	0.276042 man	0.230953 person
0.4512 dog	0.275582 cow	0.229124 pet
0.357814 animal	0.264269 fox	0.228973 lizard
0.336883 rat	0.260912 girl	0.228406 chicken
0.331284 rabbit	0.26071 sheep	0.223872 monster
0.329772 pig	0.258142 boy	0.218094 people
0.309073 monkey	0.255272 elephant	0.216812 tiger
0.307839 bird	0.248803 deer	0.215497 mammal
0.302241 horse	0.247423 woman	0.212786 bat
0.296586 mouse	0.245761 fish	0.2122 duck
0.292734 wolf	0.243787 squirrel	0.209441 cattle
0.292047 creature	0.243725 dragon	0.208839 dinosaur
0.287286 human	0.243714 frog	0.207969 character
0.286601 goat	0.234795 baby	0.207257 kid
0.282235 snake	0.233694 child	0.206511 turtle
0.279406 bear	0.231072 lion	0.2049 robot

But what is similarity?

- In distributional semantics, very broad notion. Includes synonyms, near-synonyms, hyponyms, taxonomical siblings, antonyms, etc.
- The broad notion does correlate with a psychological reality. One of the favourite tests of the distributional semantics community is the calculation of correlation between a distributional similarity system and human judgments on the Miller & Charles (1991) test set.

Miller & Charles 1991

3.92 automobile-car	3.05 bird-cock	0.84 forest-graveyard
3.84 journey-voyage	2.97 bird-crane	0.55 monk-slave
3.84 gem-jewel	2.95 implement-tool	0.42 lad-wizard
3.76 boy-lad	2.82 brother-monk	0.42 coast-forest
3.7 coast-shore	1.68 crane-implement	0.13 cord-smile
3.61 asylum-madhouse	1.66 brother-lad	0.11 glass-magician
3.5 magician-wizard	1.16 car-journey	0.08 rooster-voyage
3.42 midday-noon	1.1 monk-oracle	0.08 noon-string
3.11 furnace-stove	0.89 food-rooster	
3.08 food-fruit	0.87 coast-hill	

- Miller & Charles experiment: re-run of Rubenstein & Goodenough (1965). Correlation coefficient = 0.97.

Distributional methods are discursive

- Distributions are a good conceptual representation if you believe that ‘the meaning of a word is given by its usage’.
- Corpus-dependent, culture-dependent, register-dependent. Example: similarity between *policeman* and *cop*: 0.232632.

Distributions are register-dependent

policeman
0.586482::ball_n+poss_rel
0.47911::and_c+civilian_n
0.424271::soldier_n+and_c
0.409217::and_c+soldier_n
0.384081::secret_a
0.370919::people_n+include_v
0.36834::corrupt_a
0.358544::uniformed_a
0.352538::uniform_n+poss_rel
0.349553::civilian_n+and_c
0.315058::iraqi_a
0.311442::lot_n+poss_rel
0.307535::chechen_a
0.303514::laugh_v
0.286281::and_c+criminal_n
0.285162::incompetent_a
0.284202::pron_rel_+shoot_v
0.279526::hat_n+poss_rel
0.276776::terrorist_n+and_c
0.272654::and_c+crowd_n
0.271465::military_a

cop
0.450031::crooked_a
0.448631::corrupt_a
0.439307::maniac_a
0.380065::dirty_a
0.373174::honest_a
0.357623::uniformed_a
0.350859::tough_a
0.327847::pron_rel_+call_v
0.320139::funky_a
0.317952::bad_a
0.29243::veteran_a
0.290737::and_c+robot_n
0.285521::and_c+criminal_n
0.279318::bogus_a
0.276689::talk_v+to_p()+pron_rel_
0.272944::investigate_v+murder_n
0.257574::on_p()+force_n
0.251643::parody_n+of_p()
0.249137::Mason_n+and_c
0.246172::pron_rel_+kill_v
0.246089::racist_a

Extension

- In set-theoretic semantics, the meaning or **extension** of *cat*, cat′, is the set of all cats in some world.
- Sets intersect, so the meaning of *black cat* is $\text{cat}'(x) \wedge \text{black}'(x)$, the intersection of the set of cats and the set of black things.
- Some entities will be in several sets.

The classical account

- Difference between full synonymy (*eggplant/aubergine*) and near-synonymy (*city, town*).
- The extensions of two full synonyms are identical sets. $\text{eggplant}' = \text{aubergine}'$
- The extensions of two near synonyms have a high (whatever that means...) overlap. i.e. with respect to a specific context, near-synonyms will often be substitutable.

Some facts about synonymy

- Near-synonymy is frequent, absolute synonymy relates to dialect etc. (*eggplant/aubergine*)
- Word sense assumptions affect synonymy assumptions.
- Language learners tend to assume non-synonymy.
  e.g., “labeling entities with distinct words leads infants to create representations of two distinct individuals” (Carey, 2009:p 277)

Near-synonymy and meaning acquisition

- Readers only need a few uses to obtain a working idea of a new word's meaning. (Rice, 1990)
- Hypothesis: understanding a new word (without definition) can be modelled by two-phase comparison:
  - initial approximation: e.g., *rancid* is similar to *off*
  - acquisition of differentiating information **characteristic contexts**: e.g., *rancid* tends to appear with fatty foods (or dairy foods, or . . .)
- People's beliefs about low-to-medium frequency words may differ but approximation is usually good enough for communication.

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.

BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Full synonymy and meaning acquisition

- Full synonyms are probably acquired differently from near-synonyms, generally by (relatively) explicit definition:

  The aubergine (eggplant) has to be one of my favourite vegetables.

- Full synonyms may be different vocalisations for the same concept (their lexemes share a single semantic functional web in the brain).
- Contrast with near-synonyms which are separate concepts.

Frequency and synonymy

- Speakers use the most frequent term in their experience to convey a particular idea (frequency assumed to correlate with strength of neural connections).
- More frequent words tend to have broader meanings (more 'senses' . . .)
- Two words of very different frequency are unlikely to cover exactly the same semantic space.
- Many words are of too low frequency for hearers to make reliable decisions about synonymy.

Synonymy: requisites for an ideal distributional account

- Distinguishing between near-synonyms and full synonyms.
- No hard line between near-synonyms and non-synonyms.
- Degree of synonymy between two lexemes will vary between individuals.

The distribution of synonyms

- Similarity between *egglant/aubergine*: 0.114024
  Relatively low cosine. Partly due to frequency (222 for *eggplant*, 56 for *aubergine*).
- Similarity between *policeman/cop*: 0.232632
  To be expected: *policeman* and *cop* are discursively very different.
- Similarity between *city/town*: 0.735319
- So... similarity does not tell us how to distinguish between full and near-synonymy.

The classical account

- Three basic types of antonymy:
  - gradable (opposite ends of a scale: *cold/hot*, modifiable with *very*, etc)
  - non-gradable (discrete opposition: *dead/alive*)
  - multiple (non-gradable, discontinuous scale: *lecturer, reader, professor*).
- In terms of extension: the same entity cannot be described as both X and its antonym Y in a given situation. i.e. for a micro-world corresponding to a situation where I drink tea, the tea cannot be in the set of cold things and in the set of hot things.

Distributions of antonyms

- Similarities between:
  - cold/hot 0.287398
  - dead/alive 0.242078
  - large/small 0.6783
  - colonel/general 0.333739

Identifying antonyms

- Antonyms have a high distributional similarity. It is hard to distinguish them from near-synonyms.
- The identification of antonyms usually requires some heuristics to be applied to pairs of highly similar distributions.
- For instance, it has been observed that antonyms are frequently coordinated while synonyms are not:
  - a selection of cold and hot drinks
  - wanted dead or alive
  - lectures, readers and professors are invited to attend

The classical account

- Relationship between a more general term and a more specific term (*dog/poodle*).
- The extension of the more general includes the extension of the more specific (all poodles are dogs).
- The intension of the more specific includes the intension of the more general (all that can be said about dogs can be said about poodles)... in an essentialist account (see penguins).

Distributions of hyponyms

- No clear inclusion relationship. The set of contexts recorded for *cat* and *animal* overlap, but they are by no means subsets.
- Kotlerman et al (2010), however, demonstrated that in general, if X is a hyponym of Y, features with high values in X tend to have a high value in Y.
- Baroni et al (2012) learn hyponymy from distributions for adjective-noun phrases (a black cat is a cat). But they do not report on the features used by the classifier.
- Similarity between *cat* and *animal*: 0.357814.

Issues

- There is no formal definition for the standard lexical relations in distributional semantics.
- The standard definitions rely on the idea of extension, but there is no obvious correspondence between the corpora used to produce distributions and the real world.

Polysemy

- Distribution for *pot*, obtained from Wikipedia.

0.566454::melt_v
0.442374::pron_rel_+smoke_v
0.434682::of_p()+gold_n
0.40773::porous_a
0.401654::of_p()+tea_n
0.39444::player_n+win_v
0.393812::money_n+in_p()
0.376198::of_p()+coffee_n
0.33117::amount_n+in_p()
0.329211::ceramic_a
0.326387::hot_a
0.323321::boil_v
0.313404::bowl_n+and_c
0.306324::ingredient_n+in_p()
0.301916::plant_n+in_p()
0.298764::simmer_v
0.292397::pot_n+and_c
0.284539::bottom_n+of_p()
0.28338::of_p()+flower_n
0.279412::of_p()+water_n
0.278914::food_n+in_p()
0.262501::pron_rel_+heat_v
0.260375::size_n+of_p()
0.25511::pron_rel_+split_v
0.254363::of_p()+money_n
0.2535::of_p()+culture_n
0.249626::player_n+take_v
0.246479::in_p()+hole_n
0.244051::of_p()+soil_n
0.243797::city_n+become_v

Polysemy

- Distribution for *drug*, obtained from Wikipedia.

0.608869::and_c+alcohol_n
0.510397::alcohol_n+and_c
0.464624::or_c+substance_n
0.462777::alcohol_n+or_c
0.451267::over-the-counter_a
0.451249::inflammatory_a
0.448604::food_n+and_c
0.445496::addictive_a
0.428868::and_c+prostitution_n
0.42017::illegal_a
0.41921::recreational_a
0.417316::have_v+side_effect_n
0.408879::like_p()+Me_n
0.402512::side_effect_n+of_p()
0.400139::intravenous_a
0.397089::of_p()+abuse_n
0.39542::war_n+on_p()
0.393311::dose_n+of_p()
0.386679::metabolism_n+of_p()
0.369514::and_c+crime_n
0.36857::effect_n+poss_rel
0.366681::of_p()+choice_n
0.365335::and_c+substance_n
0.364455::drug_n+be_v
0.360401::anti-_a
0.359099::generic_a
0.358552::overdose_n+of_p()
0.358029::treatment_n+with_p()
0.35767::prostitution_n+and_c
0.35661::diabetic_a

Polysemy

- Distribution for *soft*, obtained from Wikipedia.

0.624533::plump_a
0.624433::drink_n
0.609981::plumage_n
0.588074::fluffy_a
0.547627::uneven_a
0.540281::silky_a
0.51885::palate_n
0.50562::tissue_n
0.477878::spine_n+and_c
0.453215::colourful_a
0.444027::hand-off_n
0.413344::pretzel_n
0.40609::call_n+be_v
0.388752::Cell_n
0.387858::feather_n
0.387565::and_c+tail_n
0.379231::become_v+and_c
0.377516::paste_n
0.373097::ray_n
0.372154::spot_n
0.367734::coral_n
0.362632::dorsal_a
0.361666::reboot_n
0.359202::acidic_a
0.358819::texture_n
0.358372::and_c+snack_n
0.352847::beer_n+and_c
0.348029::erosion_n+of_p()
0.346968::fleshy_a
0.344807::porn_n

Sense induction

Normally, single point in vector space represents all uses.

- Sense induction: cluster contexts and associate new instances with a cluster (contrast word sense disambiguation, where prior list of word senses).
- Different senses for each word (contrast topic clustering, where words are associated with a global set of topics).
- Early work by Neill (2002): automatically discovers 'seed' words which discriminate between clusters.
- Clusters are more discrete for homonyms compared to general polysemy: some uses in between senses?
- Current applications tend not to distinguish senses.
- More on Thursday on regular polysemy.

Multiword expressions (MWEs)

- 'words with spaces': e.g., *ad hoc* (in English!)
- non-decomposable: e.g., *kick the bucket*
- decomposable but non-compositional: e.g., *cat out of the bag* (meaning 'secret out of hiding place')
- idioms of encoding/collocations: e.g., *heavy shower*

MWEs and distributions:

- MWEs might be expected to obscure distributional meaning.
- But: ranking of contexts by PMI very similar to techniques for finding MWEs!
- and higher associations suggest lower compositionality.

Magnitude adjectives and non-physical-solid nouns. (Copestake, 2005)

Distributional data from the British National Corpus (100 million words)

	importance	success	majority	number	proportion	quality	role	problem	part	winds	support	rain
great	310	360	382	172	9	11	3	44	71	0	22	0
large	1	1	112	1790	404	0	13	10	533	0	1	0
high	8	0	0	92	501	799	1	0	3	90	2	0
major	62	60	0	0	7	0	272	356	408	1	8	0
big	0	40	5	11	1	0	3	79	79	3	1	1
strong	0	0	2	0	0	1	8	0	3	132	147	0
heavy	0	0	1	0	0	1	0	0	1	2	4	198

Adjectives: selected examples.

BNC frequencies:

	number	proportion	quality	problem	part	winds	rain
large	1790	404	0	10	533	0	0
high	92	501	799	0	3	90	0
big	11	1	0	79	79	3	1
heavy	0	0	1	0	1	2	198

Acceptability judgements:

	number	proportion	quality	problem	part	winds	rain
large			*			*	*
high				*	?		*
big			?				*
heavy	?	*	*	*			

Magnitude adjective distribution.

- Investigated the distribution of *heavy, high, big, large, strong, great, major* with the most common co-occurring nouns in the BNC.
- Nouns tend to occur with up to three of these adjectives with high frequency and low or zero frequency with the rest.
- 50 nouns in BNC with the extended use of *heavy* with frequency 10 or more, 160 such nouns with *high*. Only 9 with both: *price, pressure, investment, demand, rainfall, cost, costs, concentration, taxation*
- Clusters: e.g., weather precipitation nouns with *heavy*. Note *heavy shower* (weather, not bathroom).

Hypotheses about distribution.

- 'abstract' *heavy, high, big, large, strong, great, major* all denote magnitude (in a way that can be made formally precise)
- distribution differences due to collocation, soft rather than hard constraints
- adjective-noun combination is semi-productive
- denotation and syntax allow *heavy esteem* etc, but speakers are sensitive to frequencies, prefer more frequent phrases with 'same' meaning

Adjective similarities

	high	heavy	big	large	strong	major
high	-	-	-	-	-	-
heavy	0.22	-	-	-	-	-
big	0.26	0.22	-	-	-	-
large	0.40	0.30	0.45	-	-	-
strong	0.30	0.29	0.30	0.34	-	-
major	0.31	0.20	0.44	0.45	0.32	-

Applications of distributional semantics

- Many applications in natural language processing: e.g., improving search, processing scientific text, sentiment analysis.
- Also applications in philosophy and sociolinguistics: e.g., Herbelot, von Redecker and Müller (2012) 'Distributional techniques for philosophical enquiry' (gender studies and intersectionality).
- Poetry: *Discourse.cpp* by O.S. le Si, edited by Aurélie Herbelot, available from `http://www.peerpress.de/`
- Today (very briefly)
  - Adjective and binomial ordering
  - Compound noun relations
- Logical metonymy and sense extension (Thursday)

Adjective and binomial ordering

- *gigantic striped box* not *striped gigantic box*
- *brandy and soda* not *soda and brandy*, *run and hide*
- some pairs are irreversible
- rare and novel phrases may be irreversible (*sake and grapefruit*, *armagnac and blackcurrant*)
- ordering principles partially semantic
- lots of discussion in literature about gendered examples: e.g., *boy and girl*

Adjective and binomial ordering: approaches

- adjective (pre-nominal modifier) ordering fairly well studied in CL: data-driven approaches, but still unseen pairs of adjectives. Back-off techniques include positional probabilities (later).
- binomial ordering less studied in CL (but Copestake and Herbelot, 2011)
- Benor and Levy (2006) corpus-based investigation of binomials
  - models include explicit semantic features, based on prior literature
  - e.g., Iconicity and Power

Mixed drinks: Iconicity or Power?

Gin and Bitters Drink Recipe

The Gin and Bitters cocktail is made from Gin and Angostura bitters, and served in a chilled cocktail glass.

Gin and Bitters Ingredients

- 3 oz Gin
- 1 tsp Angostura Bitters

Gin and Bitters Instructions

- Add the bitters to a cocktail glass.
- Swirl it around until the glass is fully coated.
- Fill with gin, and enjoy at room temperature.

Binomials and gender

- Male terms tend to precede female (for humans).
- e.g., *men and women* (85%), *boys and girls* (80%), *male and female* (91%) (% from Google ngram).
- Also personal names: e.g., *James and Sarah* (82%).
- Exceptions: *father and mother* (51%), *mothers and fathers* (67%), *ladies and gentlemen* (97%).
- B+L take gender as an example of the Power feature.
- BUT: possible phonological effects (female names tend to have more syllables than male).
- Animal terms often don't show a clear order: e.g., *stallion and mare* (50%), *stallion and broodmare* (54%), *ram and ewe* (50%), *sow and boar* (51%).

Binomials and gender

- Male terms tend to precede female (for humans).
- e.g., *men and women* (85%), *boys and girls* (80%), *male and female* (91%) (% from Google ngram).
- Also personal names: e.g., *James and Sarah* (82%).
- Exceptions: *father and mother* (51%), *mothers and fathers* (67%), *ladies and gentlemen* (97%).
- B+L take gender as an example of the Power feature.
- BUT: possible phonological effects (female names tend to have more syllables than male).
- Animal terms often don't show a clear order: e.g., *stallion and mare* (50%), *stallion and broodmare* (54%), *ram and ewe* (50%), *sow and boar* (51%).

Binomials and gender

- Male terms tend to precede female (for humans).
- e.g., *men and women* (85%), *boys and girls* (80%), *male and female* (91%) (% from Google ngram).
- Also personal names: e.g., *James and Sarah* (82%).
- Exceptions: *father and mother* (51%), *mothers and fathers* (67%), *ladies and gentlemen* (97%).
- B+L take gender as an example of the Power feature.
- BUT: possible phonological effects (female names tend to have more syllables than male).
- Animal terms often don't show a clear order: e.g., *stallion and mare* (50%), *stallion and broodmare* (54%), *ram and ewe* (50%), *sow and boar* (51%).

Analogical approach to binomial and adjective ordering

- our hypothesis: humans maintain order of known examples, order unseen by semantic similarity with seen
- essentially same model for binomials and adjectives
- baseline is to use positional probabilities (Malouf 2000)
- $a \prec b$

$$\begin{array}{ll} \text{if} & C(\text{a and b}) > C(\text{b and a}) \\ \text{or} & C(\text{a and b}) = C(\text{a and b}) \\ & \text{and} \\ & C(\text{a and})C(\text{and b}) > C(\text{b and})C(\text{and a}) \end{array}$$

  and conversely for $b \prec a$
- e.g., if *tea and biscuits* is known, prefer *tea and scones* over *scones and tea*

Adjective and binomial ordering: Kumar (2012)

- Same type of model used for adjectives and binomials: unseen cases ordered by k-nearest neighbour comparison to seen examples using distributional similarity.
- e.g., if ordering *coffee, cake* compare to all known binomials A and B based on similarities A:coffee, A:cake, B:coffee, B:cake, decide on basis of closest match (best k around 6 or 7).
- Distributions from unparsed WikiWoods data: significantly better than using positional probabilities.
- Expect further improvement using phonological features in addition.

Compound noun relations

- *cheese knife*: knife for cutting cheese
- *steel knife*: knife made of steel
- *kitchen knife*: knife characteristically used in the kitchen

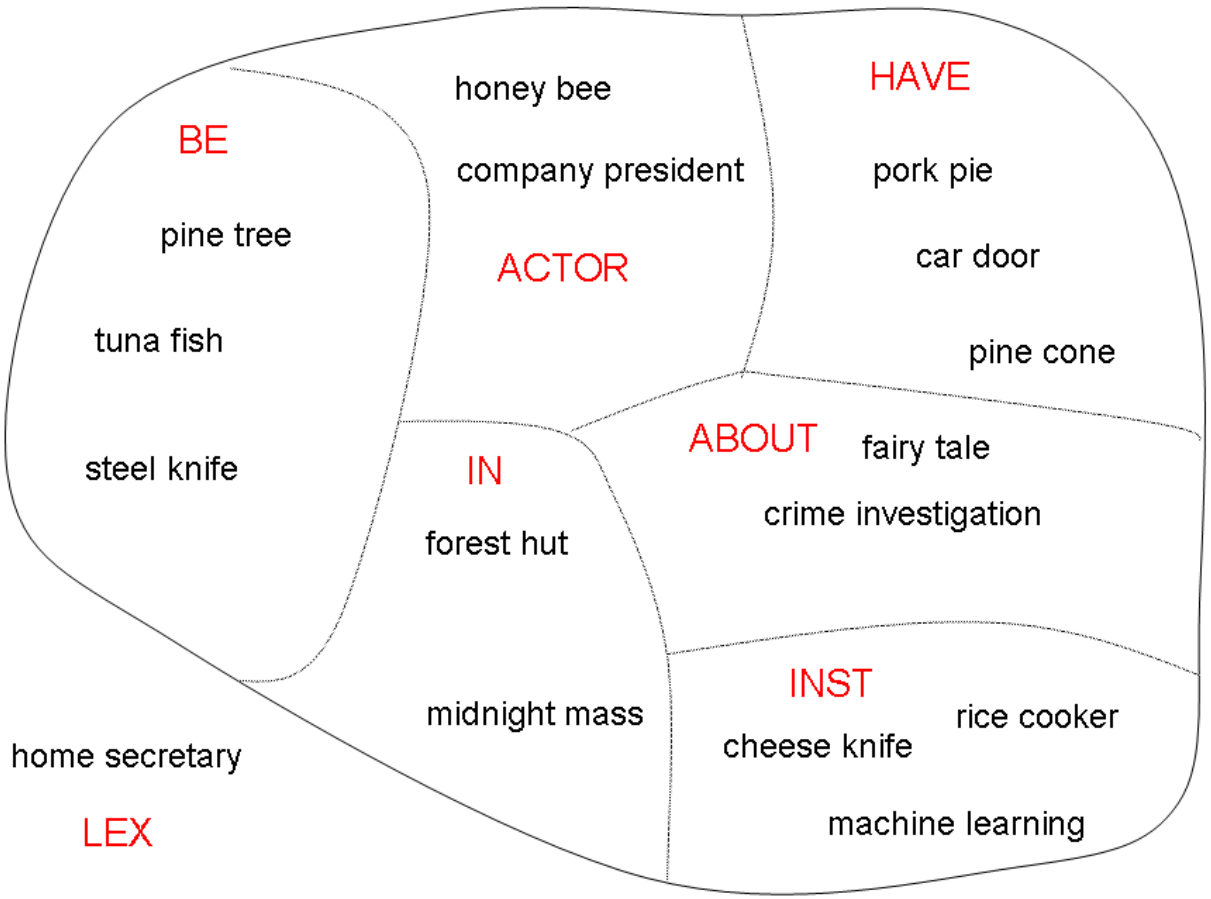
Automatic disambiguation:

- Syntactic parsers can't distinguish: N1(x), N2(y), compound(x,y)
- One approach: human annotation of compounds, use distributional techniques to compare unseen to seen examples.

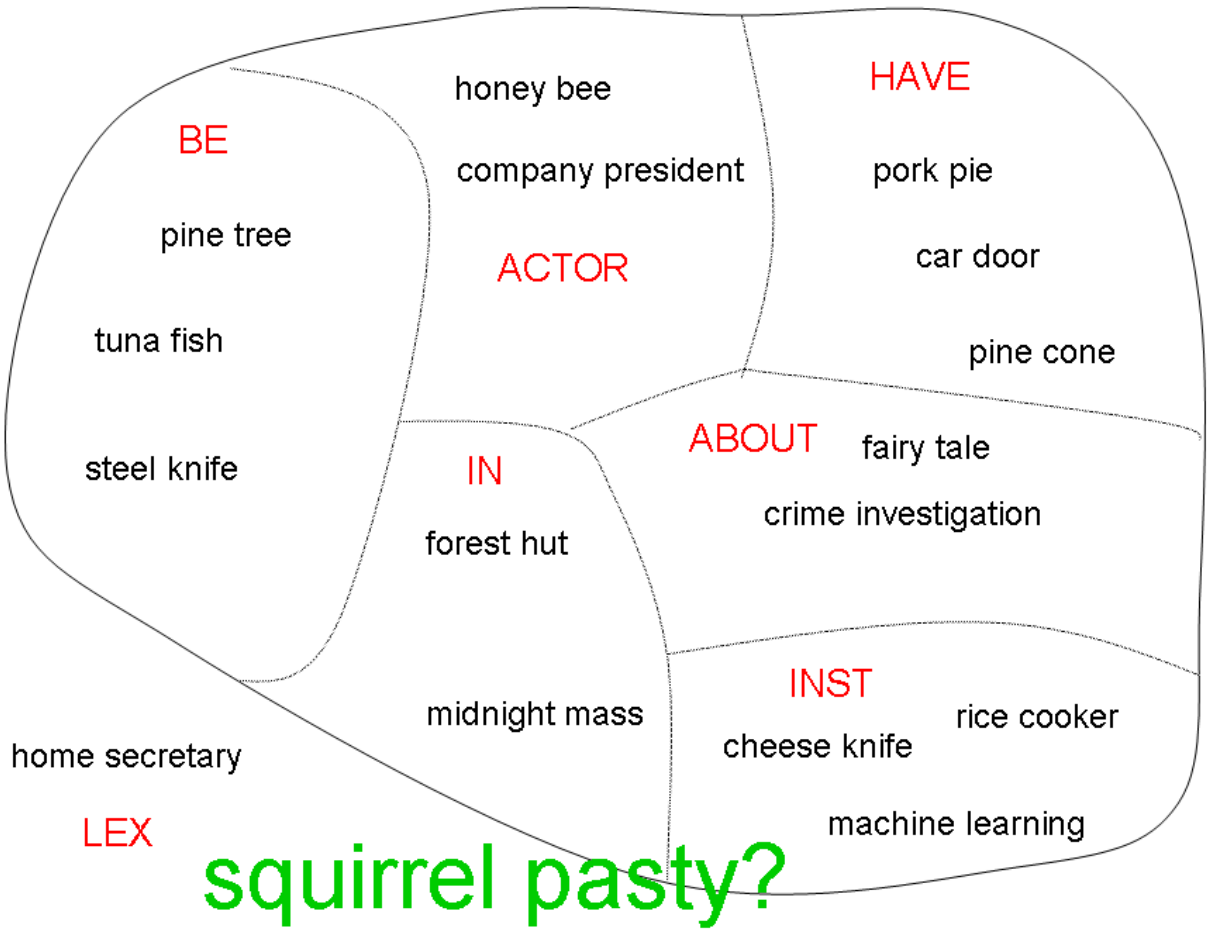
Compound noun relation schemes

- Lauer: prepositions, Lapata: verbal compounds, Girju et al, Turner.
- Ó Séaghdha, 2007: BE, HAVE, INST, ACTOR, IN, ABOUT: (with subclasses)
  LEX: lexicalised, REL: weird, MISTAG: not a noun compound.
  - Based on Levi (1978)
  - Considerable experimentation to define a usable scheme: some classes very rare (therefore not annotated reliably)
  - Annotation of 1400 examples from BNC by two annotators.

Compound noun relation learning



Compound noun relation learning



Squirrels and pasties

Compound noun relation learning

- Ó Séaghdha, 2008 (also Ó Séaghdha and Copestake, forthcoming)
- Treat compounds as single words: doesn't work!
- Constituent similarity: compounds x1 x2 and y1 y2, compare x1 vs y1 and x2 vs y2.
  squirrel vs *pork*, *pasty* vs *pie*
- Relational similarity: **sentences** with x1 and x2 vs sentences with y1 and y2.
  squirrel is very tasty, especially in a pasty vs
  pies are filled with tasty pork
- Comparison using kernel methods: allows combination of kernels.
- Best accuracy: about 65% (slightly lower than agreement between annotators) using combined kernels.

Summary

- Both applications described depend on using distributional similarity to match known cases: a type of analogical reasoning.
- Known examples may be explicitly annotated (this approach to compounds) or based on observation (adjectives and binomials).
- Techniques can be simple (k-nearest neighbours) or more complex (Ó Séaghdha's use of kernel methods).
- Range of other possible applications — we will return to some of these on Thursday.

Motivation

- Formal semantics gives an elaborate and elegant account of the productive and systematic nature of language.
- The formal account of compositionality relies on:
  - *words* (the minimal parts of language, with an assigned meaning)
  - *syntax* (the theory which explains how to make complex expressions out of words)
  - *semantics* (the theory which explains how meanings are combined in the process of particular syntactic compositions).

Motivation

- But formal semantics does not actually say anything about lexical semantics (the meaning of *cat*, cat′, is the set of all cats in particular world).
- Distributions a potential solution?
- Also, if we make the approximation that distributions are 'meaning', then we need a way to account for compositionality in a distributional setting.

Why not just look at the distribution of phrases?

- The distribution of phrases – even sentences – can be obtained from corpora, but...
  - those distributions are very sparse;
  - observing them does not account for productivity in language.
- Some models assume that corpus-extracted phrasal distributions are irrelevant data.
- Some models assume that, given enough data, corpus-extracted phrasal distributions have the status of gold standard.

Some distributional compositionality models

- Mitchell and Lapata (2010): word-based model, task-evaluated.
- Baroni and Zamparelli (2010): word-based, evaluated against phrasal distributions.
- Coecke, Sadrzadeh and Clark (2011): CCG-based model, task-evaluated.

The model

- Word-based (5 words on either side of the lexical item under consideration).
- The composition of two vectors $\vec{u}$ and $\vec{v}$ is some function $f(\vec{u}, \vec{v})$. M & L try:
  - addition $p_i = \vec{u}_i + \vec{v}_i$
  - multiplication $p_i = \vec{u}_i.\vec{v}_i$
  - tensor product $p_{ij} = \vec{u}_i.\vec{v}_j$
  - circular convolution $p_{ij} = \sigma_j \vec{u}_j.\vec{v_{i-j}}$
  - ... etc
- Task-based evaluation: similarity ratings. Multiplication is best measure.

Example

early_j	**age_n**	**early_j age_n**
africa::9.75873	africa::3.56225	africa::34.76303
african::6.87337	african::1.88733	african::12.97231
aftermath::3.40748	aftermath::1.37812	aftermath::4.69591
afternoon::42.2096	afternoon::1.9041	afternoon::80.3712
afterwards::7.46585	afterwards::3.86807	afterwards::28.87843
again::9.00563	again::2.78339	again::25.06618
age::15.6464	age::0	age::0
aged::5.99896	aged::24.6173	aged::147.67819
agencies::4.91747	agencies::1.57129	agencies::7.72677
agency::7.28471	agency::3.13776	agency::22.85767
agent::4.63014	agent::2.24935	agent::10.41480
agents::4.21793	agents::1.68319	agents::7.09957
ages::45.003	ages::0	ages::0
ago::18.8909	ago::19.2306	ago::363.2833
agree::5.05183	agree::3.67157	agree::18.54814
agreed::6.36066	agreed::2.61272	agreed::16.61862
agreement::7.64836	agreement::0.912126	agreement::6.976268
agricultural::11.3745	agricultural::2.66057	agricultural::30.26265

Difference in top-rated contexts for *early age*

multiplication
1990s
1980s
1970s
20th
1960s
childhood
1950s
age
1940s
1920s
1930s
19th
late
century
morning
stages
settlers
warning

phrase
talent
interested
showed
learned
piano
studying
exposed
ages
parents
encouraged
singing
educated
interest
uncle
violin
baronet
eldest
raised

Discussion: the meaning of *f*

- How do we interpret $f(\vec{u}, \vec{v})$ linguistically?
- Intersection in formal semantics has a clear interpretation:
  $\exists x[\text{cat}'(x) \wedge \text{black}'(x)]$
  There is a cat in the set of all cats which is also in the set of black things.
- But what with addition, multiplication (let alone circular convolution)??

Addition

- Addition is not intersective: the whole meaning of both $\vec{u}$ and $\vec{v}$ are included in the resulting phrase.
- No sense disambiguation and no indication as to how an adjective, for instance, modifies a particular noun (i.e. the distributions of *red car* and *red cheek* both include high weights on the *blush* dimension).
- Too much information

Multiplication

- Multiplication is intersective.
- But it is commutative in a word-based model:

$$\overrightarrow{\text{The cat chases the mouse}} = \overrightarrow{\text{The mouse chases the cat}}.$$

Assumptions

- Given enough data, distributions for phrases should be obtained in the same way as for single words.
- There is no single composition operation for adjectives. Each adjective acts on nouns in a different way.

Adjective types, Partee (1995)

- **Intersective:** carnivorous mammal
  $||\text{carnivorous mammal}|| = ||\text{carnivorous}|| \cap ||\text{mammal}||$
- **Subsective:** skilful surgeon
  $||\text{skilful surgeon}|| \subseteq ||\text{surgeon}||$
- **Non-subsective:** former senator
  $||\text{former senator}|| \neq ||\text{former}|| \cap ||\text{senator}||$
  $||\text{former senator}|| \not\subseteq ||\text{senator}||$

System

- For each adjective, a matrix is learned from actual AN phrases using partial least squares regression.
- Test by measuring distance between a given adjective-noun combination and the corresponding phrasal distribution.

Outline

Overview

- Based on pregroup grammar.
- Composition involves tensor product and point-wise multiplication.
- Evaluated on similarity task.

Thanks to Steve Clark for some of the slides!

Pregroup grammar

- A pregroup is a partially ordered monoid in which each element a has a *left adjoint* a^l and a *right adjoint* a^r such that

$$a^l \cdot a \to 1, \quad a \cdot a^r \to 1$$

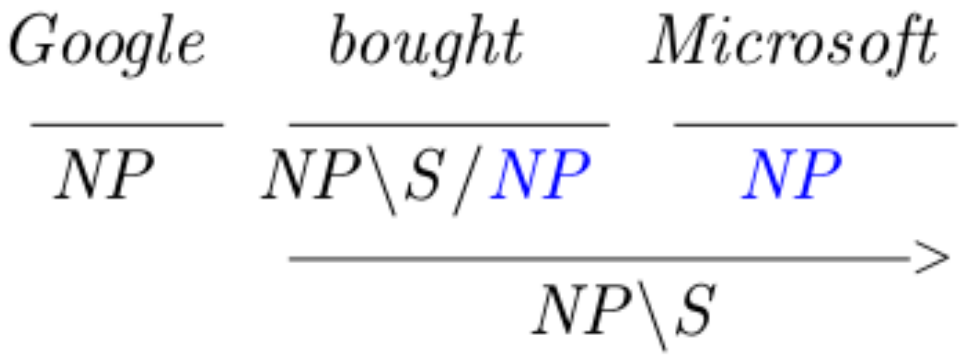
- The monoid is the set of grammatical types (NP, NP^r, NP^l, NP^{rr}, NP^{ll}, S, PP, ...) with the juxtaposition operator ($\cdot$) used to derive complex types and the empty string as unit (1)

$$NP \cdot (NP^r \cdot S \cdot NP^l) \cdot NP$$

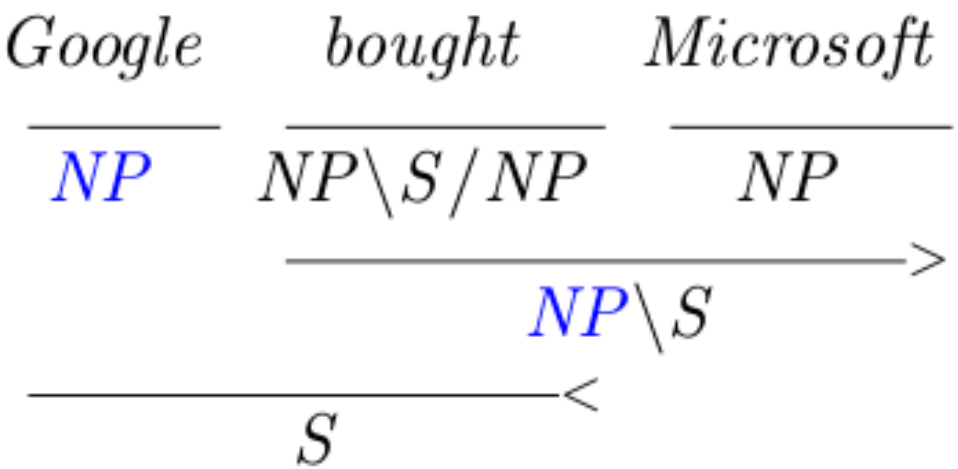
Categorial Grammar Derivation

$$\frac{Google}{NP} \quad \frac{bought}{NP\backslash S/NP} \quad \frac{Microsoft}{NP}$$

Categorial Grammar Derivation



Categorial Grammar Derivation



Pregroup Derivation

$$\frac{Google}{NP} \quad \frac{bought}{NP^r \cdot S \cdot NP^l} \quad \frac{Microsoft}{NP}$$

Pregroup Derivation

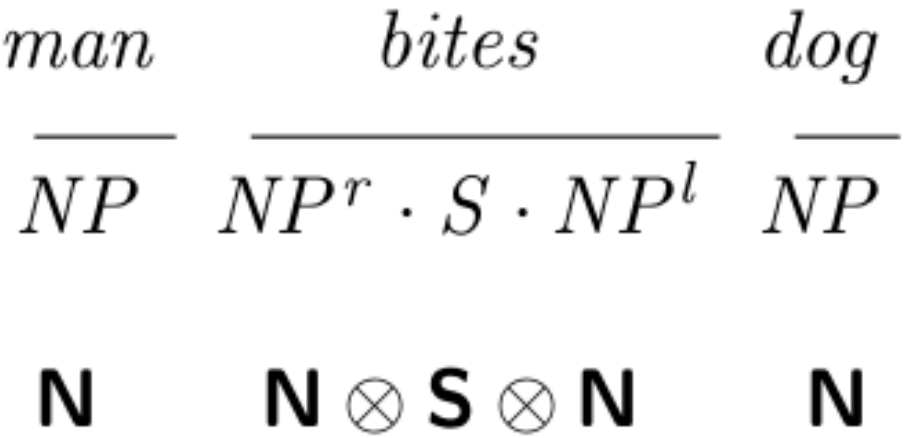
$$\dfrac{\dfrac{Google}{NP} \quad \dfrac{bought}{NP^r \cdot S \cdot NP^l} \quad \dfrac{Microsoft}{NP}}{NP^r \cdot S}$$

Pregroup Derivation

$$\dfrac{\dfrac{\dfrac{Google}{NP} \quad \dfrac{bought}{NP^r \cdot S \cdot NP^l} \quad \dfrac{Microsoft}{NP}}{NP^r \cdot S}}{S}$$

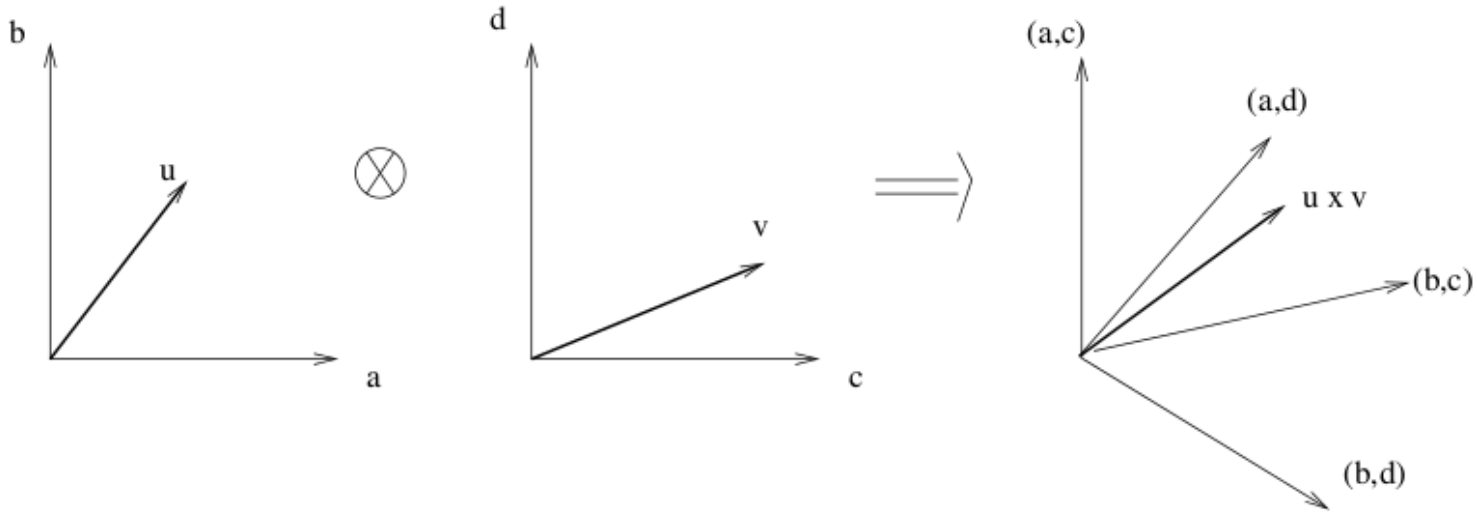
Various semantics spaces

- Lexical items of various grammatical types live in different 'spaces'.



- Representations can be vectors or matrices.
  e.g. a transitive verb may be a matrix represented in a tensor product space **N** $\otimes$ **S** $\otimes$ **N**.
- Basic types like nouns are vectors with components equal to TF*IDF values.
- Composition involves point-wise multiplication.

The tensor product



$$(u \otimes v)_{(a,d)} = u_a . v_d$$

The sentence space

- What is the sentence space?
- Truth-theoretic interpretation: sentence space has two dimensions, **True** and **False**.
- Distributional interpretation: a point in the distributional space used for verbs. But what does this really mean (in particular in the case of complex sentences)??

Truth in a 2-dimensional space

dog chases cat

	⟨fluffy,T,fluffy⟩	⟨fluffy,F,fluffy⟩	⟨fluffy,T,fast⟩	⟨fluffy,F,fast⟩	⟨fluffy,T,juice⟩	⟨fluffy,F,juice⟩	⟨tasty,T,juice⟩ . . .
$\overrightarrow{chases}$	0.8	0.2	0.75	0.25	0.2	0.8	0.1
dog,cat	0.8,0.9	0.8,0.9	0.8,0.6	0.8,0.6	0.8,0.0	0.8,0.0	0.1,0.0

$$\overrightarrow{dog\ chases\ cat}_{\mathbf{T}} = 0.8 \,.\, 0.8 \,.\, 0.9 + 0.75 \,.\, 0.8 \,.\, 0.6 + 0.2 \,.\, 0.8 \,.\, 0.0 + 0.1 \,.\, 0.1 \,.\, 0.0 + \ldots$$

Sentence meaning in a multi-dimensional space

dog chases cat

	⟨fluffy,fluffy⟩	⟨fluffy,fast⟩	⟨fluffy,juice⟩	⟨tasty,juice⟩	⟨tasty,buy⟩	⟨buy,fruit⟩	⟨fruit,fruit⟩ . . .
$\overrightarrow{chases}$	0.8	0.75	0.2	0.1	0.2	0.2	0.0
dog,cat	0.8,0.9	0.8,0.6	0.8,0.0	0.1,0.0	0.1,0.5	0.5,0.0	0.0,0.0
$\overrightarrow{dog\ chases\ cat}$	0.576	0.36	0.0	0.0	0.01	0.0	0.0

The meaning of the sentence

- In formal semantics, meaning is denotational and truth-theoretic.
- *Kim sleeps* is true iff Kim is in the set of sleeping things.
- Distributions are more about intension than extension, so should we talk of truth?
- If not, what should the meaning of a sentence be?

Beyond intersection

- What about non-intersective composition? (*fake, small, alleged*...)
- Even the semantics of intersective phrases is more than the intersection of their parts.

Is intersection enough?

A big city: just a city which is big?
See *loud, underground, advertisement, crowd, Phantom of the Opera*...

What should we compose?

> *one has the common intuition that there is a perceived difference between [...] "Indian elephant" and "friendly elephant". [...] an Indian elephant is one of a recognized variety of elephants, and their properties are not simply those of being an elephant, and being from India, but something more (such as disposition, size of ears, etc. etc.) – it's a (sub)species. In this sense, "Indian elephant" differs from "friendly elephant" because a friendly elephant is no more than an elephant that is friendly, and that's it.*
>
> *Carlson (2010)*

- What is the best representation for *Indian elephant*? The phrase or the composed form? Or both? (But how to do both??)

Logical operators

- Treatment of logical operators is unclear.
- In formal semantics, a quantifier 'counts' over the elements of a set.
  $Q(x)[rstr(x) \wedge scp(x)]$
  $\exists(x)[\text{cat}'(x) \wedge \text{run}'(x)]$
- No set in distributional semantics...

Generative lexicon and distributional semantics

Introduction to the Generative Lexicon

Contextual coercion

- GL account of contextual coercion
- Corpus-based approach to logical metonymy

English compound noun relations

Polysemy

The Generative Lexicon (Pustejovsky, 1991, 1995)

- Polysemy is pervasive.
- Sense enumeration is not an adequate treatment.
- Several types of polysemy including: regular polysemy, constructional polysemy and sense modulation.
- Lexical semantic information in qualia structure (also argument structure, event structure and inheritance structure).
- Later work emphasizes dot object: not discussed here.

Qualia structure in GL

Certain aspects of meaning are highly salient:

- FORMAL ROLE
- CONSTITUTIVE ROLE
- TELIC ROLE (purpose)
- AGENTIVE ROLE

From Pustejovsky (1991):

```
novel(*x*)
  Const: narrative(*x*)
  Form: book(*x*), disk(*x*)
  Telic: read(T,y,*x*)
  Agentive: artifact(*x*), write(T,z,*x*)
```

Qualia structure in GL

- GL: qualia structure provides metonymic interpretations in a range of contexts.
- Also, perhaps, controls application of certain processes.
- Computational approaches to qualia: encode on lexical entries (via feature structures), default inheritance over a semantic hierarchy.
- Semi-automatic acquisition from Machine Readable Dictionaries.

GL and distributional semantics

Today: GL account and distributional experiments:

- Contextual coercion
- Compound nouns
- Regular polysemy

Encoding semantics: feature structures vs distributions?

Contextual coercion

- After 6pm, most of the bars are full.
- After running for an hour, Kim was very thirsty.
- After the talk, we could go for a drink.
- After three martinis, Kim felt much happier.

Contextual coercion:

- After drinking three martinis, Kim felt much happier.

Adjectives

- fast runner: someone who runs fast
- fast typist: someone who (can) type fast
- fast car: car which can go fast

Not plausible that there are different senses of fast for each different context.

enjoy

- Mary enjoyed the book.
- Mary enjoyed reading the book.
- * Mary enjoyed that she read the book.
- ? Mary enjoyed the table.

‘Sylvie and Bruno Concluded’ (Lewis Carroll)

“You seem to enjoy that cake?” the Professor remarked.
“Doos that mean ‘munch?” Bruno whispered to Sylvie.
Sylvie nodded. “It means ‘to munch and ‘to like to munch.”
Bruno smiled at the Professor. “I doos enjoy it,” he said.
The Other Professor caught the word. “And I hope you’re enjoying yourself, little Man?” he enquired.
Bruno’s look of horror quite startled him. “No, indeed I aren’t!” he said.

Logical metonymy

Additional meaning systematically arises for some verb/noun or adjective/noun combinations:

- Kim enjoyed the cake
- Kim enjoyed eating the cake
- semantically, *enjoy*, *finish* etc always take an eventuality: different syntactic variants are thus closely related.
- contextual coercion of object-denoting NP to an appropriate eventuality

enjoy the cake would mean something like:

$$enjoy(e, x, e') \wedge cake(y) \wedge R(e', x, y)$$

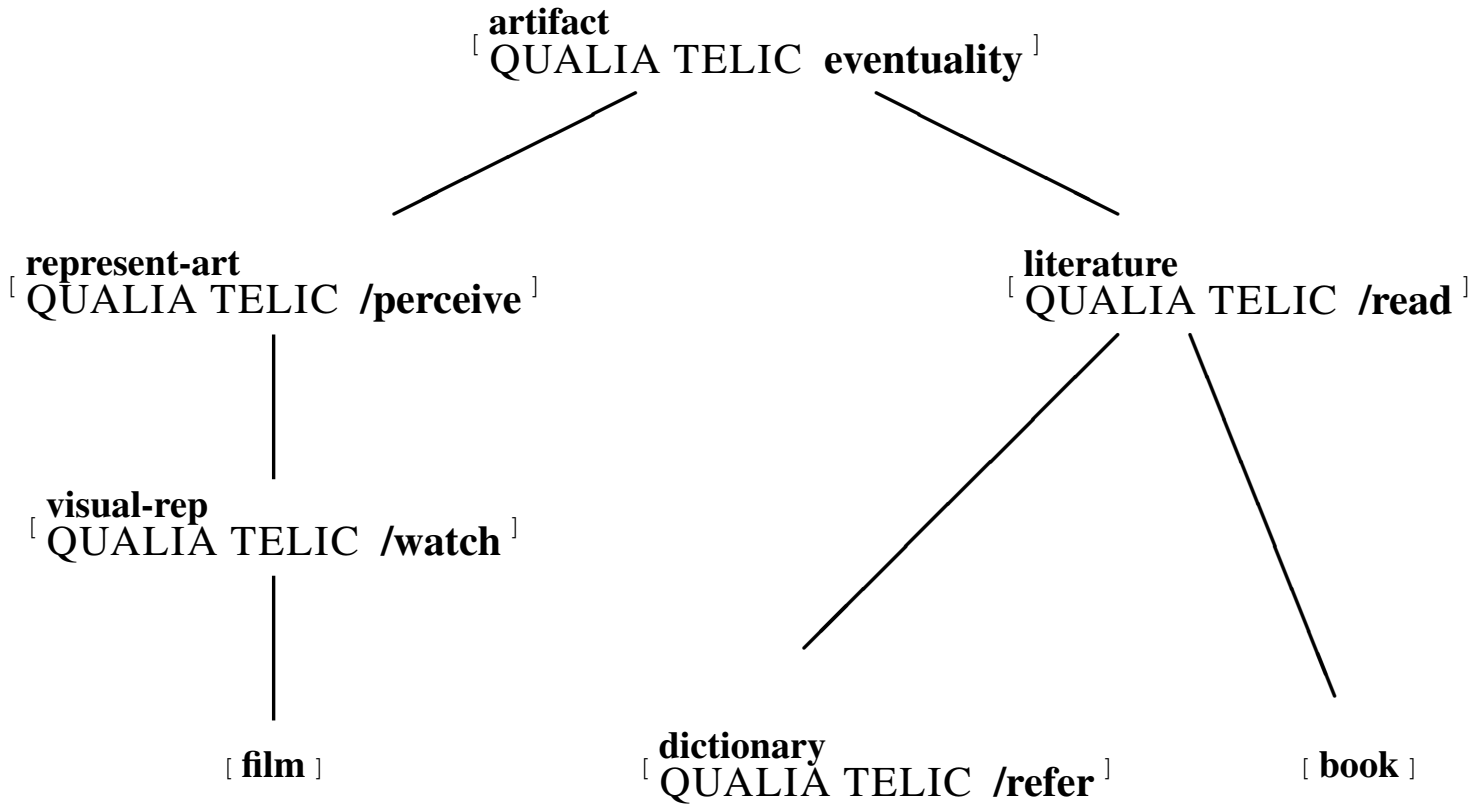
The metonymic event

- The default interpretation is supplied lexically
- A noun has qualia structure, e.g., telic role of *cake* is 'eat', agentive role is 'bake'

$$enjoy(e, x, e') \wedge cake(y) \wedge purpose(cake)(e', x, y)$$

where $purpose(cake) = eat$.

Qualia structure in feature structures



Interaction with pragmatics 1

Assume lexical defaults because of interpretation in unmarked contexts (Briscoe et al, 1990).

> *Willie enjoyed the hot sweet tea, standing on the deck in the cool of the night.*

Less common to find explicit verb when default is meant (e.g., don't tend to find *enjoy drinking the tea*).

But unusual verbs are specified.

Interaction with pragmatics 2

Defaults can be contextually overridden:

- John Grisham enjoyed that book.
- The goat enjoyed the book.

Override if Grisham is an author and goats don't read.
Similarly:

- After that book, Kim felt her knowledge of semantics had greatly improved.
- After that book, John Grisham became a household name.
- After that book, the goat had indigestion.

Formalisation in terms of typed default feature structures and non-monotonic logic (e.g., Lascarides and Copestake, 1998).

Context and adjectives

All the office personnel took part in the company sports day last week. One of the programmers was a good athlete, but the other was struggling to finish the courses. The fast programmer came first in the 100m.
cf Pollard and Sag discussion of *good linguist*

Logical metonymy restrictions

Logical metonymy has restrictions:

? Kim enjoyed the pebble.

? Kim enjoyed the dictionary.

Differences between verbs:

Kim began the salad. (eating, making)

Kim enjoyed the salad. (eating)

Kim began the book. (reading, writing)

Kim enjoyed the book. (reading)

Kim began the tunnel / bridge / path / road (constructing only)

(examples first pointed out by Godard and Jayez, 1993)

tunnels

Why is the tunnel example strange with a telic interpretation, when the explicit version is impeccable?

? Kim began the tunnel.

Kim began driving through the tunnel.

Context does not make this better:

The drive to the Alps had been long and tiring, and Kim was prone to claustrophobia.

*Therefore it was with considerable trepidation that Kim began the first tunnel.

and again . . .

tunnel can have a telic interpretation with other verbs:
But after the first tunnel, Kim felt much happier.
But much to his surprise, Kim enjoyed the first tunnel.
'telic' interpretation for *begin* in corpora seems to be almost completely restricted to foodstuffs, drinks and books (Verspoor, 1997)

Lapata and Lascarides (2003)

- corpus-based model that can retrieve plausible verbs for logical metonymy
- rough idea: given a metonymic verb (*begin, enjoy* etc) and a candidate object noun, find most frequent verbs that occur with the metonymic verb, and most frequent verbs that occur with the noun
- reasonable agreement with human judgements
- model is better than noun-only baseline
- alternative to qualia (arguably, even more lexical!)

Lapata and Lascarides (2003)

(9) a. Siegfried bustled in, muttered a greeting and began to pour his coffee.
b. She began to pour coffee.
c. Jenna began to serve the coffee.
d. Victor began dispensing coffee.

(10) a. I was given a good speaking part and enjoyed making the film.
b. He's enjoying making the film.
c. Courtenay enjoyed making the film.
d. I enjoy most music and enjoy watching good films.
e. Did you enjoy acting alongside Marlon Brando in the recent film The Freshman?

Lapata and Lascarides (2003): estimating probabilities

$P(e, o, v) = P(e).P(v|e).P(o|e, v)$
v = verb (enjoy), o = object, e = event
Maximum likelihood estimates via frequencies for $P(e)$ and $P(v|e)$, but not $P(o|e, v)$ because 'usual' verbs are not made explicit.
Assume $P(o|e, v) \approx P(o|e)$
i.e., count frequencies of verbs with objects regardless of enjoyment etc

Lapata and Lascarides (2003): some results

begin book - read (15.49) /write (15.52)
enjoy book - read (16.48) / write (16.48)
But: begin sandwich - bite into (18.12) / eat (18.23)
Adding in subject:
author book - write - 14.87
author book - read - 17.98
student book - read - 16.12
student book - write - 16.48
Correlation with human judgements: verbs r=.64, adjectives r=.40
Lower probabilities for weird examples. (but compare with infrequent good examples)?

Lapata and Lascarides (2003)

Problems:

- disambiguation: e.g., *fast plane*
- the model recovers individual verbs, but this is too specific:

  Kim enjoyed the soup

- sparse data (trained on BNC)
- titles etc:

  Sandy did not enjoy 'Sylvie and Bruno'

  back-off to semantic classes required in such cases
- not clear that it fully accounts for the semi-productivity facts model gives interpretations for 'enjoy the ice cream' but also 'begin the rock'

General distributional approach to logical metonymy?

- Still require a syntax-semantics interface component: distributions as a replacement for qualia, rather than whole of GL account.
- Ideally, want the metonymic interpretation to 'fall out' of a general distributional model of meaning.
  Distributional models with more complex feature spaces might allow this.
- Metonymic interpretation should be a distribution, approximately realizable as word(s)?
- How to allow for restrictions on *enjoy* etc, given that 'usual' verbs aren't explicit?
- Pragmatic overriding based on distribution associated with individual entity?

Compound noun relations

- *cheese knife*: knife for cutting cheese
- *steel knife*: knife made of steel
- *kitchen knife*: knife characteristically used in the kitchen

Very limited syntactic/phonological cues in English, so assume parser gives: N1(x), N2(y), compound(x,y).

Overgeneration: German compounds with non-compound translations

Arzttermin	*doctor appointment
Terminvorschlag	* date proposal
Terminvereinbarung	* date agreement
Januarhälfte	* January half
Frühlingsanfang	* spring beginning

1. **doctor appointment*/*doctor's appointment* — possessive compounds
2. **date agreement*/*agreement on a date* — head derived from a PP-taking verb (*water seeker* vs **water looker*)
3. **spring beginning*/*beginning of spring* — relational nouns as heads?

GL approach

Johnston and Busa (1996)

- *bread knife*: telic interpretation. Italian *da* (*coltello da pane*)
- *lemon juice*: origin interpretation. Italian *di* (*succo di limone*)
- *glass door*: constitutive interpretation. Italian *a* (*porta a vetri*)

Problems:

- Only applicable to a limited number of English compounds.
- How are readings selected?

Data-driven approaches to compound relation learning

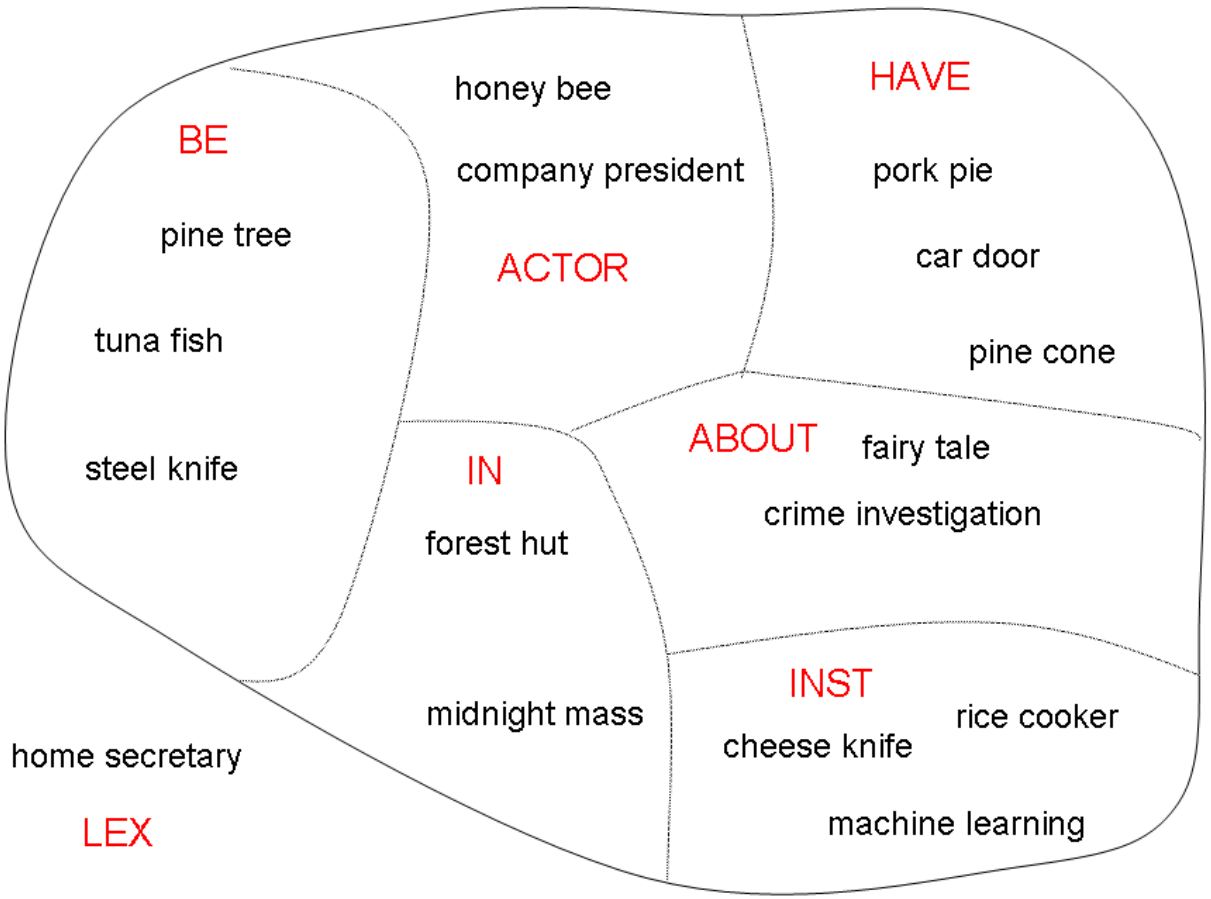
- Find paraphrases by looking for explicit relationships (Lauer: prepositions, Lapata: verbal compounds)
- OR human annotation of compounds, use distributional techniques to compare unseen to seen examples. Girju et al, Turner, Ó Séaghdha among others.

Relation schemes for learning experiments

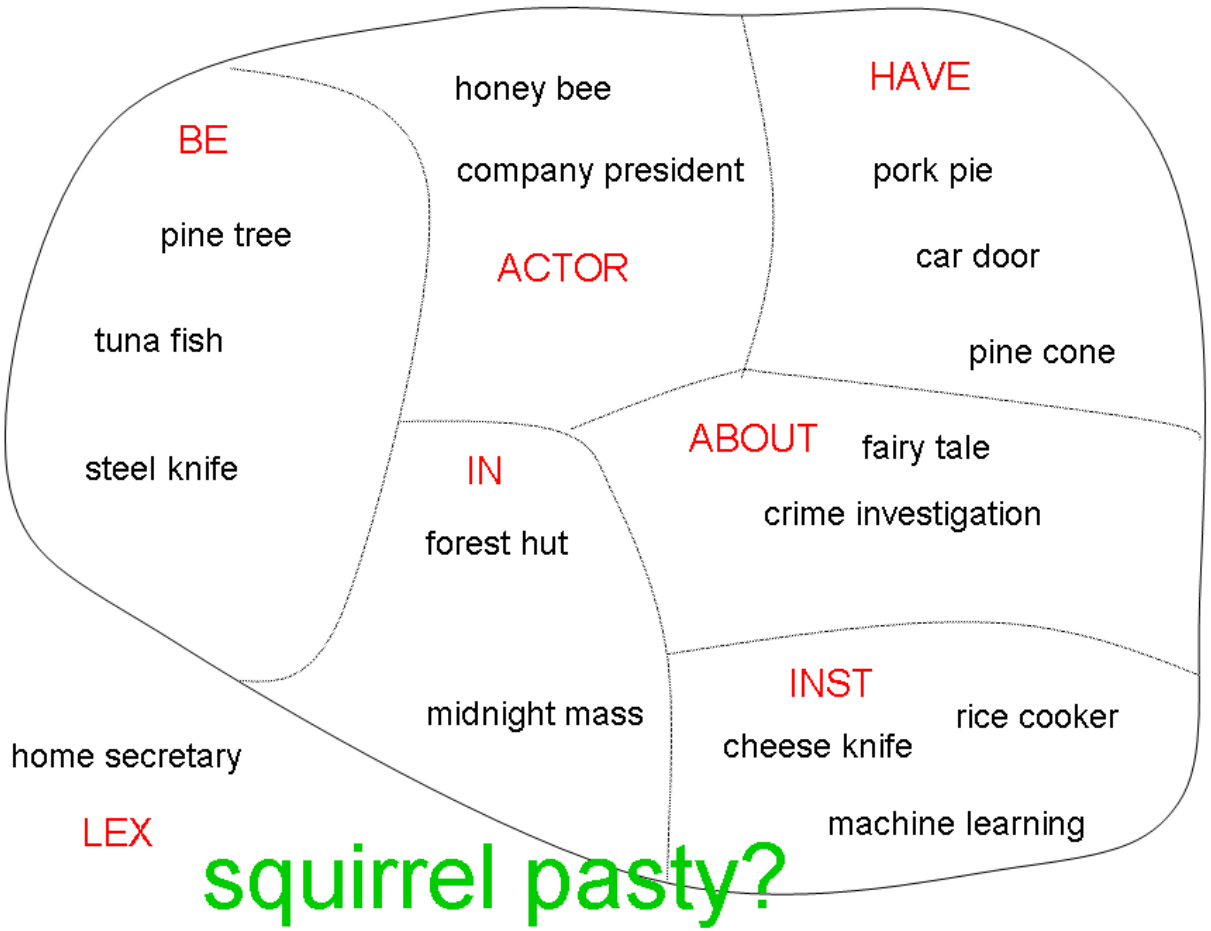
Ó Séaghdha (2007):

- BE, HAVE, INST, ACTOR, IN, ABOUT: (with subclasses) LEX: lexicalised, REL: weird, MISTAG: not a noun compound.
  - Based on Levi (1978)
  - Considerable experimentation to define a usable scheme: some classes very rare (therefore not annotated reliably)
  - Annotation of 1400 examples from BNC by two annotators.

Compound noun relation learning



Compound noun relation learning



Compound noun relation learning

- Ó Séaghdha, 2008 (also Ó Séaghdha and Copestake, forthcoming)
- Treat compounds as single words: doesn't work!
- Constituent similarity: compounds x1 x2 and y1 y2, compare x1 vs y1 and x2 vs y2.
  squirrel vs *pork*, *pasty* vs *pie*
- Relational similarity: **sentences** with x1 and x2 vs sentences with y1 and y2.
  squirrel is very tasty, especially in a pasty vs
  pies are filled with tasty pork
- Comparison using kernel methods: allows combination of kernels.
- Best accuracy: about 65% (slightly lower than agreement between annotators) using combined kernels.

Analogical reasoning using distributions

- Using distributional similarity to match known cases: a type of analogical reasoning.
- Known examples explicitly annotated (this approach to compounds) or based on observation (adjectives and binomials).
- Relatively sophisticated techniques allow combination of evidence types (Ó Séaghdha's use of kernel methods).
- Explicit relations could be thought of as a label for distributions?

Polysemy and regular polysemy

- homonymy — unrelated meanings — *bank* (river bank) vs *bank* (financial institution)
- general polysemy — related meanings but no systematic connection — *bank* (financial institution) vs *bank* (in a casino)
- regular polysemy — regularly related meaning — *bank* (N) (financial institution) vs *bank* (V) (put money in a bank), compare *store*, *cache* etc
- vague/general terms — e.g. *teacher* may be male or female (in English, other languages may distinguish)

Some types of regular polysemy/sense extension

- Count/mass: animal/meat, tree/wood . . . (generically thing/derived substance grinding)
  After several lorries had run over the body, there was rabbit splattered all over the road.
- Mass/count: portions , kinds.
  two beers: 'two servings of beer' or 'two types of beer'
- Verb alternations: causative/inchoative . . .
- Noun-verb conversions: *sugar*, *hammer*, *tango* (cf derivational endings *-ize*)

Established senses tend to have additional conventional meaning.

More regular polysemy/sense extension

- Container-contents: *bottle*
  He drank a bottle of whisky
  paralleled by suffixation with -ful
  He drank a bottleful of whisky
- Plant/fruit: *olive, grapefuit* (cf *aceituna/aceituno, pomela/pomelo*)
- Broadening of senses (maybe):
  cloud: normal sense is weather, but also e.g., *cloud of dust*
  forest, *bank* . . .
- Figure/ground (maybe):
  Kim painted the door. vs *Kim walked through the door.*
  ? Kim painted the door but got paint on herself when she walked through it.

Systematic polysemy and translation

- If similar polysemy patterns: no need to disambiguate.
- Marked: disambiguate but straightforward translation.
  Plant/fruit examples: grapefuit (pomela/pomelo).
- Translation mismatch due to differences in systematic polysemy patterns.
  - hammer a nail into a frame
  - enfoncer un clou dans un cadre avec un marteau
    Literally: drive a nail into a frame with a hammer
  - mettre un clou dans un cadre avec un marteau
    Literally: put a nail into a frame with a hammer

Metonymy: country names

- Location: About 300 Australians will remain inside Iraq on logistical and air surveillance duties.
- Government: The US and Libya have agreed to work together to resolve compensation claims . . .
- Teams: England are comfortable 3-0 winners in their end-of-season friendly against Trinidad and Tobago. (football)
  Stand-in England boss Rob Andrew said Sunday's laboured win over the Barbarians was a "useful" exercise. (rugby union)

Metonymy: object to person

- *The cello is playing badly.*
  (the person playing the cello)
- *His Dad was a Red Beret.*
  (i.e., someone who wore a red beret: here a British paratrooper)
- *Chester serves not just country folk, but farming, suburban and city folk too. You'll see Armani drifting into the Grosvenor Hotel's exclusive (but exquisite) Arkle Restaurant and C&A giggling out of its streetfront brasserie next door.*
  (i.e., people who wear Armani/C&A clothes)

Nunberg (1978)

Restaurant context:

- The ham sandwich wants his check
  (meaning ‘person who ordered a ham sandwich’)
- The french fries wants his check
- * The ham sandwich wants a coke and has gone stale
- * The brown suit is in the microwave

Compare:

- I’m parked out back.

GL accounts of regular polysemy

- Regular polysemy often involves syntactic effects.
- Lexical rules in a feature structure framework can capture syntax and semantics.
- This requires a generative account of the derived meaning.
- Application of rules has to be controlled: probabilities and productivity metrics (but practical problems with deriving these from corpora).

Metonymy disambiguation

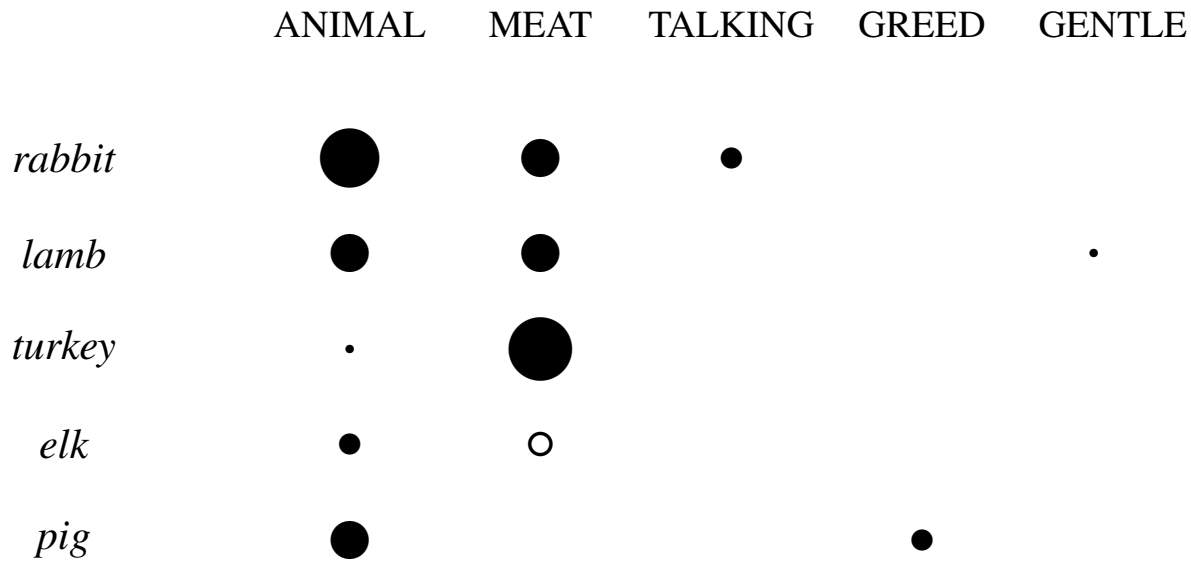
Regular sense distinctions/metonymy (e.g., place/government for countries):

> *China_org admits to climate failings.*
> *The company already owns nearly 50 stores in China_place, . . .*

Generalisation is possible, human agreement is often better that for WSD: better disambiguation performance.

Distributional approaches to regular polysemy

LC account: subspaces correspond to distributions for individuals and for groups of individuals (senses/usages) (also Rapp, 2004 on clustering; Boleda and Padó (2012) on regular polysemy).



Summary: qualia structure versus distributions

Issues with qualia structure:

- What types of thing are the fillers for qualia roles? (disjunctions of) lexemes, concepts? *smoke*: cigars, fish?
- Should they have associated probabilities?
- Can symbolic qualia values account for the observed data?
- How are qualia learned by humans?

Distributional accounts look more promising (but only if we can build single models, and don't use implausible amounts of data).

Feature structures versus distributions in general

- Feature structures are appropriate when:
  - Small number of relevant roles.
  - Role fillers can be isolated.
  - Defined processes (e.g., in grammars) which access those roles.
- Distributions are appropriate when:
  - No fixed set of roles or no role/filler distinction. Abstraction over any concept is possible (can't abstract over features).
  - Data source for distributions exists.
  - Sometimes: as a way of learning appropriate role fillers.
- Interfaces are necessary and not yet well understood.

Overview

- Introducing a relation between quantification and lexical semantics.
- A short recap on Lexicalised Compositionality.
- Doing model-theoretic quantification with distributions.
- Moving away from truth theory into a model of language comprehension.
- How to learn quantification? A real example.