# More past tense!

LINGUIST611
Spring 2022

# The legacy of the PDP model

→ Uniform procedure: No **qualitative** distinction between regulars and irregulars.

→ Novel account of generalization.

→ Learning by exposure to examples - no hypothesis testing or explicit rule learning.

→ Domain-general. Consists of units that are not unique to language; Language in the PDP model is a reflection of more general cognitive properties.

→ No distinction between competence and performance: Knowledge arises only in the context of solving a particular task.

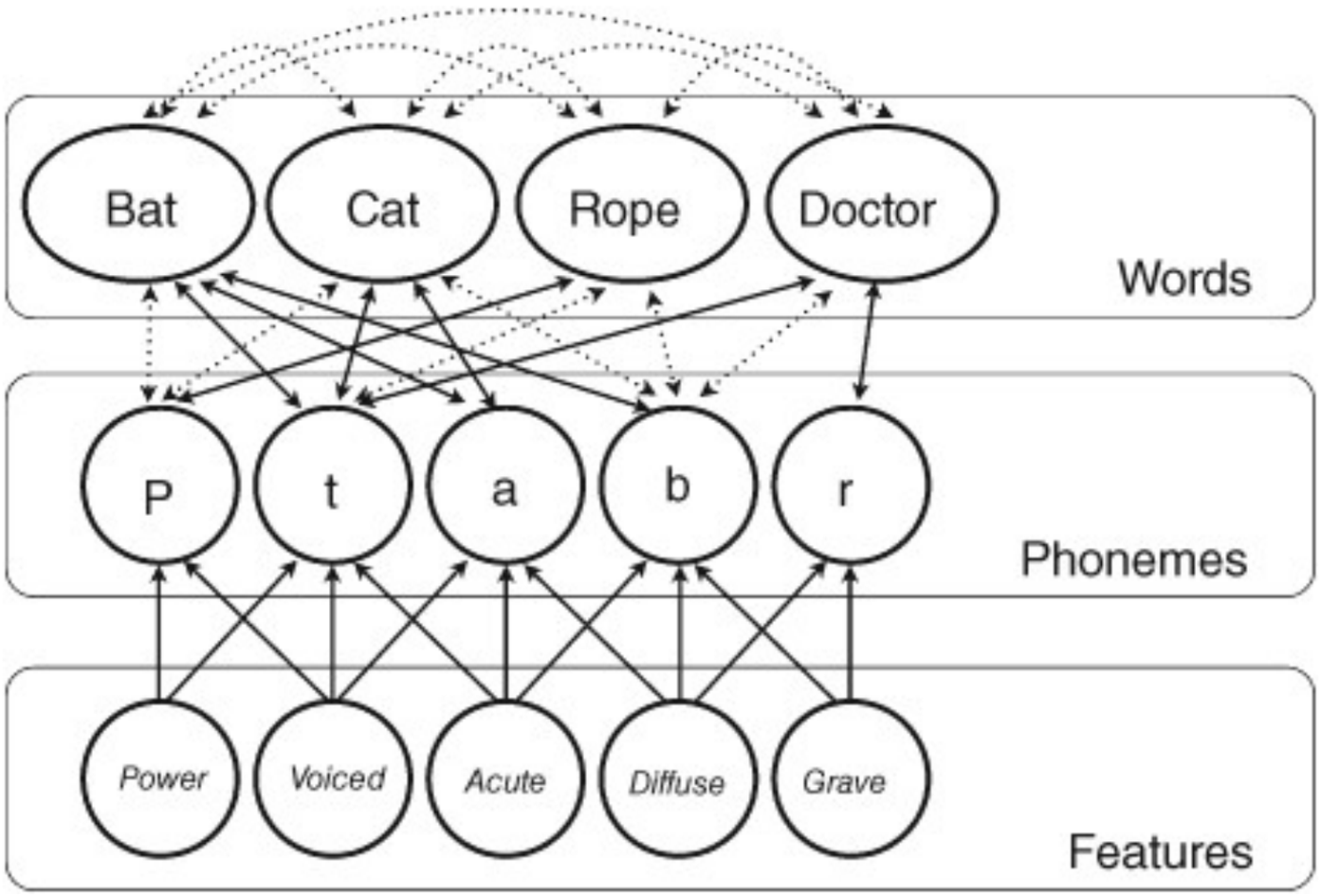
→ Understanding the nature of language requires understanding (cognitive) neurobiology, rather than analysis of primary linguistic data.

→ Computational theories are a key part of psychological theorizing.

# The legacy of the PDP model

**→ Graded effects arise from the interaction of multiple soft / violable constraints!**

- Optimality Theory et al.
- TRACE model of speech recognition
- Constraint-based sentence processing (to be seen)



# Issue #1: What do we need to explain?

| **Regulars:** | **Irregulars:** |
|---|---|
| kick - kicked | sleep - slept |
| pull - pulled | ring - rang |
| groan - groaned | buy - bought |
| vie - vied | go - went |
| introduce - introduced | come - came |
| bloviate - bloviated | eat - ate |
| ossify - ossified | take - took |
| hand - handed | bring - brought |
| yelp - yelped | drink - drank |

**<u>View #2: There's no in principle distinction between these two types of forms - the theory must explain how speakers achieve both types.</u>**

# Issue #2: What makes a psycholinguistic theory?

Words and rules**:** Verbal theory. Makes commitments on central distinctions- there is an associative memory, and a rule-based mechanism. Predictions are broad/qualitative and derived by reasoning from these premises.

Rumelhart & McClelland**:** Computationally implemented theory. Makes commitments on central distinctions - there is a single pattern associator - as well as less central features, necessary to derive predictions (e.g. Wickelphones). Predictions are precise/ quantitative and derived by simulation.

**→ What are the strengths and weaknesses of each approach? Is it a fair comparison?**

# Pinker & Prince (1988)

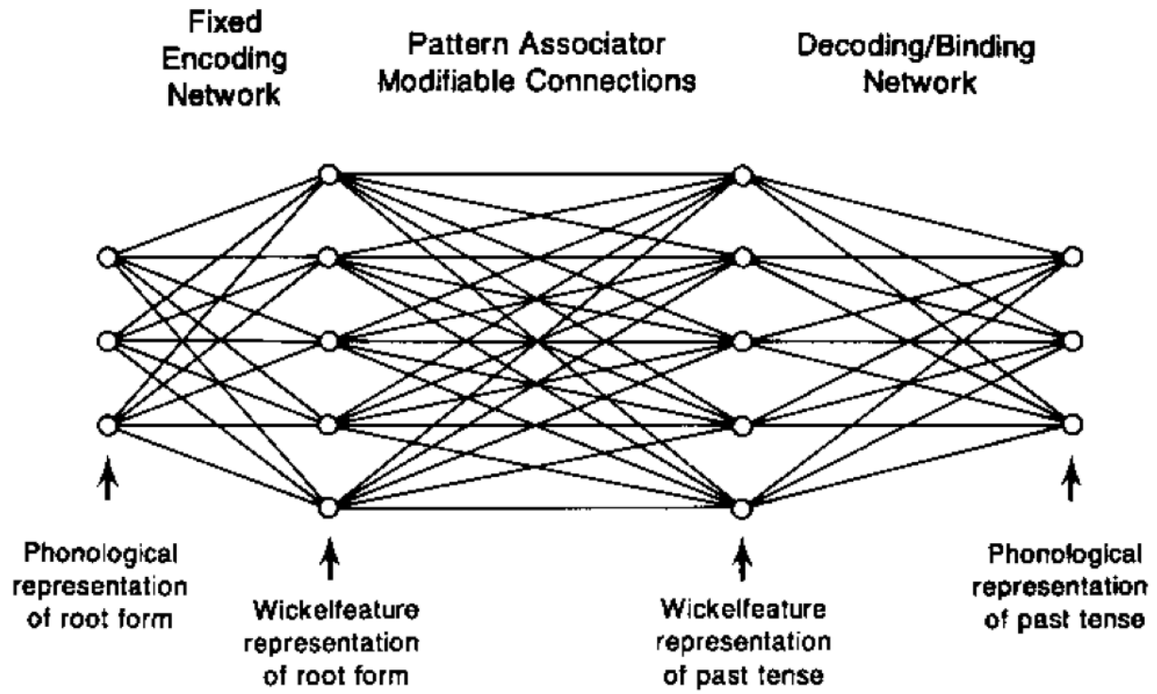
- Unable to account for linguistic structure (denominal verbs are regular: *ringed* vs *rang*)

- Unable to account for differences in compounding: *mice-eater* but not **rats-eater*.

- Unable to account for homophony *break* vs *brake*

- Unable to represent temporal order (*rapata* vs *ratapa*)

- Predicts that there should be typicality effects for regulars and irregulars alike.



# Pinker & Prince (1988)

→ **Predictions of an abstract rule:**

- There should be no effect of stem typicality: The abstract rule should apply equally well no matter what value supplies the variable.
- The rule is a regular **elsewhere** case: It applies when there is not a form that lexical memory supplies. It does not need to be the most frequent exponent / morpheme.

# Hybrid models

→ Since the original debate, much interest in **hybrid models** that realize some or all of the original PDP claims, but often with symbolic representations / structure:

- Uniform procedure: No **qualitative** distinction between regulars and irregulars.

- Computational theories are a key part of psychological theorizing.

- Graded effects arise from the interaction of multiple soft / violable constraints

# Generalization

***Usually I [zek], but yesterday I [zekt]***

| **Present Tense** | ***-ed*** | **Vowel Change** | **No Change** |
|---|---|---|---|
| 1. [zek] | a. [zekt] | b. [zʊk] | c. [zek] |
| 2. [zet] | a. [zetəd] | b. [zʊt] | c. [zet] |
| 3. [zep] | a. [zept] | b. [zʊp] | c. [zep] |
| 4. [zik] | a. [zikt] | b. [zʊk] | c. [zik] |
| 5. [zit] | a. [zitəd] | b. [zʊt] | c. [zit] |

# Generalization: Rules vs Examples

→ **Phonological rule:** Highlights **structured** similarity, similarity based on a particular structural description / rule format.

$$\text{I} \longrightarrow \Lambda \,/\, \underline{\quad\quad}\ \eta]_{[+\text{past}]}$$

→ **Analogy:** Highlights **variegated** similarity, similarity based on any arbitrary aspect of similarity to another token.

| Model form | s | p | l | ɪ | ŋ |
|---|---|---|---|---|---|
| *fling-flung* | | f | l | ɪ | ŋ |
| *sting-stung* | s | t | | ɪ | ŋ |
| "*plip*"-"*plup*" | | p | l | ɪ | p |
| "*sliff*"-"*sluff*" | s | | l | ɪ | f |

# Rules all the way down

→ **Minimal Generalization Learner (Albright & Hayes, 2003):**
Learns the most specific phonological rule necessary to capture a given alteration, and its reliability/validity. The grammar consists of rules all the way down - from rules that govern a single verb to extremely general (regular) rules:

| | Change | Variable | Shared features | Shared segments | Change location | |
|---|---|---|---|---|---|---|
| a. | | | | | | |
| b. | ∅ → d / [ | | ʃ | aɪn | ___ ]$_{[+past]}$ | (*shine-shined*) |
| c. | ∅ → d / [ | kən | s | aɪn | ___ ]$_{[+past]}$ | (*consign-consigned*) |
| d. | ∅ → d / [ | X | $\begin{bmatrix} +\text{strident} \\ +\text{continuant} \\ -\text{voice} \end{bmatrix}$ | aɪn | ___ ]$_{[+past]}$ | (generalized rule) |

# Rules all the way down

→ **Minimal Generalization Learner (Albright & Hayes, 2003):**
Key aim is to generalize, but not generalize too much. Phonological feature vocabulary facilitates this:

(6) a. $\varnothing \rightarrow$ əd / [vot ___]$_{[+\text{past}]}$

b. $\varnothing \rightarrow$ əd / [nid ___]$_{[+\text{past}]}$

c. $\varnothing \rightarrow$ əd / [X ___]$_{[+\text{past}]}$ (too general)

d. $\varnothing \rightarrow$ əd / [X $\begin{bmatrix} +\text{coronal} \\ +\text{anterior} \\ -\text{nasal} \\ -\text{continuant} \end{bmatrix}$ ___]$_{[+\text{past}]}$ (appropriately restricted)

# Rules all the way down

→ Rules differ in their **scope** (i.e. portion of lexicon they could in principle apply to) and their **confidence** (i.e. the proportion of their scope that they do apply to). **Confidence** is **adjusted** to account for intuition that more data means a higher precision estimate of confidence by taking the lower end of the confidence interval around point estimate.

Table 1
Past tenses for *gleed* derived by the rule-based model

| Output | Rule | Hits/Scope | Raw confidence | Adjusted confidence | Hits/Failures |
|---|---|---|---|---|---|
| *gleeded* | ∅ → əd / [X {d, t} ___]$_{[+\text{past}]}$ | 1146/1234 | 0.929 | 0.872 | *want, need, start, wait, decide, etc.* / **get, *find, *put, *set, *stand,* etc. |
| *gled* | i → ɛ / [X {l, r} ___ d]$_{[+\text{past}]}$ | 6/7 | 0.857 | 0.793 | *read, lead, bleed, breed, mislead, misread* / **plead* |
| *glode* | i → o / [X C ___ [+cons]]$_{[+\text{past}]}$ | 6/184 | 0.033 | 0.033 | *speak, freeze, weave, interweave, bespeak* / **leak, *teach, *leave,* etc. |
| *gleed* | No change / [X {d, t} ___]$_{[+\text{past}]}$ | 29/1234 | 0.024 | 0.014 | *shed, spread, put, let, set, cut, hit, beat, shut, hurt, cost, cast, burst, split,* etc. / **get, *want, *need,* etc. |

# Exemplar theory

→ **Generalized Context Model (Nosofsky, 1986):** An example of an **exemplar theory** of categorization: Categories are represented through examples in memory, rather than abstract summary statistics, prototypes, or other compact descriptions of the category structure.

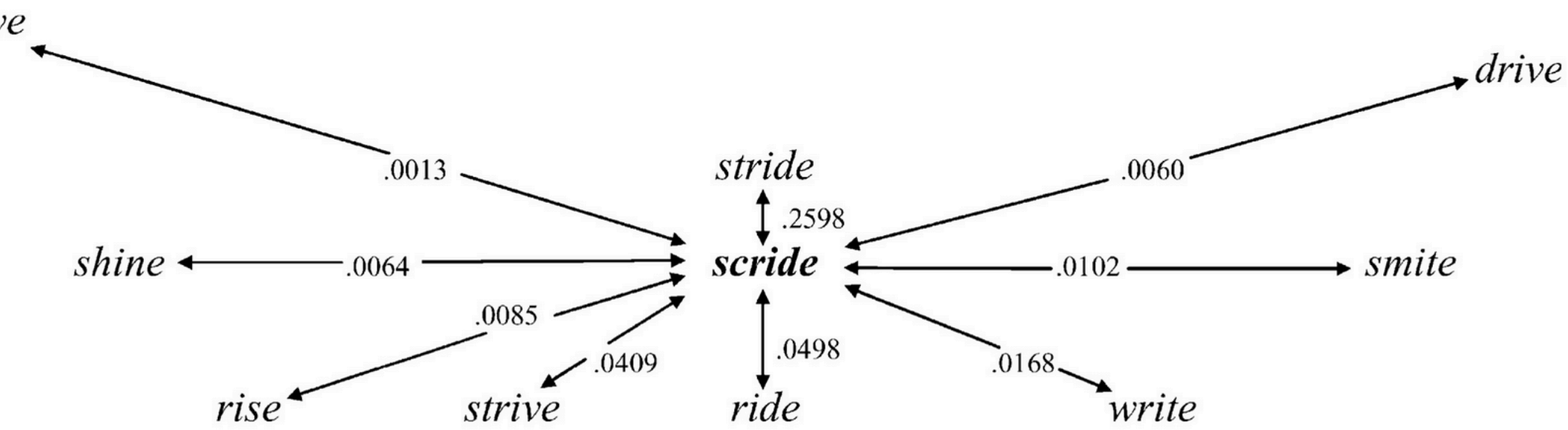



Fig. 1. Similarity of all [aɪ] → [o] forms to *scride*.

(11)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *shine*: | ʃ | | null | | null | | a | | ɪ | | n | | | |
| penalty: | 0.155 | + | 0.6 | + | 0.6 | + | 0 | + | 0 | + | 0.667 | = | 2.022 | $\eta_{ij} = e^{(-d_{ij}/s)^p}$ |
| *scride*: | s | | k | | r | | a | | ɪ | | d | | | |

# Albright and Hayes (2003)

→ **`Islands of reliability`:**

> “We will refer to phonological contexts in which a particular morphological change works especially well in the existing lexicon as "Islands of Reliability”" (p. 127)

→ Their search of the English lexicon reveals islands of reliability for both regulars and irregulars:

- **Irregulars:** ɪŋ (e.g. fling, spring, ring) is highly reliable for

$$\text{I} \rightarrow \Lambda \; / \; \underline{\quad} \; \text{ŋ}]_{[+\text{past}]}$$

- **Regulars:** all 352 stems ending in a voiceless fricative are regular.

# Albright and Hayes (2003)

Table 3
Design of the Core set of wug stems

| | |
|---|---|
| Stem occupies an island of reliability for both the regular output and at least one irregular output. | Stem occupies an island of reliability for the regular output only. |
| Stem occupies an island of reliability for at least one irregular output, but not for the regular output. | Stem occupies no island of reliability for either regular or irregular forms. |

(14) a. Island of reliability for both regulars and irregulars
***dize*** [daɪz] (*doze* [doz]); ***fro*** [fro] (*frew* [fru]); ***rife*** [raɪf] (*rofe* [rof], *riff* [rɪf])

b. Island of reliability for regulars only[12]
***bredge*** [brɛdʒ] (*broge* [brodʒ]); ***gezz*** [gɛz] (*gozz* [gaz]); ***nace*** [nes] (*noce* [nos])

c. Island of reliability for irregulars only
***fleep*** [flip] (*flept* [flɛpt]); ***gleed*** [glid] (*gled* [glɛd], *gleed*); ***spling*** [splɪŋ] (*splung* [splʌŋ], *splang* [splæŋ])

d. Island of reliability for neither regulars nor irregulars
***gude*** [gud] (*gude*); ***nung*** [nʌŋ] (*nang* [næŋ]); ***preak*** [prik] (*preck* [prɛk], *proke* [prok])

# Albright and Hayes (2003)

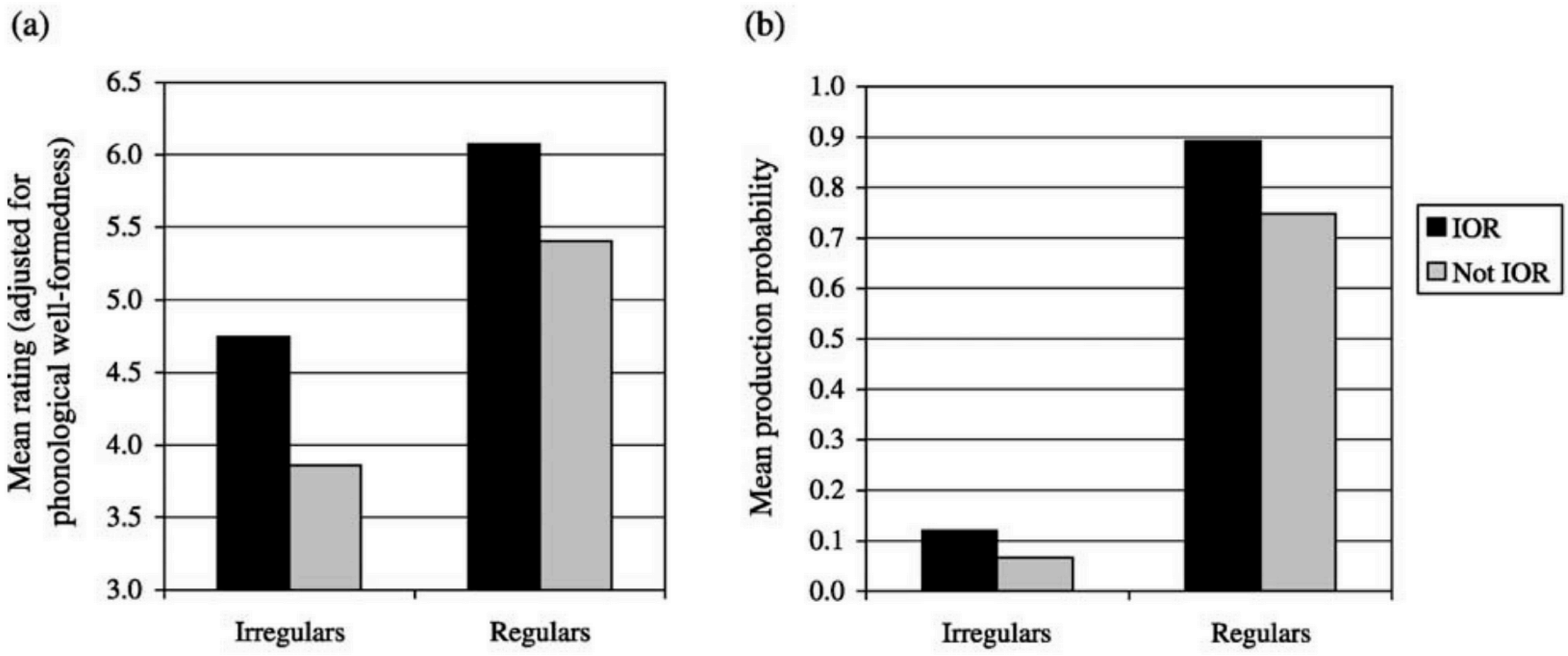



Fig. 2. Effect of islands of reliability (IOR) for irregulars and regulars. (a) IOR effect on ratings (adjusted). (b) IOR effect on production probabilities.

(21) Correlations ($r$) of participant responses to model predictions: Core verbs ($n = 41$)

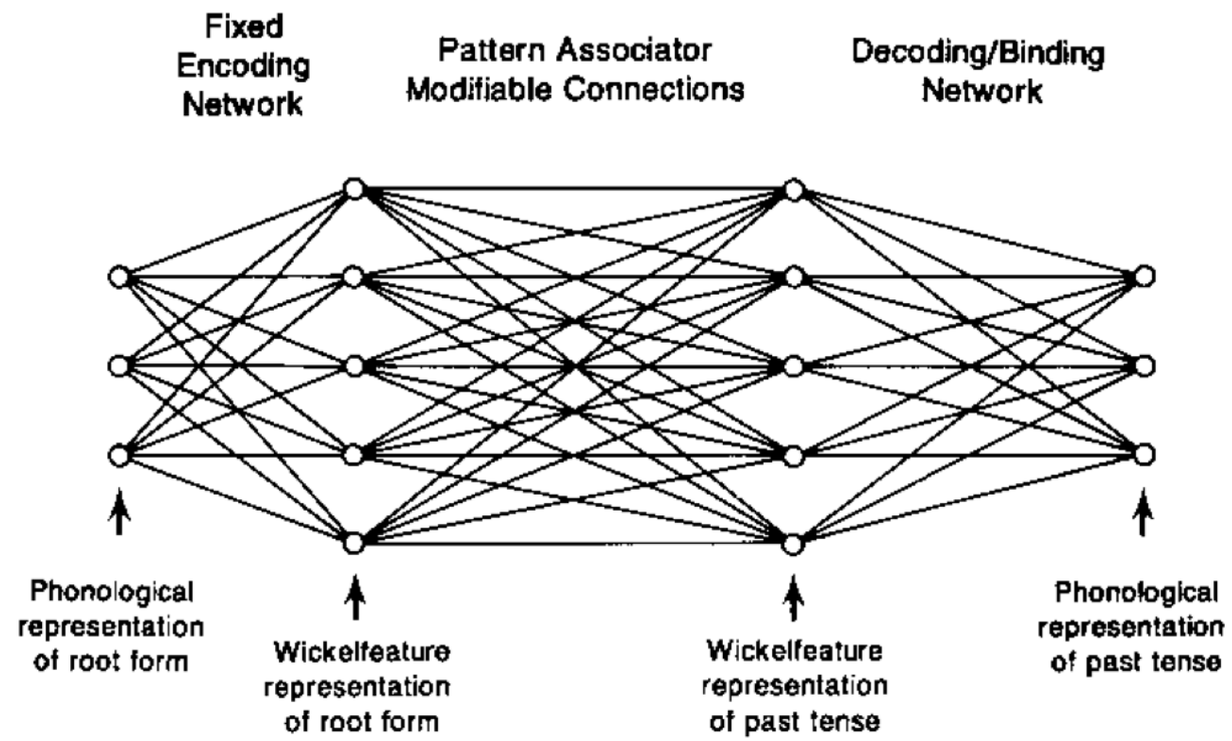
| | Rule-based model | | Analogical model | |
|---|---|---|---|---|
| | Ratings | Production probabilities | Ratings | Production probabilities |
| Regulars | 0.745 ($P < 0.0001$) | 0.678 ($P < 0.0001$) | 0.448 ($P < 0.01$) | 0.446 ($P < 0.01$) |
| Irregulars | 0.570 ($P < 0.0001$) | 0.333 ($P < 0.05$) | 0.488 ($P < 0.001$) | 0.517 ($P < 0.0001$) |

# Albright and Hayes (2003)

Locating the final consonant to determine the correct ending is a canonical case where structured similarity is required: the past tense allomorph depends solely on the final segment of the stem, in particular on just a few of its features. Our analogical model, however, is inherently unable to focus on these crucial structural elements. Instead, it gets distracted by variegated similarity, and makes wrong guesses. For instance, for the existing verb *render*, the analogical model guesses **renderèd* [rɛndərəd], based largely on the following analogical set (the ten most similar forms): *rend*, *end*, *rent*, *vend*, *raid*, *fend*, *mend*, *tend*, *round*, and *dread*. These stems bear an irrelevant similarity to *render*, which (in this case) suffices to outweigh the influence of legitimate model forms like *surrender*. The analogical model also invoked variegated similarity to overgeneralize the allomorph [-t], e.g. *whispert* [wɪspərt], based on forms like *whip*, *wish*, *whisk*, *wince*, *quip*, *lisp*, *swish*, *rip*, *work*, and *miss*.

The participants in our experiment misattached [-əd] precisely once, in the volunteered form *bliggèd* [blɪgəd]. This may be compared to the 936 responses in which the correct [-d] was attached to stems ending in non-alveolar voiced segments. We conjecture that the basis for [blɪgəd] may have been archaic forms of English (e.g. *banishèd*), encountered in music and poetry, or perhaps it was merely a speech error.

# Back to Connectionism




McLelland and Patterson (2002:471)

> We do not claim that it would be impossible to construct a rule-based model of inflection formation that has all of the properties supported by the evidence. However, such an account would not be an instantiation of Pinker's symbolic rule account. In fact, rule-based models with some of the right characteristics are currently being pursued ([45]; Albright and Hayes, unpublished). If such models use graded rule activations and probabilistic outcomes, allow rules to strengthen gradually with experience, incorporate semantic and phonological constraints, and use rules within a mechanism that also incorporates word-specific information, they could become empirically indistinguishable from a connectionist account. Such models might be viewed as characterizing an underlyingly connectionist processing system at a higher level of analysis, with rules providing descriptive summaries of the regularities captured in the network's connections.

# Maximum Entropy

→ **Related formalism:** Probabilistic models of grammar popular in phonology for modeling well-formedness judgments (e.g. Hayes et al. 2009), variation (Coetzee and Pater 2011), and learning (see Jarosz 2010). Of these, perhaps the most popular is **Maximum Entropy Grammar** (MaxEnt Goldwater and Johnson 2003).

→ Maximum Entropy Grammars realize a number of key features of connectionist networks: i) trainable via perceptron update rule and ii) gradient outputs reflecting the interaction of multiple, soft, violable constraints

# Maximum Entropy

→ Given an input (here, verb stem), the model assigns a probability distribution over potential outputs.

→ Step one: **Compute Harmony score:** *H* . This is the weighted sum of all features / constraints of an output.

→ Step two: **SoftMax the Harmony to get probability of an output**. SoftMax here is just 'proportion of exponential Harmony over all candidates' exponential Harmony'

| hit | Past -> "ed"<br>3 | "Ted"<br>-5 | Past=hit<br>1 | *H* | *exp(H)* | *p* |
|---|---|---|---|---|---|---|
| hit | | | 1 | 1 | 2.72 | 0.95 |
| hitted | 1 | 1 | | -2 | 0.135 | 0.05 |

# Maximum Entropy

→ Learning can be approximated by gradually adjusting weights to bring model's predictions in line with observed data: The **perceptron update rule** (for details see Johnson, 2007).

→ Step one: **Compute difference between predicted and observed:**

| | | Past -> "ed" | "Ted" | Past=hit |
|---|---|---|---|---|
| Observed datum | hit | | | 1 |
| Learner's prediction | hitted | 1 | 1 | |
| Difference | | –1 | –1 | +1 |

→ Step two: **Adjust weights**

| hit | Past -> "ed"<br>2.9 | "Ted"<br>-5.1 | Past=hit<br>1.1 | *H* | *exp(H)* | *p* |
|---|---|---|---|---|---|---|
| hit | | | 1 | 1.1 | 3 | 0.96 |
| hitted | 1 | 1 | | -2.2 | 0.11 | 0.04 |

# German plurals

| | Masculine | Feminine | Neuter |
|---|---|---|---|
| Common | [-e]<br>[¨-e]<br>[ - ] | [-en]<br>[-n]<br>[-nen] | [-e]<br>[ - ] |
| Less common | [-en]<br>[-n]<br>[ ¨ ] | [¨-e] | [-er]<br>[ ¨ -er] |
| Adopted foreign words | [-s] | | |

| singular | plural |
|---|---|
| das Haus (the house) | die Häuser (the houses) |
| der Student (the student) | die Studenten (the students) |
| die Hand (the hand) | die Hände (the hands) |
| das Hobby (the hobby) | die Hobbys (the hobbies) |
| die Mutti (the momma) | die Muttis (the mommies) |

- Applies to ~7% of nouns.
- But is (over)extended in childhood, and productively applied to unusual nouns, and ‘exocentric’ nouns (as in e.g. English denominal ring)

# German plurals

| | Masculine | Feminine | Neuter |
|---|---|---|---|
| Common | [-e]<br>[¨-e]<br>[ - ] | [-en]<br>[-n]<br>[-nen] | [-e]<br>[ - ] |
| Less common | [-en]<br>[-n]<br>[ ¨ ] | [¨-e] | [-er]<br>[ ¨ -er] |
| Adopted foreign words | | [-s] | |

| singular | plural |
|---|---|
| das Haus (the house) | die Häuser (the houses) |
| der Student (the student) | die Studenten (the students) |
| die Hand (the hand) | die Hände (the hands) |
| das Hobby (the hobby) | die Hobbys (the hobbies) |
| die Mutti (the momma) | die Muttis (the mommies) |