Language Processing Models

15-486/782: Artificial Neural Networks
David S. Touretzky

Fall 2006

(This lecture is based in part on notes by David Plaut)

1

Outline

- Deep and surface dyslexia
- Mapping words to meaning
- Distributed representations
- Hinton-Shallice-Plaut model

- Understanding English sentences
- Rohde's CSCP model

2

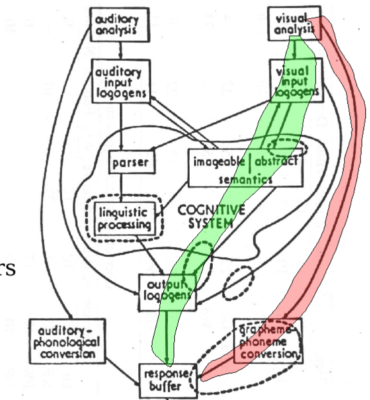
Dyslexia

- Deep dyslexia: reading via semantic route only
  - Cannot read simple non-words like “brund”
  - Semantic errors: see PEACH, say “apricot”
  - Fewer errors on concrete vs. abstract words
  - Part of speech effect:
    nouns > adjectives > verbs > function words
  - Subtypes of deep dyslexia: input, central, output.
- Surface dyslexia: map letters directly to sound
  - Can read pronounceable non-words
  - Cannot read irregular words: “pint”, “yacht”

3

Dual-Route Model of Reading

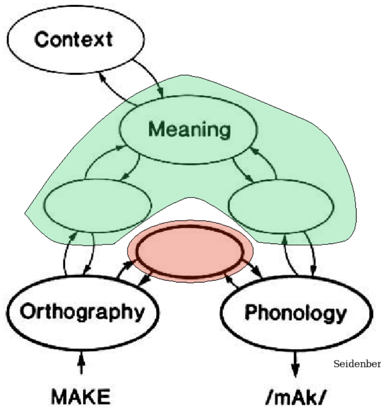
- **Phonological** route: letters to sounds
  - “cat”
  - nonwords (“brund”)
  - Damage produces deep dyslexia
- **Semantic** route: letters to meaning to sounds
  - “pint”
  - “yacht”
  - Damage produces surface dyslexia



Morton & Patterson 1980

4

Connectionist (Neural Network) Dual-Route Model



Seidenberg & McClelland 1989

5

Error Types in Deep Dyslexia

- Semantic (RIVER ⇒ “ocean”)
- Visual (SCANDAL ⇒ “sandals”)
- Mixed visual/semantic (TROUBLE ⇒ “terrible”)
- Visual-then-semantic (SYMPATHY ⇒ “orchestra”)
- Morphological/derivational (LOVELY ⇒ “loving”)
- Function-word substitution (FROM ⇒ “with”)

Dyslexic patients always produce a mixture of error types. But the distribution varies based on the type of dyslexia.

6

Patient Error Distributions

Error Proportions

Type		Semantic	Mixed	Visual	Derivational	Other
	VS	19	16	48	10	7
input	PS	10	7	51	9	23
	KF	4	10	61	19	6
central	DE	23	6	35	32	4
	WS	21	17	35	4	23
output	PW	54	4	13	22	6
	GR	56	?	22	11	11

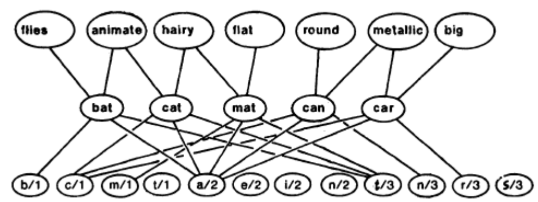
Correct Performance (percent)

	Concrete	Abstract	Adjective	Verb	Functor	Nonword
GR	50	10	16	6	1	0
KF	73	14	32	7	10	—
PW	67	13				0
DE	70	10				10

7

Mapping Words to Meaning: Local Word Representation

Damage to hidden layer knocks out individual words.

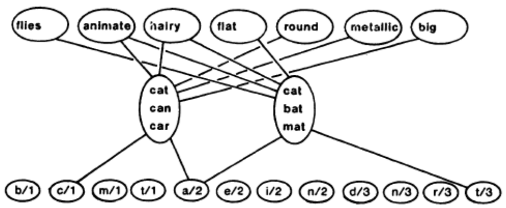


Hinton, McClelland, and Rumelhart 1986

8

Representing an Arbitrary Mapping in Distributed Form

- Hidden units represent clusters of visually similar words.
- Sememe (output) units must be thresholded to prevent false positives.

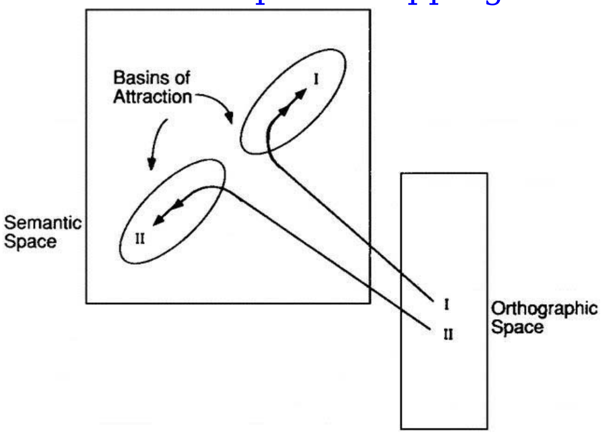


Hinton, McClelland, and Rumelhart 1986

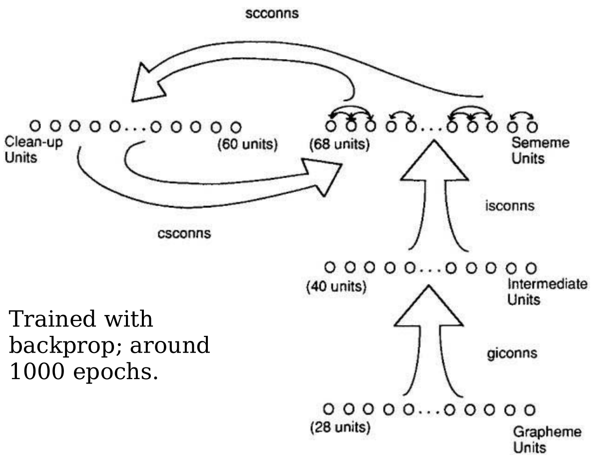
Distributed Representations in a Feed-Forward Network: Error Patterns

- Damage to hidden layer does not knock out single words.
- Instead, damage raises the overall error rate.
- Can produce a variety of error types.
- But output patterns and distribution of error type does not match human data.
- What's needed: a way to “clean up” the semantics.

Attractor Dynamics Can Clean Up the Mapping



Hinton & Shallice Model (1991)



Trained with backprop; around 1000 epochs.

68 Semantic Features

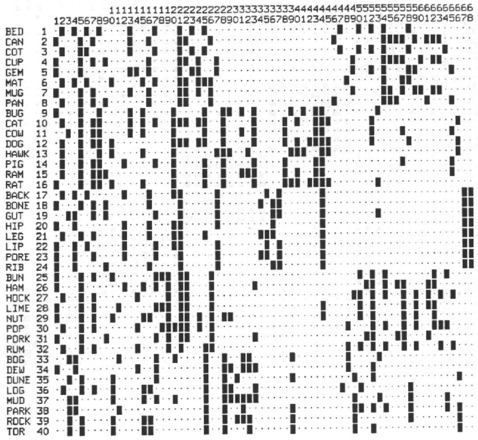
Appendix B

Table B1
Semantic Features

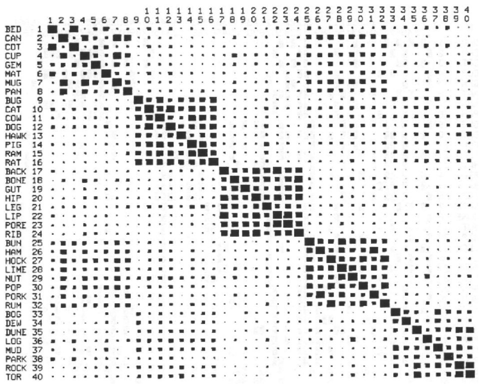
No.	Feature	No.	Feature
1	max-size-less-foot	34	partof-limb
2	max-size-foot-to-two-yards	35	surfaceof-body
3	max-size-greater-two-yards	36	interiorof-body
4	main-shape-2D	37	above-waist
5	main-shape-3D	38	mammal
6	cross-section-rectangular	39	wild
7	cross-section-circular	40	fierce
8	has-legs	41	does-fly
9	white	42	does-swim
10	brown	43	does-run
11	green	44	living
12	color-other-strong	45	carnivore
13	varied-colors	46	madeof-metal
14	transparent	47	madeof-wood
15	dark	48	madeof-liquid
16	hard	49	madeof-other-nonliving
17	soft	50	gotfrom-plants
18	sweet	51	gotfrom-animals
19	tastes-strong	52	pleasant
20	moves	53	unpleasant
21	indoors	54	man-made
22	in-kitchen	55	container
23	in-bedroom	56	for-cooking
24	in-living-room	57	for-eating-drinking
25	on-ground	58	for-other
26	on-surface	59	used-alone
27	otherwise-supported	60	for-breakfast
28	in-country	61	for-lunch-dinner
29	found-woods	62	for-snack
30	found-near-sea	63	for-drink
31	found-near-streams	64	particularly-assoc-child
32	found-mountains	65	particularly-assoc-adult
33	found-on-farms	66	used-for-recreation
		67	human
		68	component

Note. Directly interconnected sememes occur within the same section.

40 Word Vocabulary



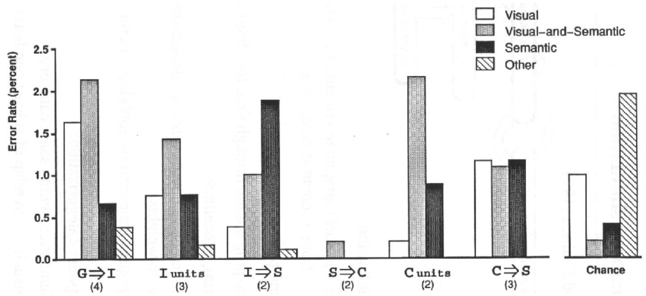
Semantic Similarity



Simulating Brain Damage

- Various methods:
  - Remove some hidden units
  - Remove some cleanup units
  - Add noise to the weights
- Similar effects, but some differences.

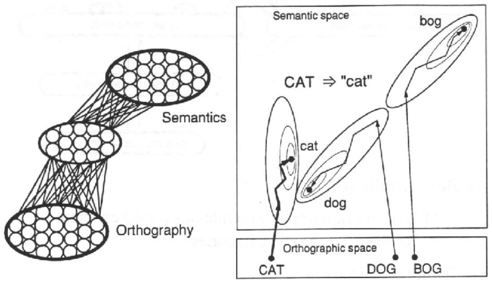
Distribution of Error Types



(G)rapheme, (I)ntermediate, (S)emantic, (C)leanup

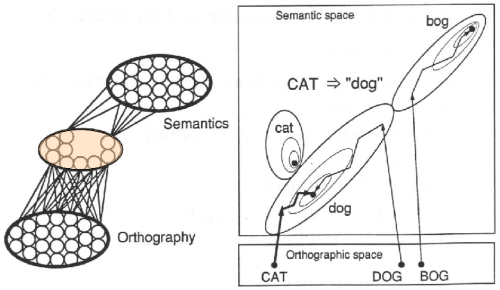
Chance = pick a word at random

Normal Network



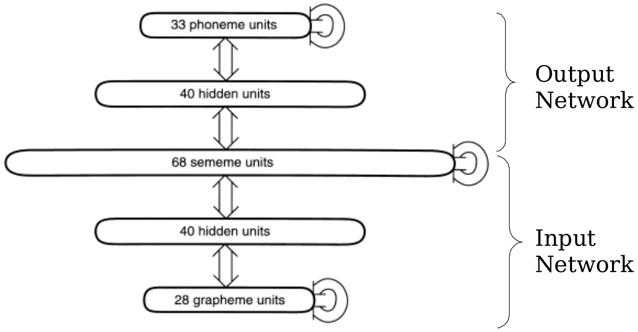
Lesioned Network

Lesions change the shape of the attractor space.



Hinton-Shallice-Plaut Model

Add a phonological layer to address more error types.



Deterministic Boltzmann Machine with Noise (Instead of Backprop)

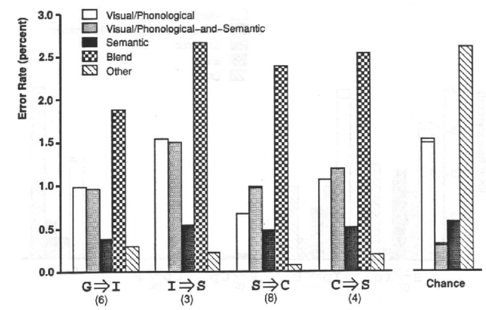
$$\Delta w_{ij} = \langle s_i s_j \rangle^{+} - \langle s_i s_j \rangle^{-}$$

$$s_j^{(t)} = \lambda s_j^{(t-1)} + (1-\lambda)\tanh\left(\frac{1}{T}\left(\nu_\sigma + \sum_i s_i^{(t-1)} w_{ij}\right)\right)$$

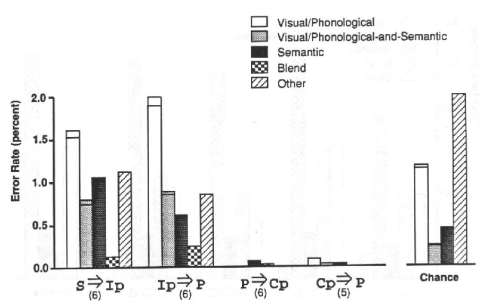
where $\nu_\sigma \in N(0,\sigma)$

T is a gradually lowering temperature parameter:
- high T → flat sigmoid
- low T → steep sigmoid (step function)

Error Rates for Damage to Input Network



Error Rates for Damage to Output Network



Concrete vs. Abstract Words

- Assume that concrete/visualizable words have more semantic features than abstract words.
- More features = more recurrent connections: greater opportunity for cleanup.
- Result: model produces fewer errors for concrete words.

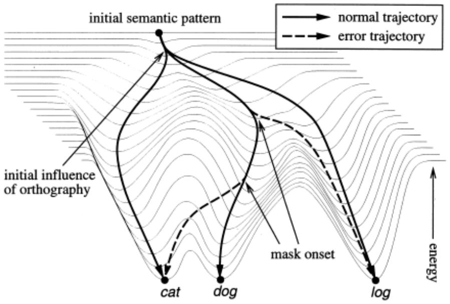
Recovery From Damage

- Retraining the network after damage produced faster recovery than the original training.
- Retraining on only a subset of the words produced improvements in all the words.
  - Distributed representations produce transfer among words
- How should patients be trained to help them recover from damage?
  - One possibility: focus on words that are non-prototypical for their semantic category.

Errors in Normal Subjects

- McLeod, Shallice, and Plaut (2000): normal subjects can be induced to make reading errors by rapid presentation of stimuli.
- Flash a sequence of words in rapid succession (160 msec); each asks as a masking stimulus for the preceding word. Ask subjects to identify words they saw.
- Results: subjects make a mixture of visual and semantic errors

Attractor Basins



Rohde's Sentence Comprehension and Production Model (CSCP)

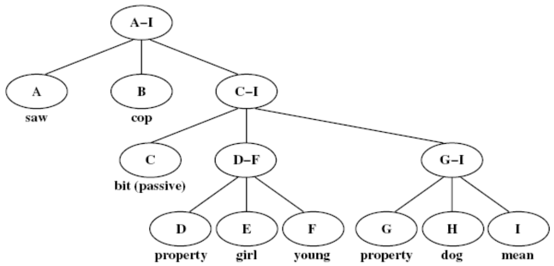
- Could understand simple sentences and answer queries about them.
- Vocabulary around 300 words.
- Complex grammar: various types of embedded clauses, parallel constructions, ambiguities.
- Giant network composed of multiple modules.
- Several SRNs (Simple Recurrent Nets).
- Trained with crosss-entropy; zero-error radius 0.1.
- Took 2 months to train! (500 MHz DEC Alpha)

Comprehension Measured by Query Answering

- Input: "The boy ate the soup."
- Queries:
  - Who ate?
  - What did the boy eat?
  - What was done to the soup?

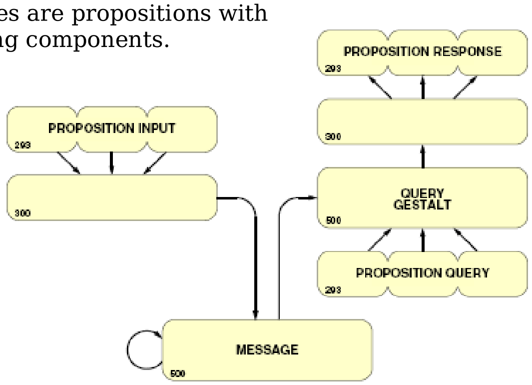
Propositional Encoding

"A cop saw the young girl that was bitten by that mean dog."

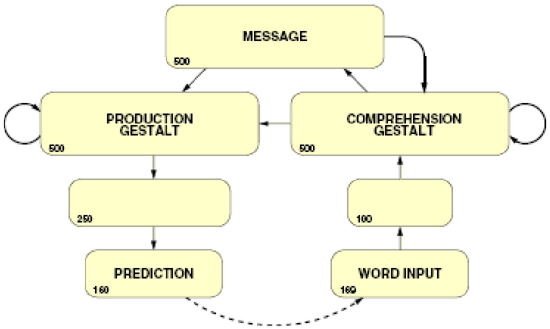


Message Encoder/Decoder Network

- Queries are propositions with missing components.



Comprehension/Production Network



The Full Model

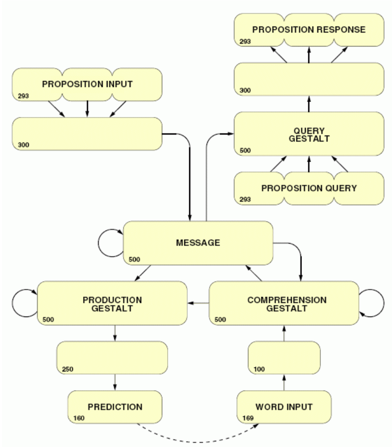


Figure 6.1: The full CSCP model.

33

Results

- The model learned a basic (300 word) subset of English with rich grammatical structure.
- Generalization to novel sentences.
- Some ability to resolve ambiguities, e.g., attachment.
- Training time was substantial.
- Some non-human-like errors.
- Could estimate reading time, grammaticality.

34