Extracting Conceptual Knowledge from Text Corpora

Marco Baroni

Center for Mind/Brain Sciences
(University of Trento)

Napoli
April 16, 2008

Collaborators

- Brian Murphy, Eduard Barbu, Massimo Poesio (CIMeC)
- Alessandro Lenci (University of Pisa and ILC/CNR)
- Marco Baroni and Alessandro Lenci, 2008. Concepts and properties in word spaces, *Italian Journal of Linguistics*, special issue on distributional models of the lexicon
- Marco Baroni, Eduard Barbu, Brian Murphy and Massimo Poesio, in preparation. StruDEL: A distributional semantic model based on properties and types

Outline

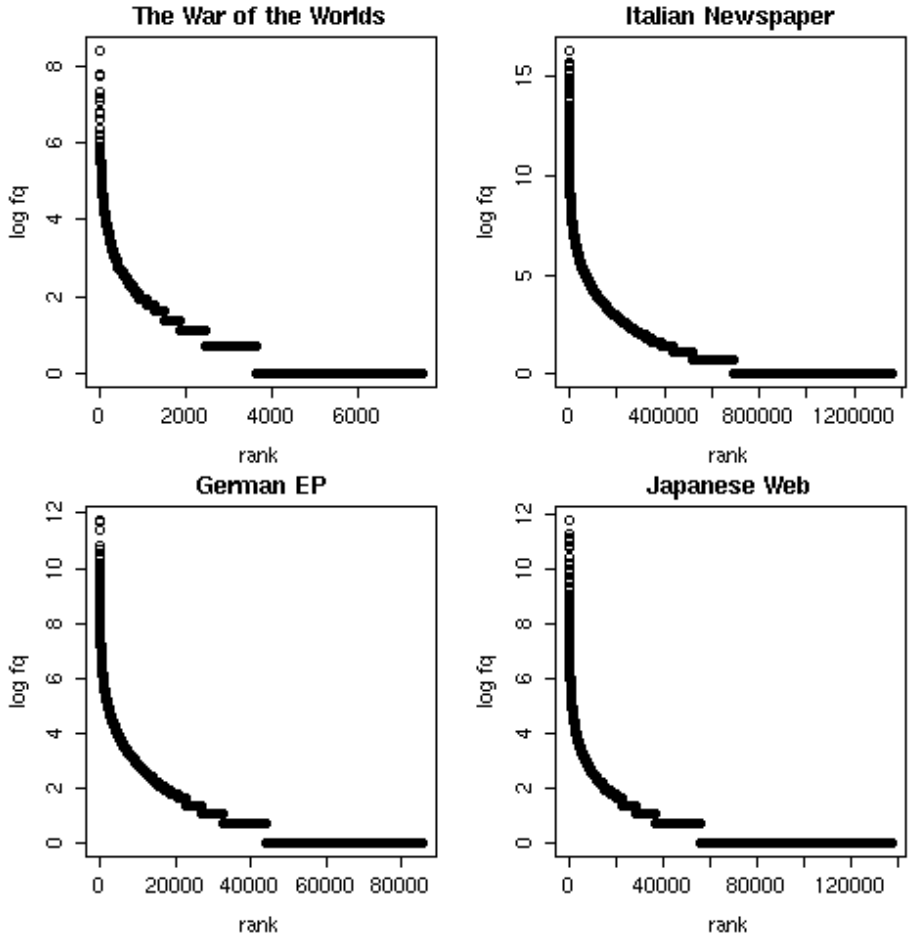
Text corpora

- Corpora are large electronic collections of texts produced in natural communicative settings
- Popular in language engineering, lexicography, language teaching/learning since (at least) the nineties
- Typology and some famous exemplars:
  - Balanced, representative, 'reference' corpora: Brown/LOB (1M tokens), COBUILD (10M, . . .), BNC (100M)
  - Opportunistic: WSJ, la Repubblica-SSLMIT, Gigaword (1B)
  - Web-derived corpora (WaCky project)
  - Parallel: Hansard, OPUS, EuroParl
  - Specialized (CHILDES), comparable, diachronic. . .

Corpora and the human experience

- Computational modelers in cognitive science (e.g., Rogers and McClelland 2004) typically work with hand-crafted input
- Corpora are “real”, natural input, akin to what humans hear/read, with same problems of noise and skewed input distribution (Zipf’s law) humans must face
- Computer seen as statistics-driven agent that “learns” from its environment: can it teach us something about human learning?

Zipf's law


The War of the Worlds
log fq
rank
0
2000
4000
6000
Italian Newspaper
log fq
rank
0
400000
800000
1200000
German EP
log fq
rank
0
20000
40000
60000
80000
Japanese Web
log fq
rank
0
40000
80000
120000

The contextual view meaning

- Acquisition/representation of meaning/conceptual knowledge is core issue in cognitive science
- Corpus-based simulations can help!

The contextual view meaning

- Acquisition/representation of meaning/conceptual knowledge is core issue in cognitive science
- Corpus-based simulations can help!
- “You should tell a word by the company it keeps” (Firth, 1957)
- “[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts” (Cruse, 1986)

Outline

Distributional semantics
Word space models (WSMs)

- Meaning of word/concept defined by *set of contexts* in which word occurs in corpus
- Similarity of words represented as *geometric distance* among *context vectors*

Distributional semantics

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark
park
owner
leash

Distributional semantics

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	+
park	
owner	
leash	

Distributional semantics

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	+
park	+
owner	
leash	

Distributional semantics

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	+
park	+
owner	+
leash	

Distributional semantics

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	+
park	+
owner	+
leash	+

Distributional semantics

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

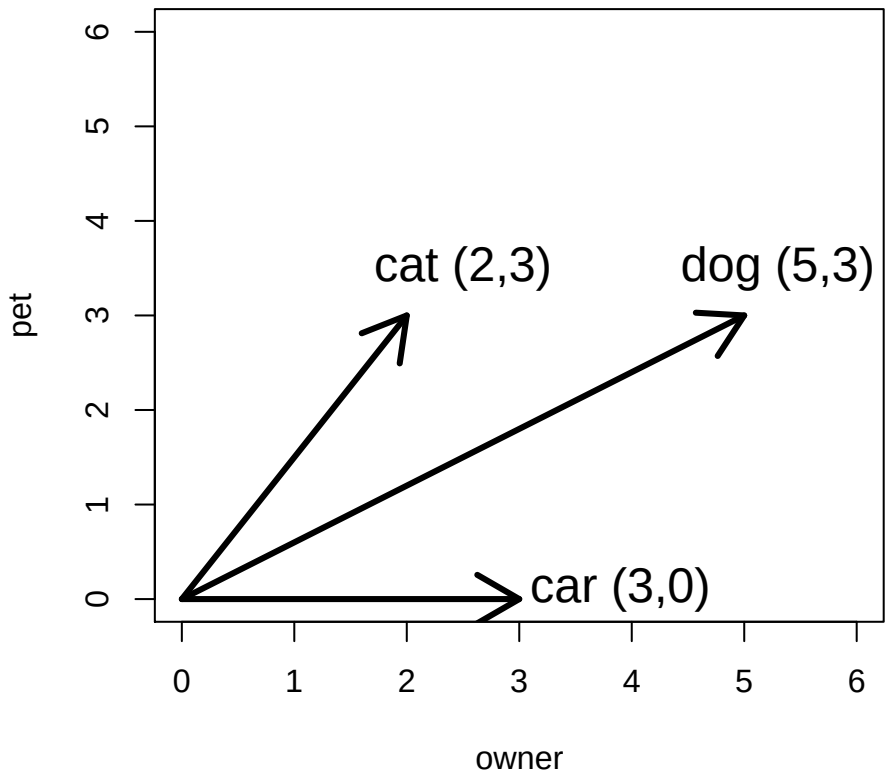
Distributional semantics

Meaning as co-occurrence

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

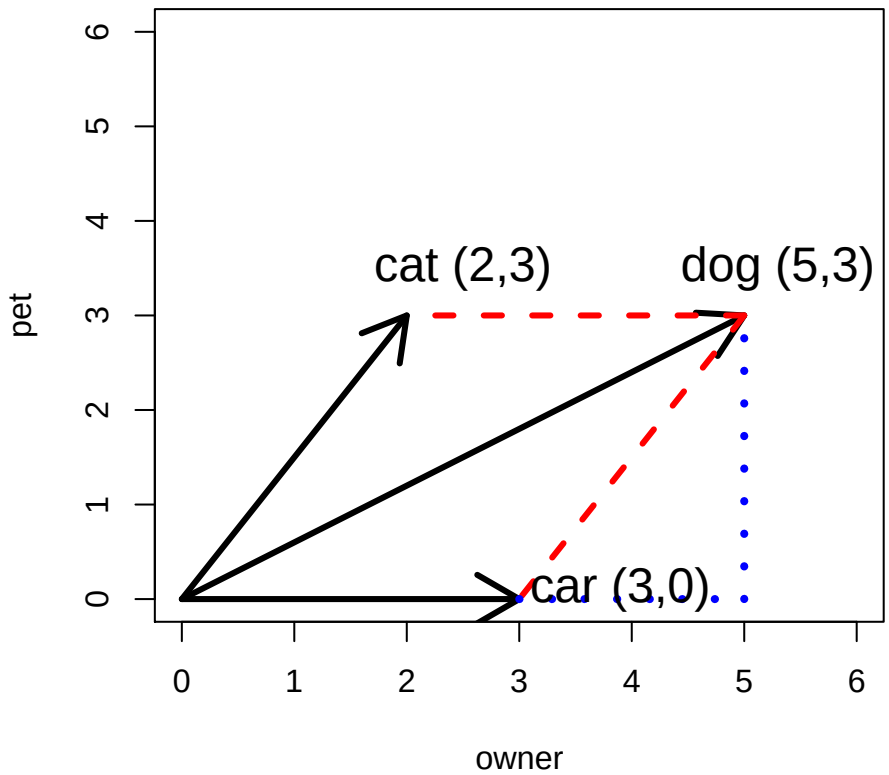
Distributional semantics

Similarity in word space



Distributional semantics

Similarity in word space



Distributional semantics
Which context?

- Documents/large textual spans
- All words in a narrow window
- Lemmatized content words in a narrow window
- Content words in specific syntactic constructions or specific surface patterns
- Context needs not be linguistic! Vectors could include, e.g., co-occurrence counts with sensory stimuli

Distributional semantics
Success in cognitive simulations

- synonym identification (Landauer and Dumais 1997)
- text coherence (Landauer and Dumais 1997)
- categorization (Burgess and Lund 1997)
- semantic priming (Lowe 2000, McDonald and Brew 2002, Vigliocco et al. 2004)
- substitution errors (Vigliocco et al. 2004)
- child lexicon acquisition (Li et al. 2004, Baroni et al. 2007)

Outline

The TOEFL synonym match task

- 80 items

The TOEFL synonym match task

- 80 items
- Target: *levied*
  Candidates: *imposed, believed, requested, correlated*

The TOEFL synonym match task

- 80 items
- Target: *levied*
  Candidates: *imposed*, *believed*, *requested*, *correlated*

Human performance on the synonym match task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: 97.75%

TOEFL results

- Humans:
  - Foreign test takers: 64.5%
  - Macquarie non-natives: 86.75%
  - Macquarie natives: 97.75%

TOEFL results

- Humans:
  - Foreign test takers: 64.5%
  - Macquarie non-natives: 86.75%
  - Macquarie natives: 97.75%
- Machines:
  - Rapp's 2003 SVD-based model trained on lemmatized BNC: 92.5%

Outline

The *flat* model of semantic similarity

- "Semantic similarity" is multi-faceted notion but a single WSM provides only one way to rank a set of words (according to distance measure of choice)
- Nearest neighbours of *motorcycle* in a standard WSM:
  - motor → component
  - car → co-hyponym
  - diesel → component?
  - to race → proper function
  - van → co-hyponym
  - bmw → hyponym
  - to park → proper function
  - vehicle → hypernym
  - engine → component
  - to steal → frame?

Murphy's objections

Murphy (2002), p. 429-430

- "Although *jugs* might be related to both *vinegar* and *bottles*, these relations are extremely different, and an overall similarity score does not represent these differences."
- In order to distinguish how jugs are related to vinegar from how they are related to bottles, one needs to know what are the *properties* of these concepts:
- "[S]ince one's concept of a jug, say, would include detailed information about its origins, parts, materials, functions and so on, the concept is more than sufficient to distinguish the meaning of *jugs* from that of *vinegar* and, for that matter, *bottles*."

Semantic relations in cognitive and applied tasks

- Property generation: humans can easily produce coherent lists of typical properties of concepts (*norms* of McRae et al., Vinson/Vigliocco and others)
- Humans are able to distinguish different *types* of relations between properties and concepts, e.g., between formal and functional properties
  - A *dish* looks like a *CD* but its function is more similar to that of a *bottle*

Semantic relations in cognitive and applied tasks

- Different relations must be extracted and identified for modeling semantic interpretation, e.g.:
  - Telic quale relation in type coercion: finish the book (*to read*) vs. the ice-cream (*to eat*) (Pustejovsky 1995)
  - Salient properties in compound interpretation: a *zebra cup* is a cup *with stripes*
  - Parts in co-reference bridging: The building faced a dark alley. A window opened
- Specific *types* of properties (e.g., visual vs. functional) play crucial role in neural organization of concepts and semantic deficits (Martin 2007, Vigliocco et al. 2004)
- Semantic relations needed in practical applications, in particular development of lexical resources

Outline

StruDEL

Structured Dimensions Extraction and Labeling

- Tries to build contextual vectors that represent *typed properties of concepts*

StruDEL

Structured Dimensions Extraction and Labeling

- Tries to build contextual vectors that represent *typed properties of concepts*
- Concept: *motorcycle*
- (Target) representation:
  - *for* traveling
  - *for* running
  - *is-a* vehicle
  - *has* motor
  - *has* two wheels
  - ...

Strategies for typed property extraction

- Automated identification of plausible concept-property *connectors*:
  - lice in a large number of dogs $\rightarrow$ YES
  - lice and leeches $\rightarrow$ NO

Strategies for typed property extraction

- Automated identification of plausible concept-property *connectors*:
  - lice in a large number of dogs → YES
  - lice and leeches → NO
- (Weighted) number of *distinct connectors* between concept and property is better indicator of true semantic relation than absolute co-occurrence frequency
  - year of the tiger is very frequent, but following are not attested: year of some tigers, the tigers have years, etc.
  - *Vice versa*, no tail/tiger connector is very frequent, but there are many of them: tail of the tiger, tail of some tigers, the tigers have tails, etc.

Strategies for typed property extraction

- Automated identification of plausible concept-property *connectors*:
  - lice in a large number of dogs → YES
  - lice and leeches → NO
- (Weighted) number of *distinct connectors* between concept and property is better indicator of true semantic relation than absolute co-occurrence frequency
  - year of the tiger is very frequent, but following are not attested: year of some tigers, the tigers have years, etc.
  - *Vice versa*, no tail/tiger connector is very frequent, but there are many of them: tail of the tiger, tail of some tigers, the tigers have tails, etc.
- The *type* of a relation can be extracted by generalizing over the connectors: of, with, of some, have point, together, to a *part/whole* relation

StruDEL baking

- Concept-property-type tuples extracted from ukWaC, a corpus of random Web pages including 2.25 billion tokens
- Property lists extracted for 1,234 (concrete) concepts
- Compared to various state-of-the-art WSMs, including SVD-based model and model using dependency parses (DV)

book

The StruDEL description

property	*type*	*LL*
to read	*verb* obj	3941.3
author	c *from* p	3772.8
to write	*verb* obj	2399.5
reader	c *for* p	2298.5
chapter	p *in* c	2259.8
library	c *in* p	2222.4
to publish	*verb* obj	1907.7
reading	p *from* c	1296.8
publisher	c *from* p	1258.0
review	p *about* c	1156.4

tiger

The StruDEL description

property	*type*	*LL*
jungle	c *from* p	132.2
cat	p *as* c	94.1
species	p *as* c	89.4
stripes	p *as* c	84.1
animal	p *as* c	75.6
to maul	subj *verb*	63.7
habitat	c *in* p	63.4
lion	p *as* c	56.0
to tame	*verb* obj	53.2
zoo	c *in* p	51.4

motorcycle

The StruDEL description

property	*type*	*LL*
to ride	*verb* obj	345.8
rider	p *on* c	199.8
vehicle	p *as* c	103.7
motorbike	p *for* c	100.2
street	c *on* p	71.3
to park	*verb* obj	69.3
scooter	p *over* c	51.6
car	p *as* c	45.7
to insure	p *for* c	39.8
bike	p *out* c	37.7

Outline

Outline

NORMS

McRae et al. 2005

- 725 participants rating 541 concepts, 30 subjects per concept
- Subjects produce list of properties that describe concept
- Manual normalization: *loud*, *noise*, *noisy* mapped to *is loud*
- NB: NORMS is a (conscious) *product* of human semantic representations, not a direct window into these representations
  - In other words, NORMS is useful comparison point, but not necessarily a “gold standard”

motorcycle

The NORMS description

property	*productions*
has wheels	22
has 2 wheels	20
is dangerous	18
has engine	13
is fast	12
used with helmet	11
made by Harley Davidson	10
is loud	10
used by 1 or 2 people	9
is a vehicle	9

Property analysis

- On average, NORMS and StruDEL share 2.4 of the 10 most salient properties of a concept
- No other distributional model we tested shares more than 1.5/10 properties with NORMS

Property analysis

- On average, NORMS and StruDEL share 2.4 of the 10 most salient properties of a concept
- No other distributional model we tested shares more than 1.5/10 properties with NORMS
- Systematic analysis of different property *types* privileged by NORMS vs. StruDEL (Baroni and Lenci 2008)
- E.g., for *car*:
  - Shared: *engine, gasoline, transportation*
  - NORMS only: *wheels, 4 wheels, doors, steering wheel, expensive, for passengers, vehicle*
  - StruDEL only: *it is driven, driver, it is parked, road, garage, race, parking*

Property types

Wu and Barsalou (Submitted), McRae et al. (2005), simplified

C **C**ategory: dog-animal, airplane-vehicle

P **P**arts: dog-tail, airplane-wing

Q **Q**ualities: dog-brown, airplane-fast

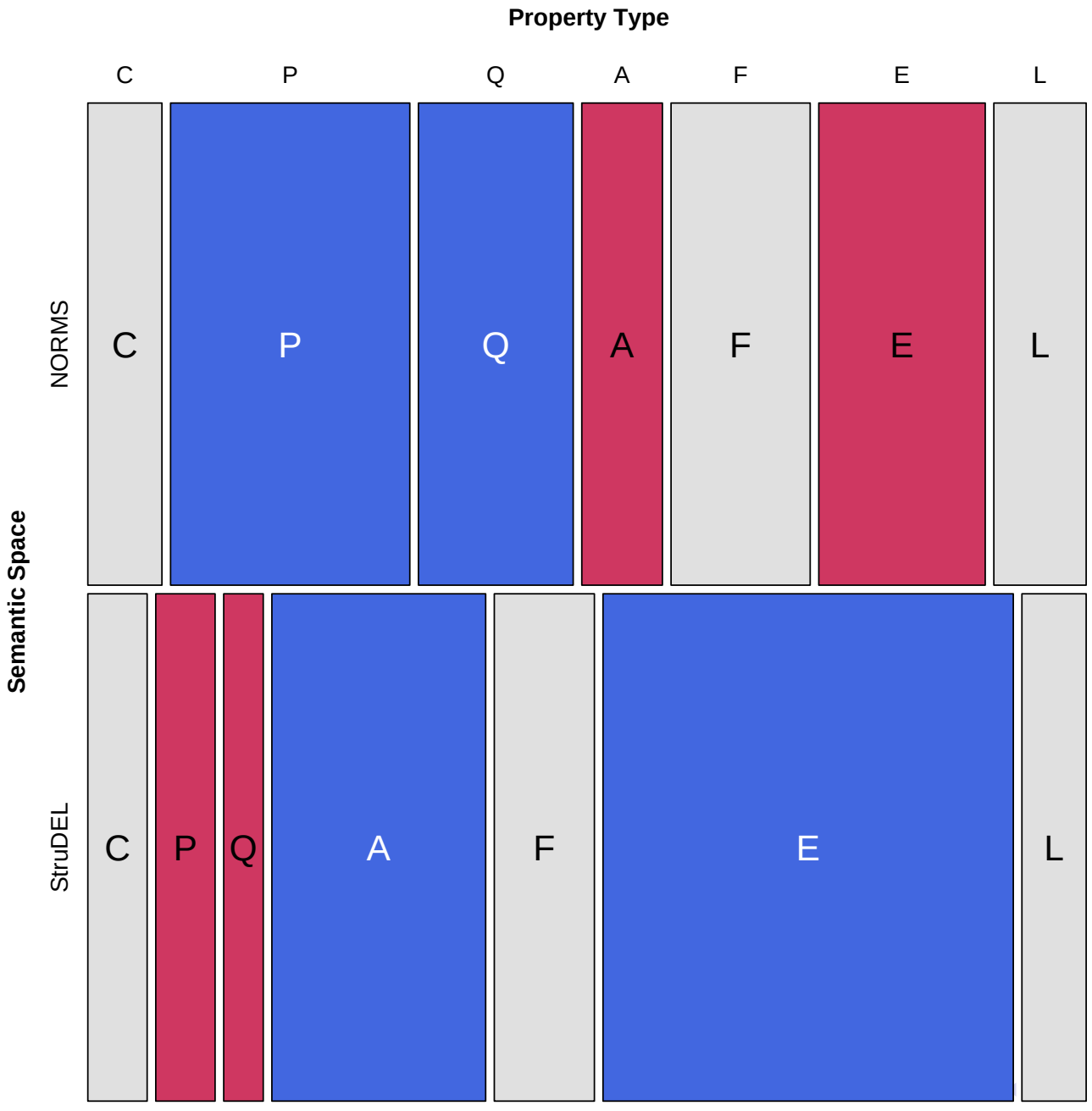
A Typical **A**ctivities and behaviours: dog-barks, airplane-flies

F **F**unction: dog-pet, dog-hunting, airplane-transportation

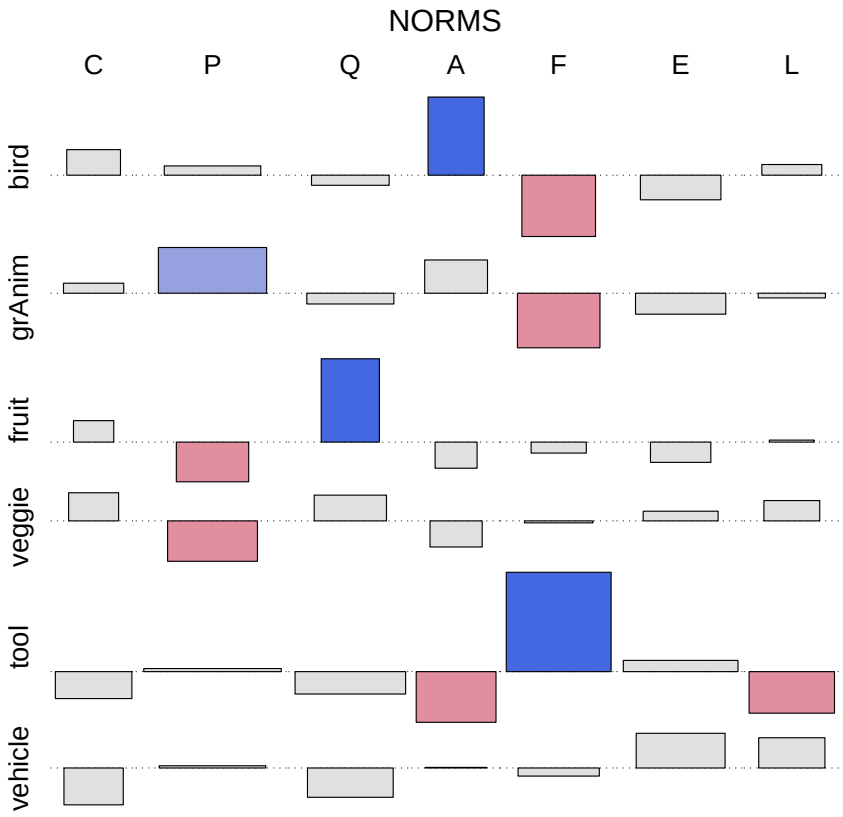
E Related **E**ntities: dog-cat, airplane-pilot

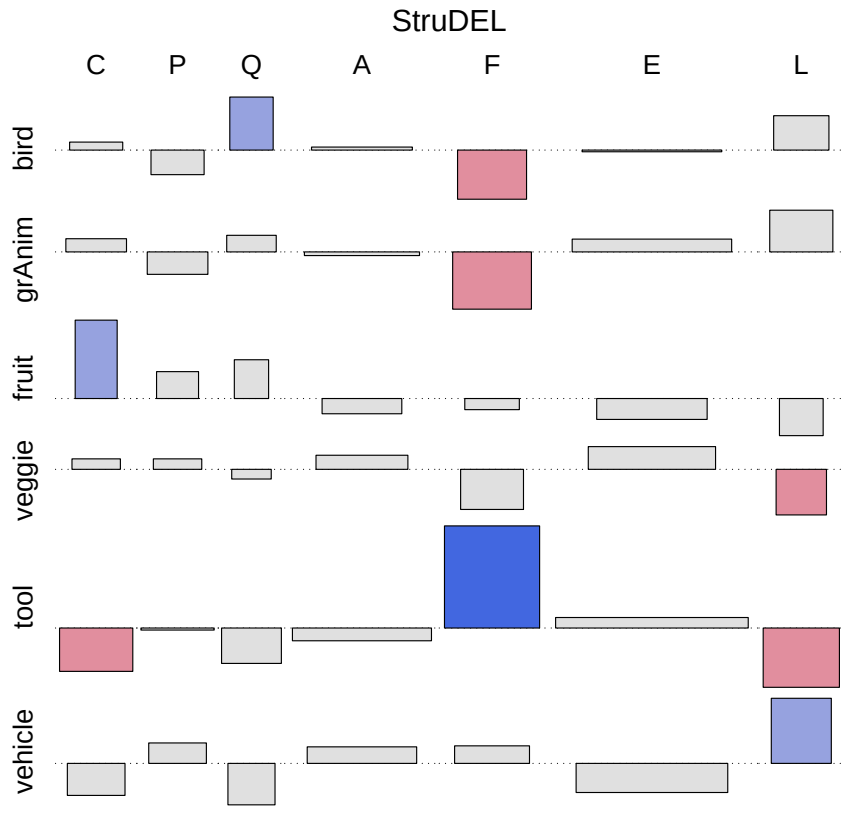
L **L**ocation: dog-kennel, airplane-sky

Property types in NORMS and StruDEL



Properties by categories





Outline

Data-set

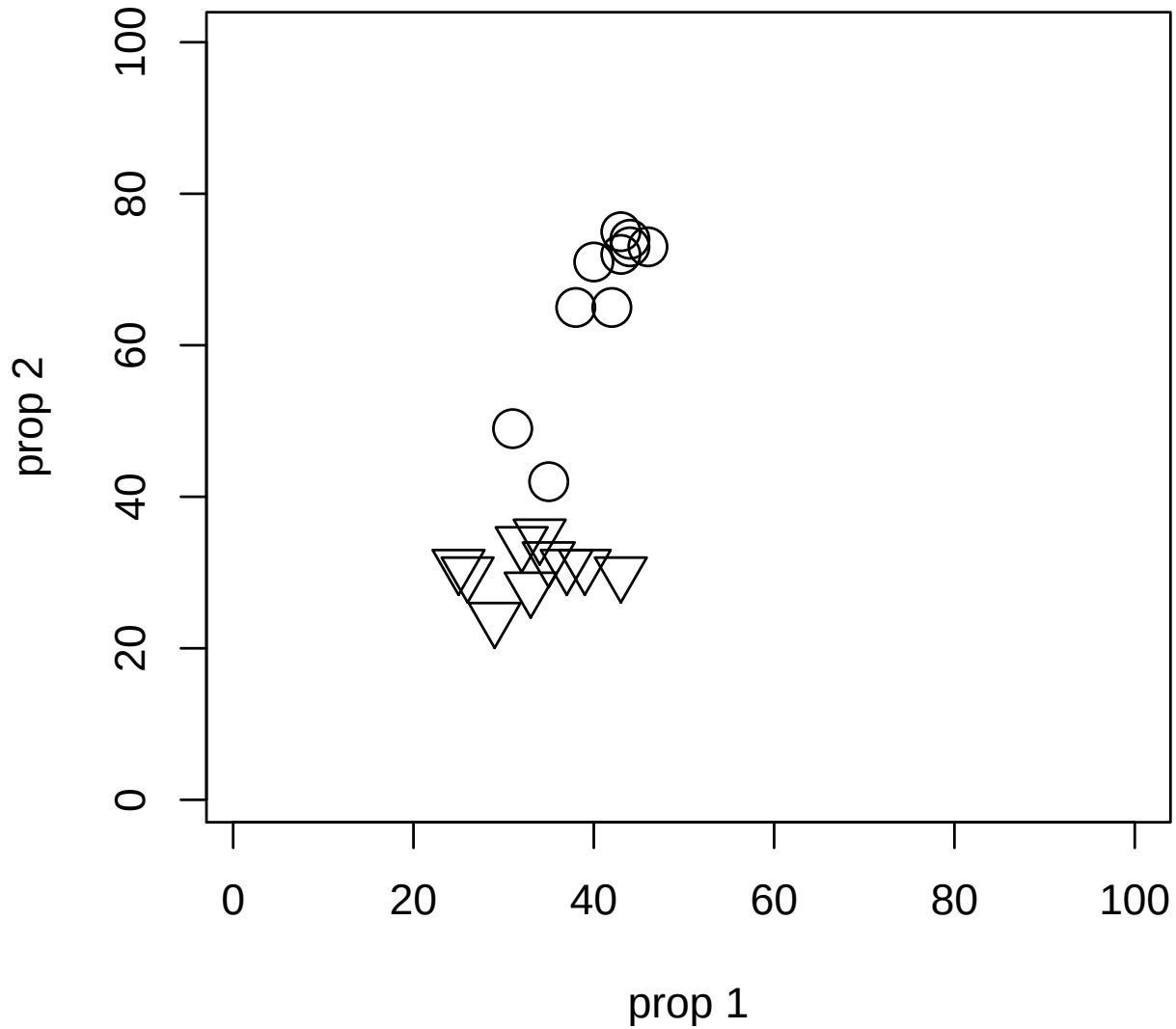
44 concrete concepts

- 24 natural concepts
  - 15 animals: 7 birds, 8 ground animals
  - 9 vegetables: 4 fruits, 5 greens
- 20 artifacts
  - 13 tools
  - 7 vehicles

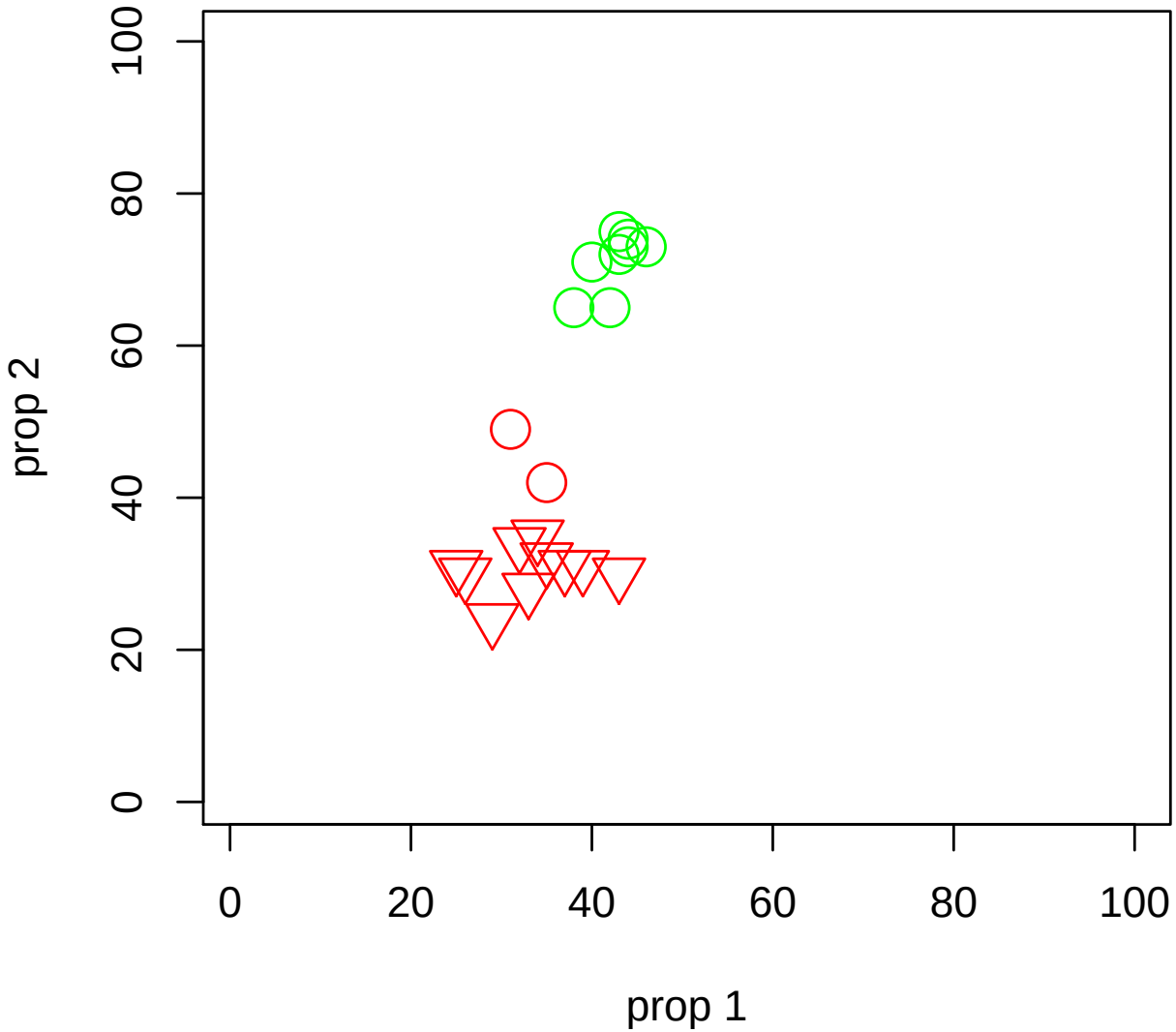
Hierarchical categorization

- 6-way: birds, ground animals, fruits, greens, tools, vehicles
- 3-way: animals, vegetables, man-made
- 2-way: natural, man-made

Categorization as clustering in semantic space



Categorization as clustering in semantic space



Results

Percentage *purity* of clusters

space	*6 categories*	*3 categories*	*2 categories*
NORMS	91	98	100
StruDEL	79	91	98
DV	73	89	95
SVD	79	75	59

“Animals”

Typical properties in the 3-way solution

NORMS

- animal
- legs
- beak
- eggs
- bird

StruDEL

- it breeds
- is seen
- is shot
- is rescued
- dies

“Vegetables”

Typical properties in the 3-way solution

NORMS

- vegetable
- sweet
- on trees
- fruit
- edible

StruDEL

- is sliced
- is minced
- is eaten
- it grows
- slice

“Man-made”

Typical properties in the 3-way solution

NORMS

- metal
- plastic
- handle
- for transportation
- wood

StruDEL

- is used
- in hands
- powered
- has use
- makes things

Outline

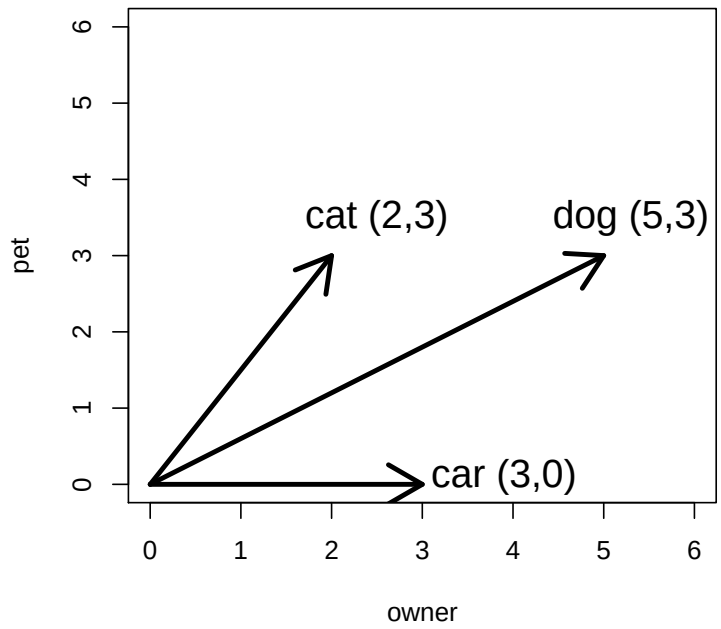
Other tasks

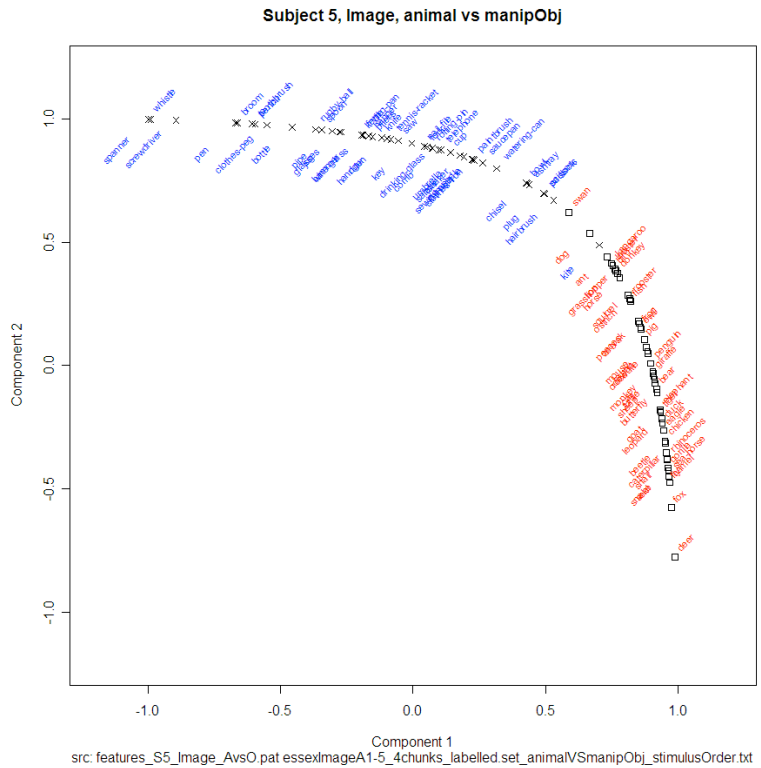
StruDEL is consistently best performer

- Predicting free association (*cat* . . . *dog*)
- Modeling prototypicality ratings (a *sparrow* is a more typical bird than a *penguin*)
- Generating specific properties (what is the typical *location* of hammers? what is their typical *function*?)

From word spaces to brain spaces

Work in progress





Outline

Bananas in adult and child-directed speech

BNC spoken-demographic section vs. Lara corpus

```
          my brother
  and I got dried <bananas> when war broke out
  three pounds of <bananas> please
  we could have a <banana>  souffle
             Those <bananas> are going a bit mm
                             well I like them when
                             they go soft

           who eats <bananas> ?
who do we have in
our zoo that eats <bananas> ?
        monkeys eat <bananas>
           shall we
give the monkey a <banana>  ?
        there's your <banana>
```

Hammers in adult and child-directed speech

BNC spoken-demographic section vs. Lara corpus

someone went
berserk with a <hammer> , that's been known
we had <hammer> drilled the blunt bit
I was tapping
it with a <hammer> wasn't I

she had a <hammer> and she was banging
on the wall
the <hammer> is the most useful tool,
Lara. Whenever the telly
goes wrong you
just hit it
with your <hammer>

The Human Experience

- We can gain useful insights about human conceptual acquisition applying fancy learning techniques to whatever corpora we have available
- But we suspect we will not make a major breakthrough until we learn from the same data humans learn from:
  - Corpora cue how adults transfer knowledge to other adults
  - Knowledge in corpora is not organized incrementally
  - Large corpora (still) lack multimodal information
- Corpora in CHILDES are too small, sparse, often record speech in special occasions
  - Lara, one of densest CHILDES corpora, contains transcripts of about 120 hours

The Human Experience

- We can gain useful insights about human conceptual acquisition applying fancy learning techniques to whatever corpora we have available
- But we suspect we will not make a major breakthrough until we learn from the same data humans learn from:
  - Corpora cue how adults transfer knowledge to other adults
  - Knowledge in corpora is not organized incrementally
  - Large corpora (still) lack multimodal information
- Corpora in CHILDES are too small, sparse, often record speech in special occasions
  - Lara, one of densest CHILDES corpora, contains transcripts of about 120 hours
- The Human Experience: record full verbal and visual experience of multiple children uninterruptedly for first 3 years of life

Some references

M. Baroni, A. Lenci and L. Onnis (2007). ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. Proceedings of ACL 2007: 49-56.
A. Cruse (1986). *Lexical semantics*. Cambridge: Cambridge University Press.
T.K. Landauer and S.T. Dumais (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2): 211-240.
T. Landauer, P. Foltz, and D. Laham (1998). An introduction to Latent Semantic Analysis. *Discourse Processes* **25**, 259-284.
P. Li, C. Burgess and K. Lund (2000). The acquisition of word meaning through global lexical co-occurrences. *Proceedings of the 31st Child Language Research Forum*: 167-178.
A. Martin (2007). The representation of object concepts in the brain. *Annual Review of Psychology 58*.
K. McRae, G. Creee, M. Seidenberg & C. McNorgan (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods 37*.

Some more references

G. Murphy. (2002). *The big book of concepts.* The MIT Press.
J. Pustejovsky (1995). *The generative lexicon*. MIT Press.
T. Rogers and J. McClelland (2004). *Semantic cognition: A parallel distributed processing approach*. The MIT Press.
C. Rowland, J. Pine, E. Lieven and A. Theakston (2005). The incidence of error in young children's wh-questions. *Journal of Speech, Language and Hearing Research*, 48(2): 384-404.
M. Sahlgren (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Stockholm University.
G. Vigliocco, D. Vinson, W. Lewis and M. Garrett (2004). Representing the meanings of objects and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48: 422-488.
L. Wu & L. Barsalou (Submitted). *Grounding concepts in perceptual simulation: I. Evidence from property generation.*