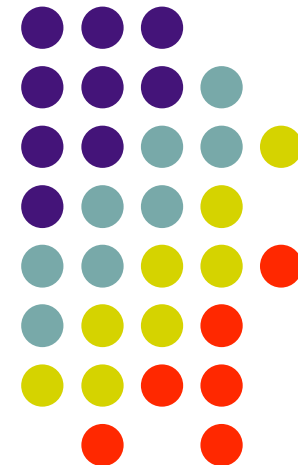


Distributional Models of Concepts and Properties

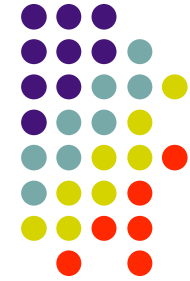


Alessandro Lenci
Università di Pisa – Dipartimento di Linguistica “T. Bolelli”

alessandro.lenci@ilc.cnr.it



Meanings and word spaces

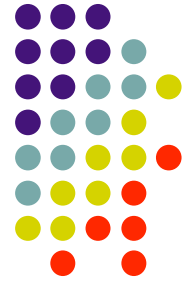


- **Word Space Models (WSMs)** represent meanings as points in a **high dimensional vector space**
 - words are associated with **distributed representations** induced from textual input and recording their global co-occurrence history
- WSMs are claimed to be able to capture a variety of facts about **human semantic learning, processing, and representation** (Landauer & Dumais 1997, McDonald & Brew 2004, Li *et al.* 2004, etc.)

“The dimensionality--optimizing method offers a promising solution to the ancient puzzle of **human knowledge induction**. It still remains to determine how wide its scope is among human learning and cognition phenomena. [...] We would suggest that applications to problems in conditioning, association, pattern and object recognition, contextual disambiguation, metaphor, concepts and categorization, reminding, casebased reasoning, probability and similarity judgment, and complex stimulus generalization are among the set where this kind of induction might provide new solutions”

(Landauer & Dumais 1997: 235)

Meanings and word spaces



- WSMs capture **distributional correlations** between words
 - close words in space have similar distributions in linguistic contexts
- WSMs “become semantic spaces” via the **Distributional Hypothesis** (Harris 1968, Miller & Charles 1991)
 - “Two words are semantically similar to the extent that their contextual representations are similar”
 - **Word Contextual Representation**
 - “An abstraction of information in the set of natural linguistic contexts in which a word occurs” [Charles 2000]

Word Space Models



- **Models of the mental lexicon**
 - lexical priming (Lund *et al.*, 1995), analogical reasoning (Ramscar and Yarlett, 2003), semantic similarity judgements (MacDonald & Ramscar, 2001), semantic deficits (Vigliocco, *et al.* 2004), etc.
- **Natural Language Processing (NLP)**
 - word sense disambiguation, Information Retrieval, ontology and thesauri learning (Lin 1998, Maedche & Staab 2004, Widdows 2003), etc.

Word space models

- **Latent Semantic Analysis (LSA)** (Landauer & Dumais 1997)
- **Hyperspace Analogue to Language (HAL)** (Burgess & Lund 1997)
- **Random Indexing (RI)** (Karlsgren & Sahlgren 2001)
- **Incremental Semantic Analysis (ISA)** (Baroni, Lenci & Onnis 2007)
- **Tupleware** (Baroni & Lenci in preparation)

Word Space Models

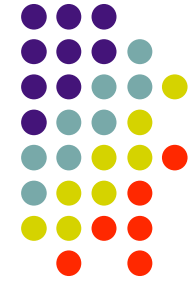
Lowe (2001), Padó & Lapata (2007)



- WSMs are formally defined by the quadruple $\langle T, B, M, S \rangle$
 - **T** is the set of “target” items
 - the elements to which the space provides a semantic representation
 - **B** is the **basis** that defines the space dimensions
 - **linguistic contexts** used to compute the distributional similarity
 - **M** is a co-occurrence matrix
 - cells contain some function of the **co-occurrence frequency** of targets with the basis contexts
 - **S** is some **measure of the distance** between points in space
 - e.g. cosine

Word Space Models

the co-occurrence matrix



- **Rows** correspond to the target words
- **Columns** correspond to the contexts defined by the basis B
 - different type of contexts
 - sliding window of adjacent words, syntactic relations, etc.
- **Cell values** correspond to (some function of) the **co-occurrence frequency** of a word in a certain context

	say	eat	open	think	country	sweet	...
king	6	2	5	4	1	0	
president	10	3	2	3	7	0	
cake	0	4	2	0	0	3	
banana	0	7	0	0	0	1	
...							

Testing distributional similarity

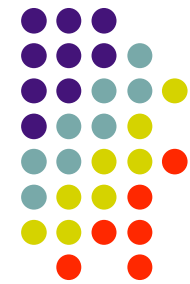
synonym identification



- Standard test set
 - synonym section in TOEFL (Test of English as a Foreign Language)
 - 80 items for which subjects must select the correct synonym among 4 candidates
 - target = furnish {supply, impress, protect, advise}
 - target = physician {chemist, pharmacist, nurse, doctor}
- Landauer & Dumais (1997)
 - LSA: 64.4% TOEFL participants: 64.5%
- Rapp (2004)
 - LSA: 92% Macquarie University (non-native): 86.75%
Macquarie University (native): 97.75%

Testing distributional similarity

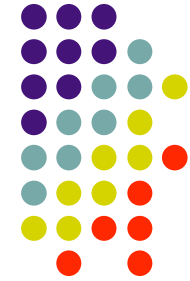
(in collaboration with Marco Baroni, CIMEC)



- “Cognitively plausible” semantic tasks
 - Concrete Noun Categorization
 - Abstract/Concrete Noun Discrimination
 - Property Generation
- Distributional model
 - Infomap (Stanford University, <http://infomap.stanford.edu>)
 - variant of Latent Semantic Analysis (LSA)
- Training corpus
 - British National Corpus (BNC)
 - 100 millions of tokens
 - corpus is lemmatized and PoS tagged
 - wordspace basis: 2K most frequent N, V, A and ADVs

1st Experiment

Concrete Noun Categorization



- In **categorization tasks**, subjects are typically asked to assign experimental items - objects, images, words - to a given category or to group together items belonging to the same category
 - categorization presupposes an understanding of the relationship between the items in a category
- Task 1- **Concrete Noun Categorization**
 - **goal**
 - evaluate whether semantic categories can emerge out of distributional similarities
 - *do close words in distributional space belong to the same semantic class?*
 - **data set**
 - **44** concrete nouns belonging to **6** semantic classes
 - subset of **McRae et al. (2005) Semantic Norms**
 - **method**
 - **Self-Organizing Maps (SOMs)** are applied to build semantic categories from the distributional vectors generated with LSA

1st Experiment

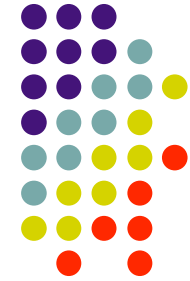
data set



chicken	bird-animal-natural
eagle	bird-animal-natural
duck	bird-animal-natural
swan	bird-animal-natural
owl	bird-animal-natural
penguin	bird-animal-natural
peacock	bird-animal-natural
dog	groundAnimal-animal-natural
elephant	groundAnimal-animal-natural
cow	groundAnimal-animal-natural
cat	groundAnimal-animal-natural
lion	groundAnimal-animal-natural
pig	groundAnimal-animal-natural
snail	groundAnimal-animal-natural
turtle	groundAnimal-animal-natural
cherry	fruitTree-vegetable-natural
banana	fruitTree-vegetable-natural
pear	fruitTree-vegetable-natural
pineapple	fruitTree-vegetable-natural
mushroom	green-vegetable-natural
corn	green-vegetable-natural

lettuce	green-vegetable-natural
potato	green-vegetable-natural
onion	green-vegetable-natural
bottle	tool-artifact
pencil	tool-artifact
pen	tool-artifact
cup	tool-artifact
bowl	tool-artifact
scissors	tool-artifact
kettle	tool-artifact
knife	tool-artifact
screwdriver	tool-artifact
hammer	tool-artifact
spoon	tool-artifact
chisel	tool-artifact
telephone	tool-artifact
boat	vehicle-artifact
car	vehicle-artifact
ship	vehicle-artifact
truck	vehicle-artifact
rocket	vehicle-artifact
motorcycle	vehicle-artifact
helicopter	vehicle-artifact

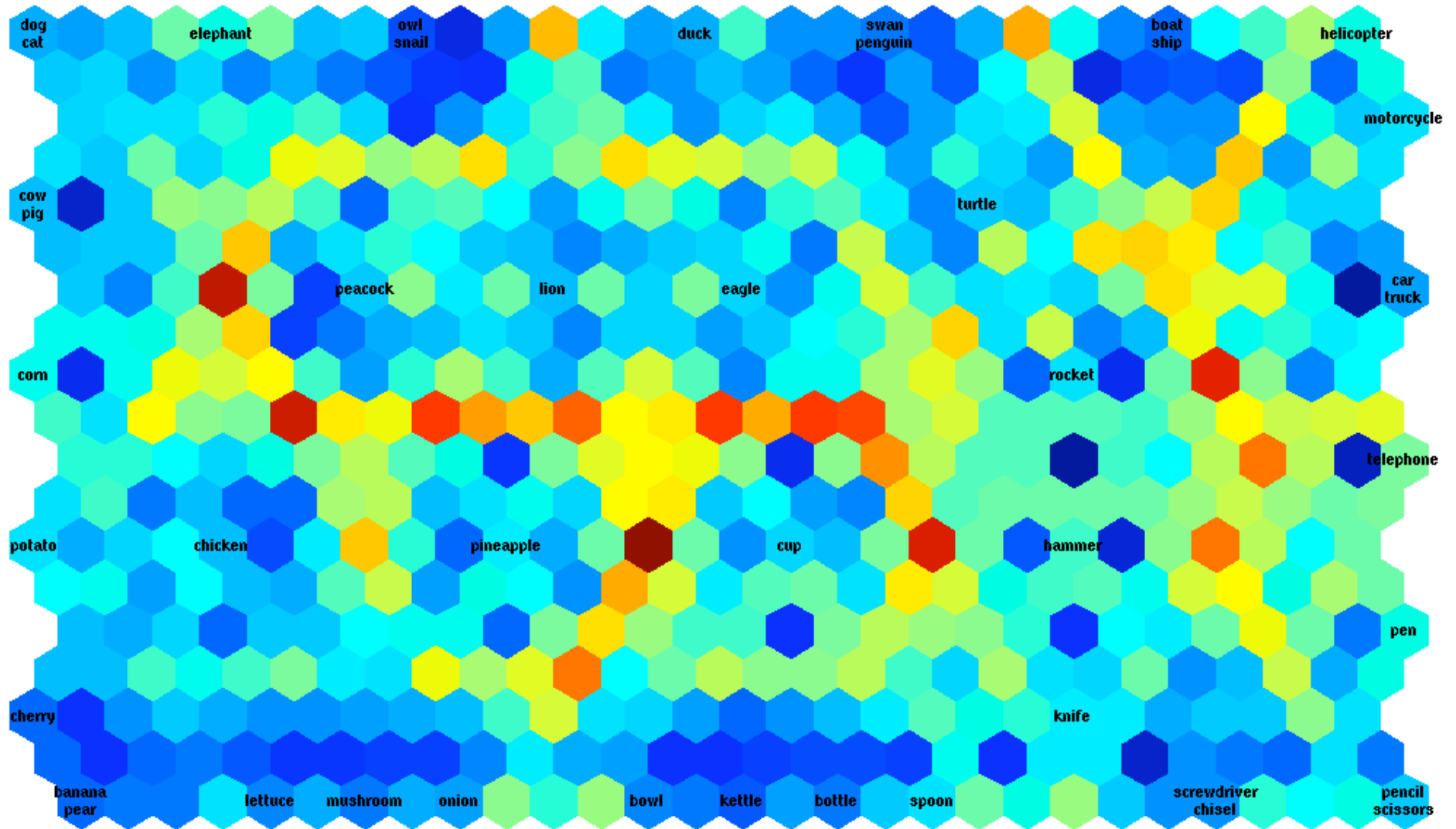
Self-Organizing Maps (SOMs)



- Self-Organized Maps
 - unsupervised neural networks based on “competitive learning” (Kohonen 2001)
 - applied to model **verb meaning learning** (Farkas & Li 2002), **lexical development** (Li *et al.* 2004), **semantic memory** (Vigliocco *et al.* 2004)
- Methodology
 - continuous classification on a topological basis
 - **vector similarity relations are turned into relations of spatial contiguity on the SOM**
 - contiguous clusters on the map can be interpreted as semantic “proto-categories”

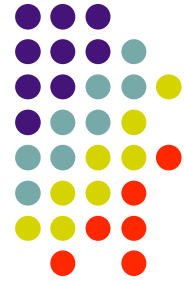
1st Experiment

Concrete Noun Categorization



2nd Experiment

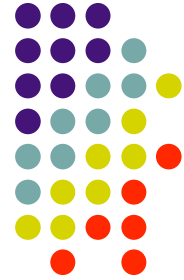
Abstract/Concrete Noun Discrimination



- Behavioral and neuropsychological evidence suggests that **abstract** and **concrete** concepts might be represented, retrieved and processed differently in the human brain
 - cf. Noppeney, U. and C. Price (2004)
- Task 2 - **Abstract/Concrete Noun Discrimination**
 - **goal**
 - evaluate the ability of WSMs to model the contrast between concrete and abstract nouns
 - **data set**
 - 40 nouns extracted from the **Medical Research Council (MRC) Psycholinguistic Database**, divided into “concreteness classes”
 - CONC index in MRC summarizes the subjects’ judgment about noun concreteness
 - **three classes**
 - **HI** - nouns with the highest CONC index in MRC
 - subset of the nouns of Experiment 1
 - **LO** - nouns with the lowest CONC index in MRC
 - **ME** - nouns whose CONC index is close to the average CONC value in MRC

2nd Experiment

data set (MRC Psycholinguistic Database)

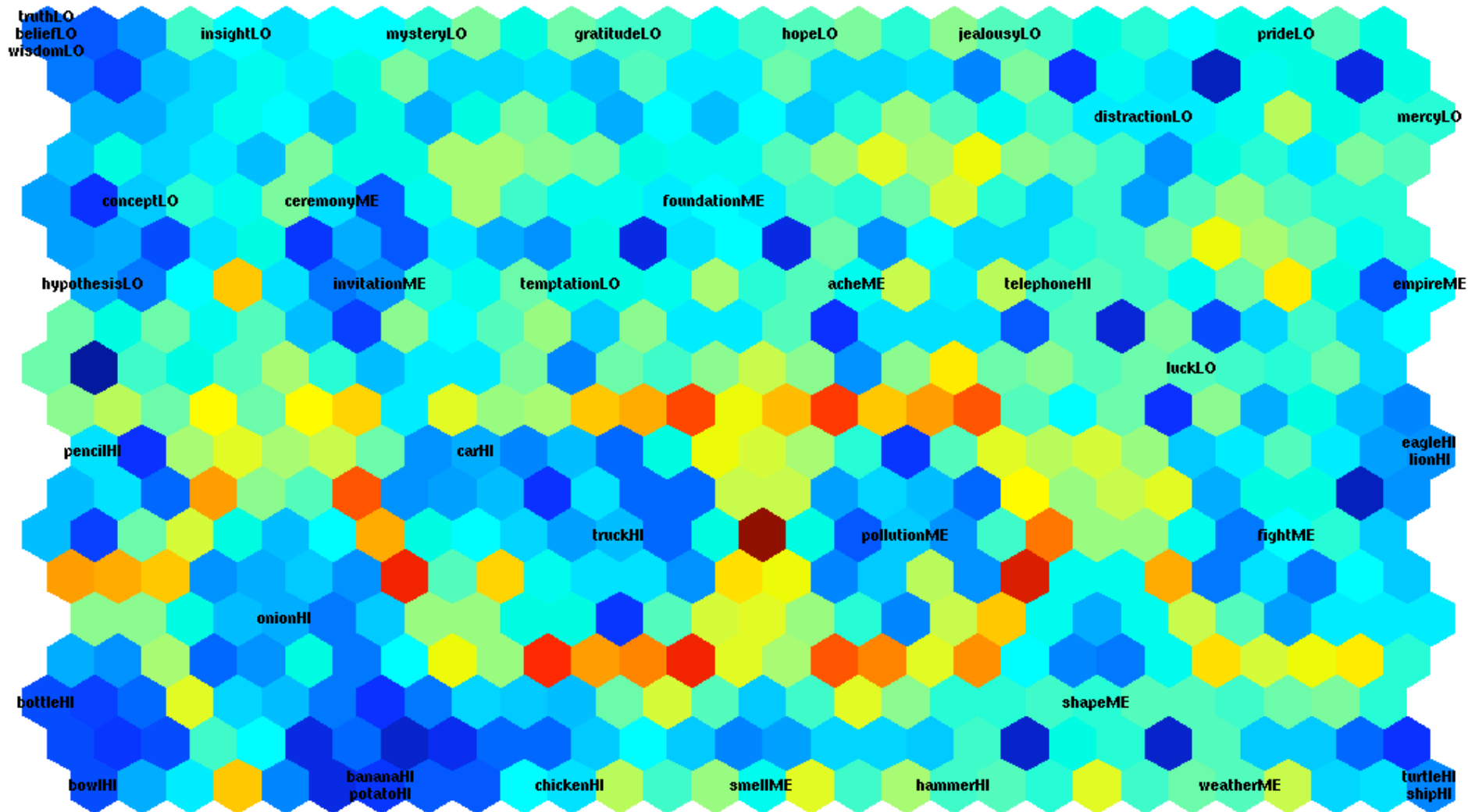
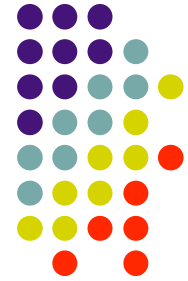


NOUN	CONC	CLASS
chicken	614	HI
eagle	616	HI
lion	627	HI
turtle	644	HI
banana	633	HI
onion	632	HI
potato	629	HI
bowl	575	HI
pencil	617	HI
telephone	619	HI
truck	595	HI
ship	615	HI
car	622	HI
bottle	591	HI
hammer	605	HI
pollution	463	ME
invitation	439	ME
shape	452	ME
empire	429	ME
foundation	462	ME

NOUN	CONC	CLASS
fight	455	ME
smell	450	ME
ache	443	ME
ceremony	430	ME
weather	439	ME
jealousy	250	LO
truth	261	LO
hypothesis	261	LO
hope	261	LO
mercy	239	LO
mystery	256	LO
gratitude	239	LO
concept	264	LO
temptation	265	LO
pride	270	LO
belief	270	LO
insight	270	LO
wisdom	275	LO
luck	275	LO
distraction	289	LO

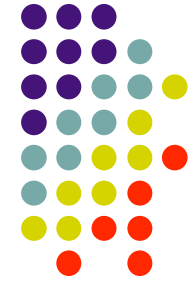
2nd Experiment

Abstract/Concrete Noun Discrimination



3rd Experiment

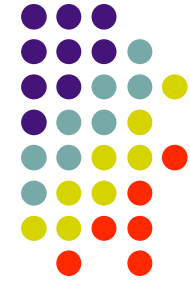
Property Generation (Baroni & Lenci in press)



- The ability to describe a concept in terms of its **salient properties** is an important feature of human conceptual cognition
 - *dog* → **is_an_animal**, **has_four_legs**, **barks**, etc.
- Task 3 – **Property Generation**
 - **goal**
 - compare the type of human-generated **properties** collected by psychologists to the types of properties generated by WSMs
 - **data set**
 - 44 concrete nouns (same as Experiment 1)
 - **properties produced by human subjects**
 - extracted from **McRae et al. (2005) Semantic Norms**

Meanings in word spaces

the problem of structure

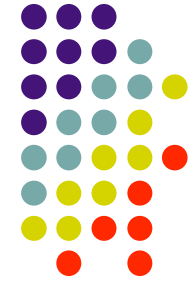


- Semantic representations in WSM lack **internal** and **external structure**
 - word meanings are represented as **points** in WSMs
 - the internal dimensions of distributional vectors are not semantically loaded
 - “Meaning [...] is a relation among words. In such a relational system, one cannot talk about the meaning of a word in isolation; words have meanings only in virtue of their relations to other words – meaning is a property of the system as a whole” [Kintsch 2007: 91]
 - relations among words are only defined in quantitative terms
 - WSMs can only express how close two words are, but they can not express the **type of relation** linking them

car	
van:0.755437	co-hyponym
driver:0.728991	agent
vehicle:0.708992	hyperonym
park:0.699351	event
motorist:0.692586	agent
motor:0.686240	meronym
lorry:0.661483	co-hyponym

Semantic Feature Norms

McRae et al. (2005)



- Lists of the properties that subjects consider important to describe a concept
 - each concept (expressed by a noun) is described by 20 subjects

	CAR	
a_vehicle	superordinate	9
causes_pollution	contingency	8
different_colours	external_surface_property	6
has_4_doors	external_component	5
has_4_wheels	external_component	18
has_a_steering_wheel	internal_component	12
has_an_engine	internal_component	13
has_doors	external_component	13
has_wheels	external_component	19
has_windows	external_component	6
is_expensive	systemic_property	11
is_fast	systemic_property	9
requires_drivers	contingency	7
requires_gasoline	contingency	12
used_for_passengers	participant	9
used_for_transportation	function	19

*feature (property)
production
frequency*

Properties in Word Spaces



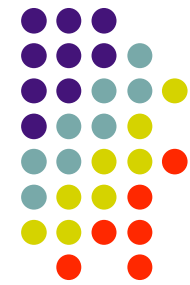
- The 10 nearest neighbors in the word space were selected as the properties associated to a target word

concept	property	semantic type	cosine
<i>car</i>	<i>van</i>	co-hyponym	0.75
	<i>driver</i>	participant	0.73
	<i>vehicle</i>	hyperonym	0.71
	<i>park</i>	action	0.70

- Each property was classified with its **semantic type**
 - cf. Wu & Barsalou (submitted) **taxonomy of property types** (the same used in the McRae Norms)
- Comparison between the human-generated properties and the properties generated by LSA was carried out at the level of their **semantic type**

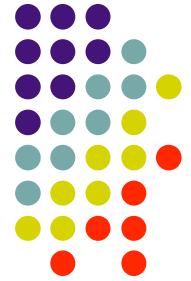
A taxonomy of property types

Wu & Barsalou (submitted)



<i>Class</i>	<i>Property Type</i>	<i>Code</i>	<i>Example</i>
Taxonomy (c)	Coordinate	cc	<i>cat-dog</i>
	Superordinate	ch	<i>cat-animal</i>
Entity (e)	Associated abstract entity	eae	<i>telephone-information</i>
	Entity behavior	eb	<i>lion-roar</i>
	External component	ece	<i>truck-wheel</i>
	External surface property	ese	<i>banana-yellow</i>
	Internal component	eci	<i>car-engine</i>
	Internal surface property	esi	<i>pineapple-crunchy</i>
	Larger whole	ew	<i>cow-cattle</i>
	Made-of	em	<i>bottle-glass</i>
	Quantity	eq	<i>pear-slice</i>
	Systemic feature	esys	<i>elephant-wild</i>
Situation (s)	Associated entity	se	<i>spoon-bowl</i>
	Associated event	sev	<i>watermelon-picnic</i>
	Function	sf	<i>scissors-cut</i>
	Action	sa	<i>banana-eat</i>
	Location	sl	<i>ship-port</i>
	Participant	sp	<i>boat-fisherman</i>
	Time	st	<i>pineapple-summer</i>
Introspective (i)	Cognitive operation	io	<i>snail-like a slug</i>
	Evaluation	ie	<i>pineapple-delicious</i>
	Negation	in	<i>penguin-cannot fly</i>

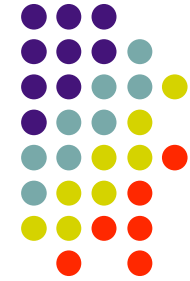
Comparing differing property spaces



- Comparison between WSM properties and two types of human-elicited property sets
- Human data
 - **Semantic Norms** (McRae *et al.* 2005)
 - properties produced on the basis of a written stimulus
 - **metalinguistic task**
 - **ESP Game** (von Ahn & Dabbish 2004)
 - spontaneous descriptions of Web images within a coordination game

ESP Game

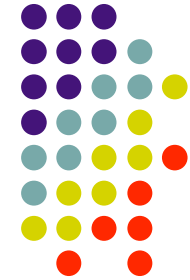
<http://www.gwap.com/>



- “Collaborative” annotation of web images
 - two partners are presented with a set of images
 - communication is not possible
 - the purpose of the game is to assign to the image the same label (word, phrase, etc.) chosen by the partner
 - every time the players converge on the same label, they gain some points
- **Players tend to converge on similar labels to describe the images**
 - “the string on which two players agree is typically a good label for the image” (Von Ahn & Dabbish 2004)
- The game is currently being used by Google
 - **Google Image Labeler**
 - <http://images.google.com/imagelabeler/>

ESP Game

<http://www.gwap.com>



10 MILLION LABELS COLLECTED

The ESP Game

As seen on CNN and newspapers around the world! beta

47 Players LOGGED in

Welcome, **ALEXENCI**
[Sign Out](#)

Today's Best Players

JWARENOH	119100
PAVILIONXYZ	86635
IMPRIMERE	72720
MINISBACK	58310
JENNINHELAB	45700
KGAP	44850

Most points in the last 24 hours
(Updated every hour)

HOW TO Play
) Play NOW (
your Profile
top scores

! Did you know?
The ESP Game is helping to label all images on the Web!
learn more...
Play our new game
NEW Phetch NEW

[Terms of Service](#) | [FAQ](#) | [ESP Image Search](#) | [Contact Us](#) | [Credits](#)

Funded in part by the National Science Foundation (NSF).
© 2005 Carnegie Mellon University, all rights reserved. Patent Pending.

ESP Game

<http://www.gwap.com/>



0:42
Time Left

The ESP Game

0250
score

Taboo Words
FOREST

Your Guesses

Type your next guess:

Pass

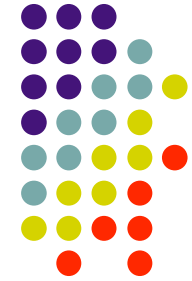
Flag

Your partner has entered a guess



© 2005 Carnegie Mellon University, all rights reserved. Patent Pending.

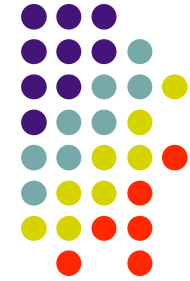
ESP Game Corpus



- Corpus di ca. 360,000 labels, corresponding to 50,000 images
 - average of 7.2 labels per image
 - es. {eat, table, people, wine, dinner}
 - the corpus was lemmatized
 - for each of 44 concrete nouns we selected the most significant labels
 - labels with the highest log-likelihood association score with the target

<i>car</i>	<i>log-likelihood</i>
wheel	12.7
road	11.4
truck	10.9
wheels	10.2
race	9.7

ESP vs. Semantic Norms



- Associations extracted from ESP represent a sort of *de facto* semantic norms
- **Semantic Norms**
 - properties are produced in a controlled situation and the task is metalinguistic
 - subjects are explicitly asked to list the most salient properties of the target concept
 - concepts are described with **words**
 - targets are described **out of contexts**
- **ESP**
 - **spontaneous** productions of the properties
 - players are only asked to try to choose the same label of the partners
 - they do not receive instructions on what to describe
 - stimuli are **pictures**
 - stimuli are described in context (**situated entities**)

3rd Experiment

Property Generation

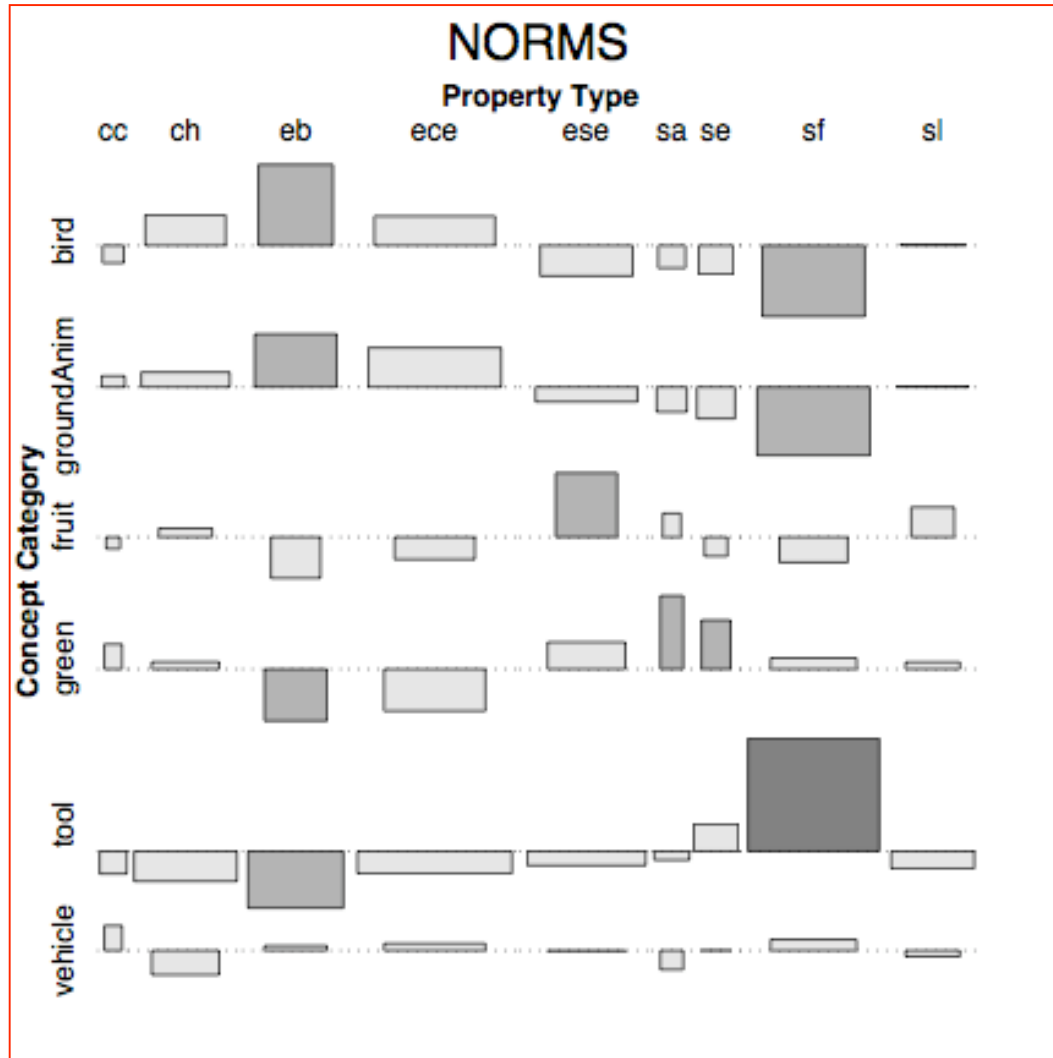
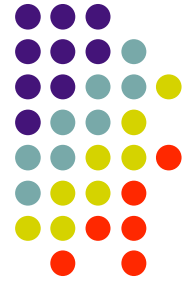


Baroni & Lenci (in press)

- cc** co-hyponym
- ch** hyperonym
- eb** prototypical behavior
- ece** meronym
- ese** perceptual attribute
- sa** prototypical action
- se** associated entity
- sf** prototypical function
- sl** prototypical location

3rd Experiment

Property Generation

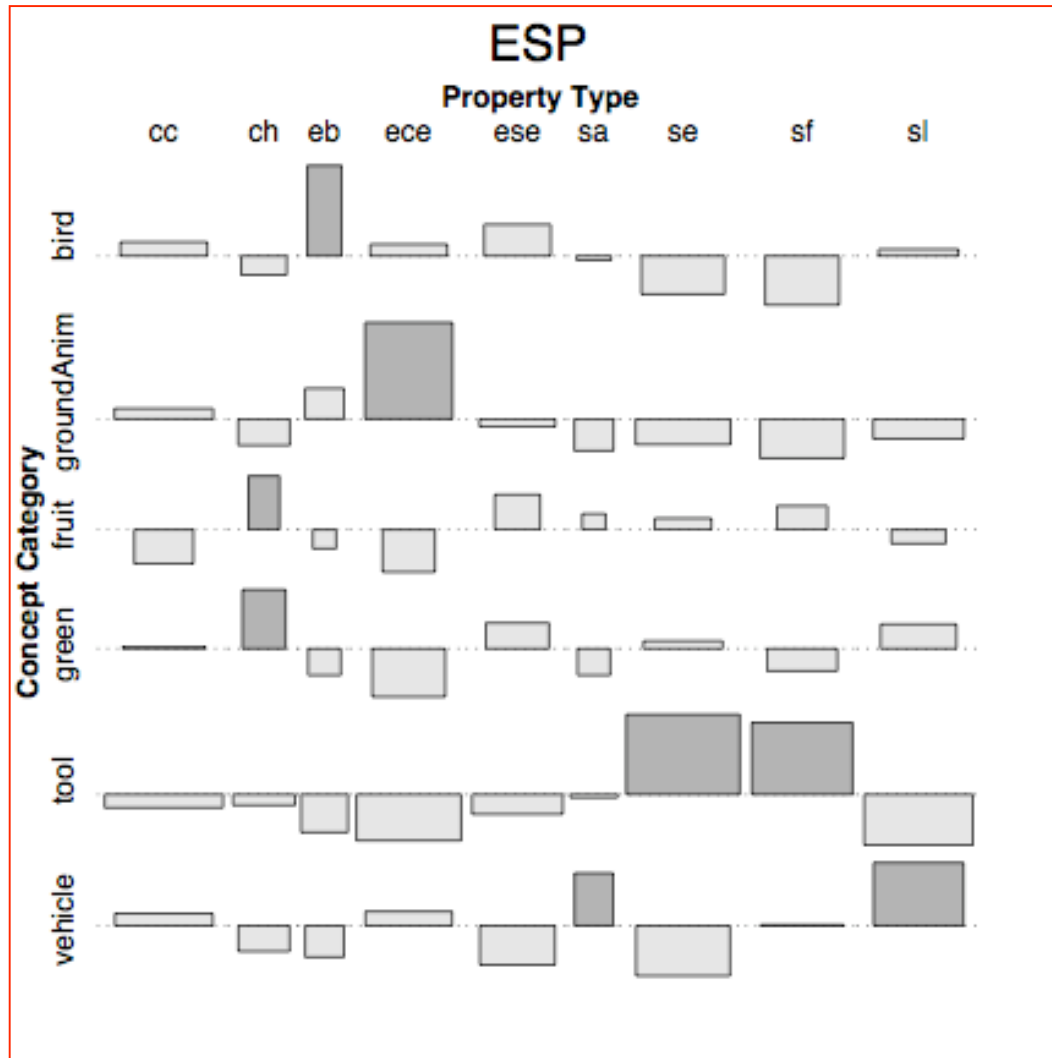
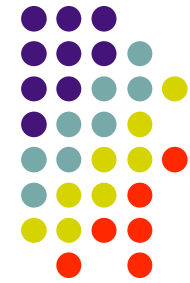


Baroni & Lenci (in press)

- cc** co-hyponym
- ch** hyperonym
- eb** prototypical behavior
- ece** meronym
- ese** perceptual attribute
- sa** prototypical action
- se** associated entity
- sf** prototypical function
- sl** prototypical location

WSM e proprietà

Baroni & Lenci (in corso di stampa)

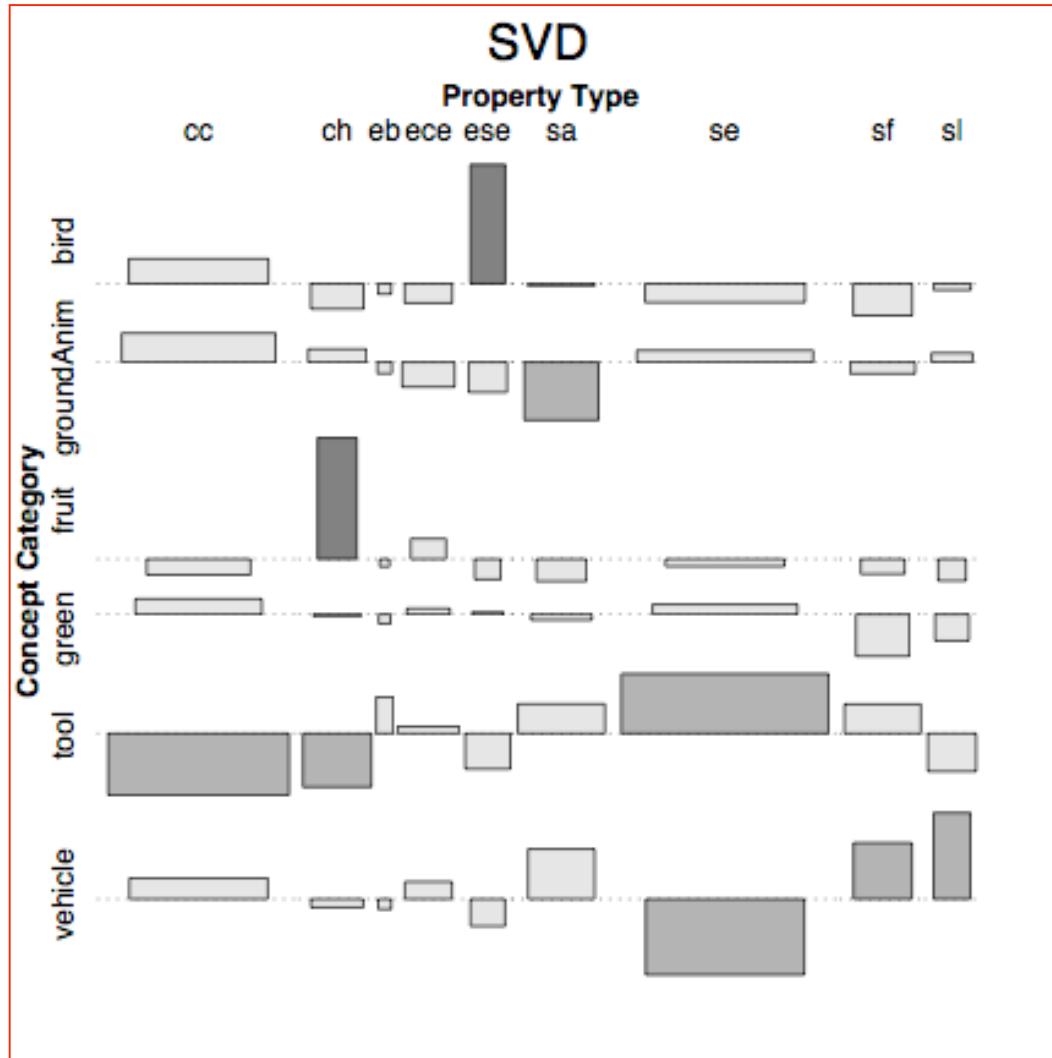


Baroni & Lenci (in press)

- cc** co-hyponym
- ch** hyperonym
- eb** prototypical behavior
- ece** meronym
- ese** perceptual attribute
- sa** prototypical action
- se** associated entity
- sf** prototypical function
- sl** prototypical location

3rd Experiment

Property Generation

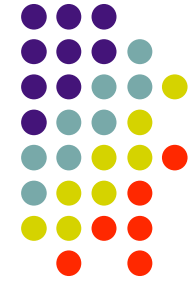


Baroni & Lenci (in press)

- cc** co-hyponym
- ch** hyperonym
- eb** prototypical behavior
- ece** meronym
- ese** perceptual attribute
- sa** prototypical action
- se** associated entity
- sf** prototypical function
- sl** prototypical location

Property generation

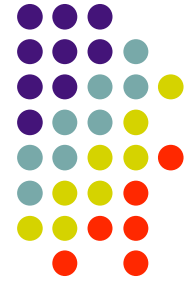
some remarks



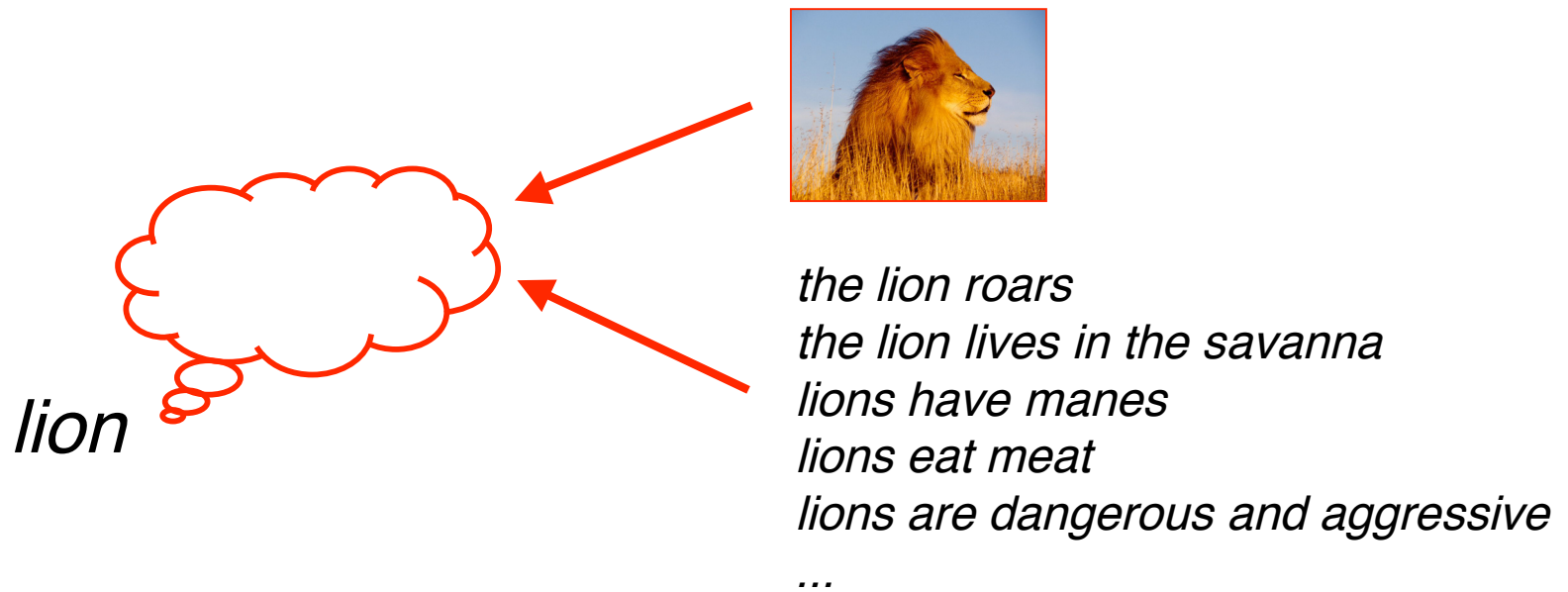
- Some property types are largely **underrepresented** in WSMs
 - e.g. meronyms, attributes
 - this information may be simply underrepresented in corpora
- WSMs have a strong bias towards **taxonomical relations**
 - co-hyponyms, hyperonyms, etc.
- Barsalou *et al.* (2008) argue that taxonomical relations (especially hyperonyms) may have a linguistic source, contrasting with other properties (e.g. meronyms) that are more related to the “**situated simulation**” of categories

Distributional semantics

some theoretical issues

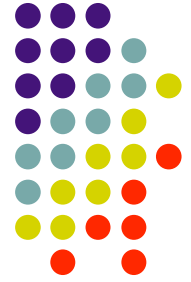


- Language is just **another input** that we use to build our conceptual representations of the external world



Distributional semantics

some theoretical issues



- **Language as a source of meaning**
 - to what extent does our lexical competence depend on the fact that we observe a word in certain (**linguistic**) **contexts**?
 - to what extent is linguistic input responsible for the **acquisition of word meaning**?
 - to what extent **internalized usage distributions** can explain human responses in behavioral semantic tasks?
 - how **paradigmatic semantic classes** are correlated with syntagmatic distributions?

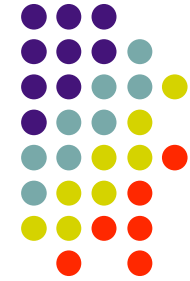
Distributional semantics

some theoretical issues



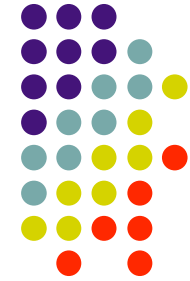
- In linguistics (both formal and cognitive approaches) there is a common tendency to explain distributional regularities in terms of semantic properties of their items
 - word distributions are the explicanda and semantic properties are the explicantia
 - w_1 and w_2 behave similarly from the linguistic point of view *because* they have similar semantic properties
- Distributional models suggest that semantic lexical properties themselves can (at least in part) be grounded in (depend on) linguistic distributions
 - caveat for using semantics as a “prior” for linguistic explanations

References



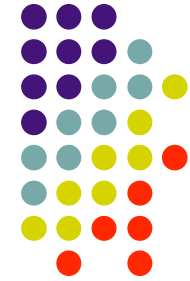
- Baroni, M. & A. Lenci (2008), “Concepts and Properties in Word Spaces”, in Lenci A., (ed.), *From context to meaning: distributional models of the lexicon in linguistics and cognitive sciences*, special issue of the *Italian Journal of Linguistics*, XX/1
- Charles, W.G. (2000), “Contextual correlates of meaning”, *Applied Psycholinguistics*, 21: 505-524
- Firth, J. R. (1957), *Papers in Linguistics*, London, Oxford University Press
- Harris Z. (1968), *Mathematical Structures of Language*, New York, Wiley
- Harris, Z.S. (1970), *Papers in Structural and Transformational Linguistics*, D. Reidel Publishing Company, Dordrecht-Holland
- Lund K., Burgess C., & R.A. Atchley (1995), “Semantic and associative priming in high-dimensional semantic space”, *Proceedings of the Cognitive Science Society*, Hillsdale, N.J., Erlbaum Publishers: 660-665
- McDonald S. & M. Ramscar (2001). “Testing the distributional hypothesis: The influence of context on judgements of semantic similarity”, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Edinburgh, LEA: 611-616
- McRae, K., Cree, G., Seidenberg, M. and C. McNorgan (2005), “Semantic feature production norms for a large set of living and nonliving things”, *Behavior Research Methods*, 37: 547-559
- Miller, G.A. & W.G. Charles (1991), “Contextual correlates of semantic similarity”, *Language and Cognitive Processes*, 6: 1-28
- Ramscar M. & D. Yarlett (2003), “Semantic grounding in models of analogy: An environmental approach”, *Cognitive Science*, 27(1): 41-71
- Vigliocco, G., Vinson, D.P, Lewis, W. & M.F. Garrett, (2004), “Representing the meanings of object and action words: The featural and unitary semantic space hypothesis”, *Cognitive Psychology*, 48: 422-488

References



- Baroni M., Lenci A., & L. Onnis (2007), “ISA meets Lara: A fully incremental word space model for cognitively plausible simulations of semantic learning”, in *Proceedings of the ACL Workshop on Cognitive Aspects of Language Acquisition*, Praha: 49-56
- Burgess, C. & K. Lund (1997), “Modelling parsing constraints with high-dimensional context space”, *Language and Cognitive Processes*, 12: 1-34.
- Karlgren, J. & M. Sahlgren (2001), “From words to understanding”, in Uesaka Y., Kanerva P. & H. Asoh (eds.), *Foundations of real-world intelligence*, Stanford, CSLI: 294-308
- Landauer, Th.K. & 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
- Lin, D. (1998), “An Information-Theoretic Definition of Similarity”, in *Proceedings of the 15th International Conference on Machine Learning*: 296-304
- Lowe, W. (2001), “Towards a theory of semantic space”, Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Philadelphia, PA, LEA: 576-581
- Padó S. & M. Lapata (2007), “Dependency-based construction of semantic space models”, *Computational Linguistics*, 33(2): 161-199
- Rapp, R (2004), “A Freely Available Automatically Generated Thesaurus of Related Words”, *Proceedings of LREC 2004*, Lisbona, ELRA: 395-398

References



- Barsalou, L.W., Santos, A., Simmons, W.K., & C.D. Wilson (2008), “Language and simulation in conceptual processing”, in M. De Vega, A.M. Glenberg, & A.C. Graesser (eds.), *Symbols, embodiment, and meaning*, Oxford, Oxford University Press
- Farkas, I. & P. Li (2002), “DevLex: A self-organizing neural network model of the development of lexicon”, in L. Wang *et al.* (Eds.), *Proceedings of the Ninth International Conference on Neural Information Processing*, Singapore
- Kintsch, W. (2007), “Meaning in context”, in Landauer, T.K., McNamara, D.S., Dennis S. & W. Kintsch, (eds.) (2007), *Handbook of Latent Semantic Analysis*, Mahwah NJ, Lawrence Erlbaum: 89-105.
- Li, P., Farkas, I. & B. MacWhinney (2004), “Early lexical acquisition in a self-organizing neural network”, *Neural Networks*, 17: 1345-1362
- Noppeney, U. & C. Price (2004), “Retrieval of abstract semantics”, *Brain and Image*, 22: 164-170
- von Ahn L. & L. Dabbish (2004), “Labeling images with a computer game”, *ACM Conference on Human Factors in Computing Systems*: 319-326
- Wu, L. & L. Barsalou (in press), “Grounding concepts in perceptual simulation: Evidence from property generation”