

LEXICON AND ONTOLOGY: THE VIEW FROM CONCEPT ACQUISITION

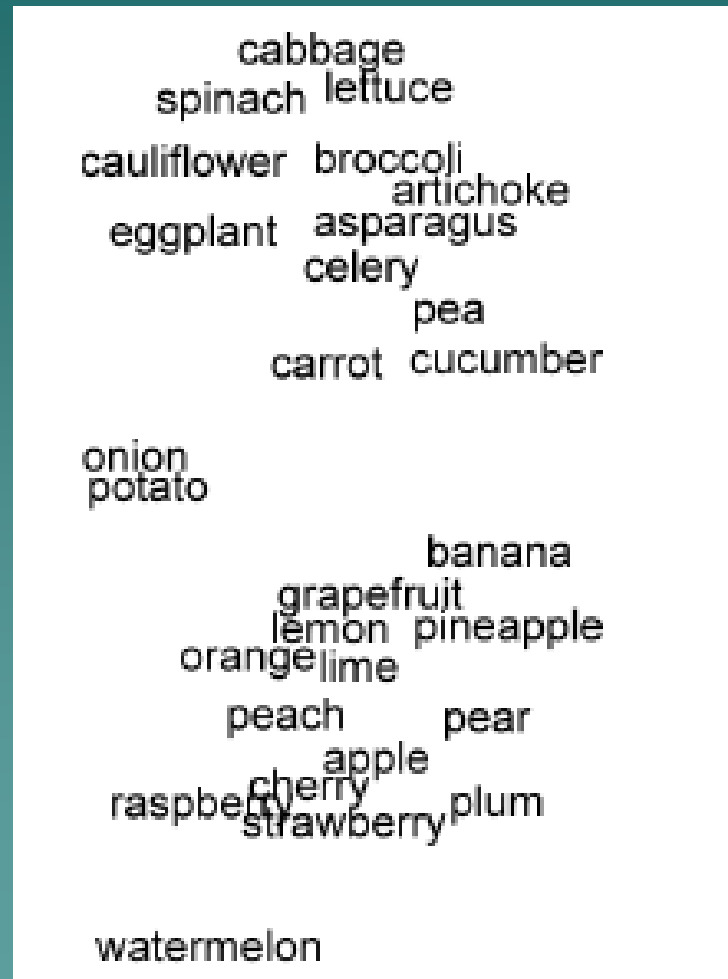
Massimo Poesio
Uni Trento, CIMEC / DISI

A stylized silhouette of a mountain range in a teal color, located at the bottom right of the slide.

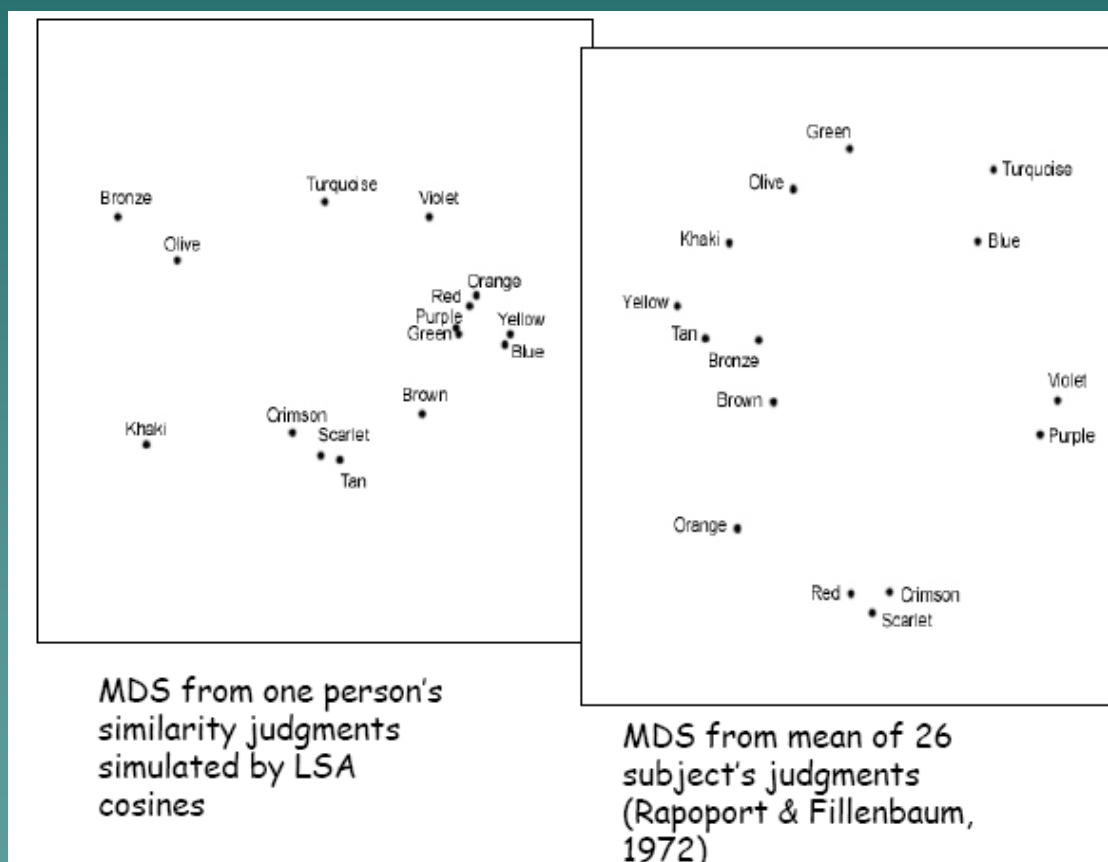
CONCEPT ACQUISITION

- ◆ Models for the acquisition of conceptual knowledge from data
- ◆ These models tend to be extremely simple, but behave well at certain tasks

THE VECTOR SPACE VIEW OF CONCEPTUAL KNOWLEDGE



STRENGTHS OF THIS APPROACH: CATEGORIZATION



CONCEPT ACQUISITION

- ◆ Models for the acquisition of conceptual knowledge from data
- ◆ These models tend to be extremely simple, but behave well at certain tasks
- ◆ We (me, Baroni, Lenci) and others (e.g. Cimiano, Schubert) attempt to build more complex models inspired by ideas from lexical and ontological research (Guarino, Vieu, Pustejovsky, etc)

TWO VIEWS OF VECTORS

WORDS

	car	hood	Chomsky	
0	1	0		car
0	0	0		Chomsky
0	0	1		corpus
0	1	0		emissions
1	1	0		engine
1	1	0		hood
0	1	0		make
0	1	0		model
0	0	1		noun
0	0	1		parsing
0	0	1		tagging
1	1	0		tires
1	1	0		truck
1	1	0		trunk
0	0	1		wonderful

‘collocational’

ATTRIBUTES

part	antler
part	haunches
quality	agility
quality	gentleness
activity	calving
activity	Leap
activity	barking
rel-obj	forest
rel-agent	hunter

‘relational’

CONCEPT ACQUISITION

- ◆ Models for the acquisition of conceptual knowledge from data
- ◆ These models tend to be extremely simple, but behave well at certain tasks
- ◆ We (me, Baroni, Lenci) and others (e.g. Cimiano, Schubert) attempt to build more complex models inspired by ideas from lexical and ontological research (Guarino, Vieu, Pustejovsky, etc)
- ◆ Indications that these models behave better at certain tasks (Poesio & Almuhareb 2005, Baroni Barbu Murphy & Poesio to appear)

CLUSTERING WITH QUALIA-INSPIRED CONCEPT DESCRIPTIONS

	All Candidate Attributes	Heuristic filtering	Filtering by classification
Purity	0.657	0.672	0.693
Entropy	0.335	0.319	0.302
Vector size	24,178	4,296	3,824
Clustered Concepts	402	402	401

CONCEPT ACQUISITION

- ◆ Models for the acquisition of conceptual knowledge from data
- ◆ These models tend to be extremely simple, but behave well at certain tasks
- ◆ Attempt to build more complex models inspired by ideas from lexical and ontological research
- ◆ This attempt raises evaluation issues

EVALUATION: PROBLEMS

◆ Attribute extraction

- WordNet only contains ISA and PART attributes (more in glosses, see Barbu & Poesio 2008)
- SIMPLE could be part of the solution

◆ Categorical distinctions

- The WordNet category structure is highly subjective

CLUSTERING: ERROR ANALYSIS

ANIMAL	bear, bull, camel, cat, cow, deer, dog, elephant, horse, kitten, lion, monkey, puppy, rat, sheep, tiger, turtle
--------	---

CLUSTERING: ERROR ANALYSIS

ANIMAL	bear, bull, camel, cat, cow, deer, dog, elephant, horse, kitten, lion, monkey, puppy, rat, sheep, tiger, turtle
EDIBLE FRUIT	apple, banana, berry, cherry, fig, grape, kiwi, lemon, lime, mango, melon, olive, orange, peach, pear, pineapple, strawberry, watermelon, (pistachio, oyster)

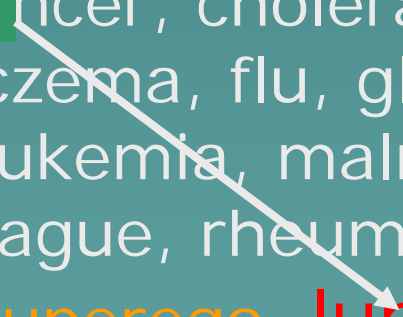
CLUSTERING: ERROR ANALYSIS

ANIMAL	bear, bull, camel, cat, cow, deer, dog, elephant, horse, kitten, lion, monkey, puppy, rat, sheep, tiger, turtle
EDIBLE FRUIT	apple, banana, berry, cherry, fig, grape, kiwi, lemon, lime, mango, melon, olive, orange, peach, pear, pineapple, strawberry, watermelon, (pistachio, oyster)
ILLNESS	acne, anthrax, arthritis, asthma, cancer, cholera, cirrhosis, diabetes, eczema, flu, glaucoma, hepatitis, leukemia, malnutrition, meningitis, plague, rheumatism, smallpox, (superego, lumbago, neuralgia, sciatica, gestation, menopause, quaternary, pain)

CLUSTERING: ERROR ANALYSIS

ANIMAL	bear, bull, camel, cat, cow, deer, dog, elephant, horse, kitten, lion, monkey, puppy, rat, sheep, tiger, turtle
EDIBLE FRUIT	apple, banana, berry, cherry, fig, grape, kiwi, lemon, lime, mango, melon, olive, orange, peach, pear, pineapple, strawberry, watermelon, (pistachio, oyster)
ILLNESS	acne, anthrax, arthritis, asthma, cancer, cholera, cirrhosis, diabetes, eczema, flu, glaucoma, hepatitis, leukemia, malnutrition, meningitis, plague, rheumatism, smallpox, (superego, lumbago, neuralgia, sciatica, gestation, menopause, quaternary, pain)

IN WORDNET: PAIN



USE EMPIRICAL DATA FROM ELSEWHERE

- ◆ Research on conceptual knowledge is carried out in Neural Science and Psychology as well
- ◆ The number of people working on concepts in these disciplines much greater
 - 1:10? 1:100?
- ◆ But there is limited interchange between CL and the other disciplines studying concepts
 - Except indirectly through the use of WordNet

EVIDENCE FROM OTHER AREAS OF COGNITIVE SCIENCE

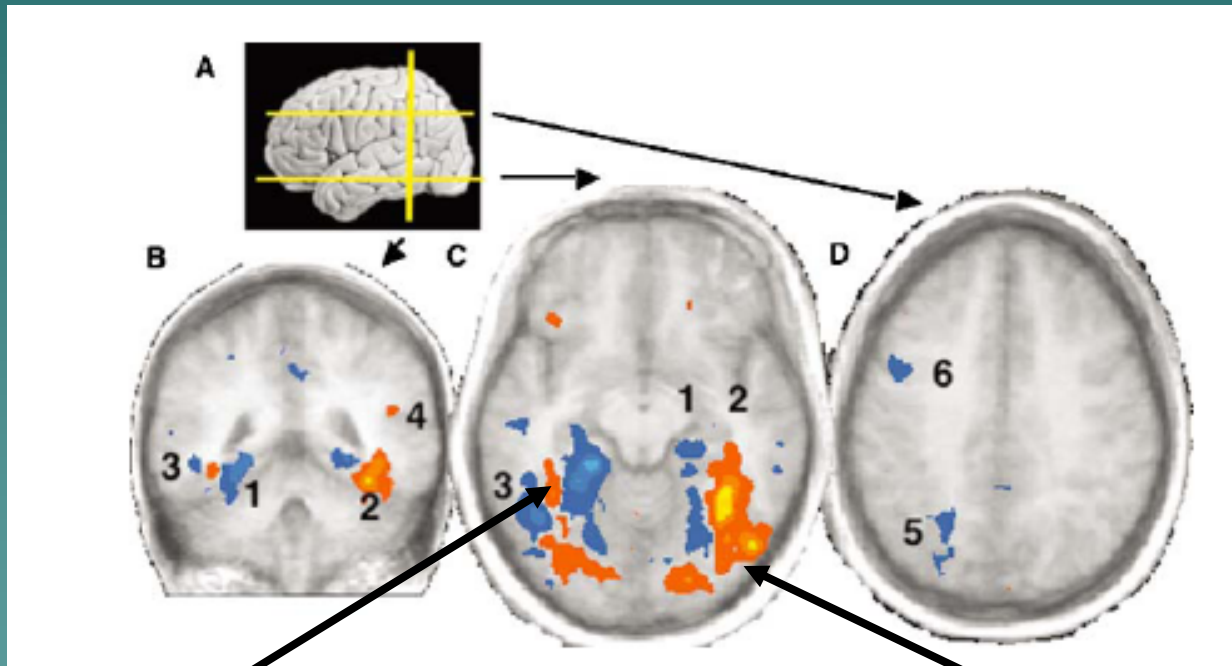
- ◆ Attributes: evidence from psychology
 - Association lists (priming)
 - ◆ E.g., use results of association tests to evaluate proximity (Lund et al, 1995; Pado and Lapata, 2008)
 - ◆ Comparison against feature norms: Schulte im Walde, 2008)
 - Feature norms
- ◆ Category distinctions: evidence from neural science

SPEAKER-GENERATED FEATURES (VINSON AND VIGLIOCCO)

the strawberry

- red (20)
- fruit (18)
- sweet (13)
- has seeds (12)
- grows (10)
- small (6)
- taste (6)
- food (5)
- from garden (5)
- juice (5)
- dessert (3)
- eat (3)...

CATEGORY DISTINCTIONS IN THE BRAIN



TOOLS

ANIMALS

PROBLEMS:

- ◆ Feature norms:
 - Great differences between the existing databases
- ◆ Brain data:
 - Much of this work tested very simple / very dubious distinctions
 - ◆ 'TOOLS': screwdriver / umbrella / pen / handkerchief

Problem: differences between feature norms

◆ motorcycle

– Vinson & Vigliocco:

◆ wheel, motor, loud, vehicle, wheel, fast, handle, ride, transport, bike, human, danger, noise, seat, brake, drive, fun, gas, machine, object, open, small, travel, wind

– Garrard et al:

◆ vehicle, wheel, fast, handlebar, light, seat, make a noise, tank, metal, unstable, tyre, coloured, sidecar, indicator, pannier, pedal, speedometer, manoeuvrable, race, brakes, stop, move, engine, petrol, economical, gears

– McRae et al:

◆ wheels, 2_wheels, dangerous, engine, fast, helmets, Harley_Davidson, loud, 1_or_2_people, vehicle, leather, transportation, 2_people, fun, Hell's_Angels, gasoline

– Mutual correlation of ranks ranges from 0.4 to 0.7

THE HOPE

- ◆ These studies won't be directly helpful to us, but we could adopt and improve the methodology
 - see e.g., Murphy et al for categorical distinctions, Baroni & Kremer 2008 for feature norms
- ◆ A science of 'empirical lexical studies' (empirical ontology?) in which empirical data from neural science / psychology / corpora feed into lexical construction (and viceversa)