Connectionist Language Processing

Lecture 5: **Acquisition of the English Past Tense**

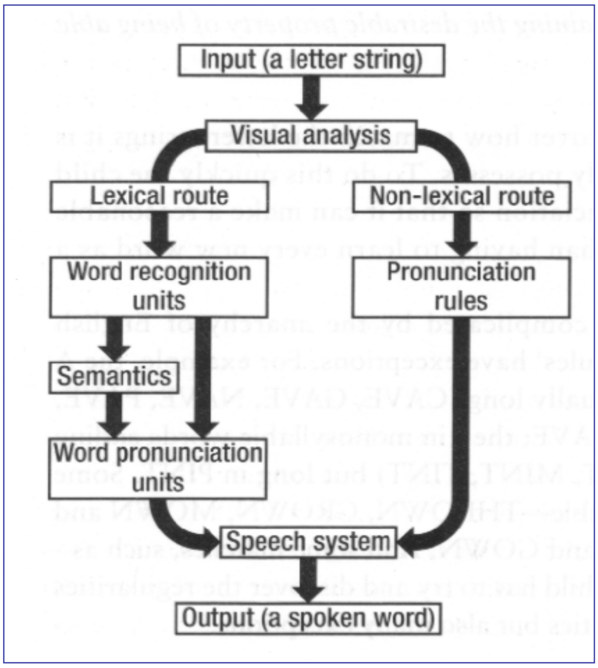
Matthew W. Crocker
crocker@coli.uni-sb.de
Harm Brouwer
brouwer@coli.uni-sb.de

Reading Aloud

- **Task**: produce correct pronunciation for a word, given its printed form
- Suited to connectionist modeling:
  - Need to learn mappings from one domain (orthography) to another (sound)
  - Multi-layer networks are good at this, even when mappings are arbitrary
  - Human learning is similar to network learning:
    - I.e. learning takes place gradually over time
    - Incorrect attempts are often corrected
- If a network can't model this linguistic task successfully, it would be a serious blow to connectionist modeling. But …

Dual Route Model

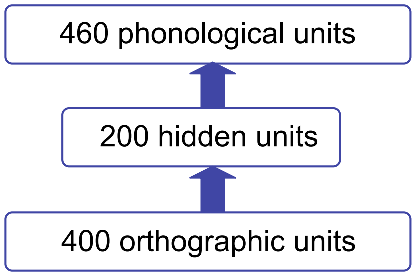
- The standard model of reading posits two independent routes leading to pronunciation of a word, because …
    - People can easily pronounce words they have never seen:
        - SLINT or MAVE
    - People can pronounce words which break the "rules":
        - PINT or HAVE
- One mechanism uses general rules for pronunciation
- The other mechanism stores pronunciation information with specific words



Evidence for Dual-Route Model

- Evidence from neuropsychology shows different patterns of behaviour for two types of brain damage that are acquired after learning
- Phonological dyslexia
    - **Symptom**: Read words without difficulty, but cannot produce pronunciations for non-words
    - **Explanation**: Damage to rule-based route; lexical route intact
- Surface dyslexia:
    - **Symptom**: Can pronounce words and non-words correctly, but tend to regularise irregulars
    - **Explanation**: Damage to the lexical route; rule-based route intact
- All Dual-Route models share:
    - A lexicon for known words, with specific pronunciation information
    - A rule mechanism for the pronunciation of unknown words

Seidenberg & McClelland (1989)



- Network behaviour is a function of experience
  - Reflects previous experience on a particular word
  - Experience with words resembling that string
  - Experience with HAVE overcomes the fact that _AVE is usually a long vowel
- Can produce a pronunciation for MAVE, but error is introduced by words like HAVE
- Performance: 97% accuracy on pronouncing learned words
  - Models: frequency & interaction with regularity, neighborhood, consistency
- Reading non-words (model gets 60%, humans 90%)
- Lexical decision (FRAME is a word, but FRANE is not)

Representations are important

- **Position specific** for inputting words of maximum length N: N groups of 26 binary inputs = word
- But consider: LOG, GLAD, SPLIT, GRILL, CRAWL
  - The model needs to learn mapping between L and /l/, for L in different positions
  - Learning pronunciations for different positions should be straightforward
  - Alignment: letters and phonemes are not in 1-to-1 correspondence
- **Non-position-specific** loses order important information: RAT = ART = TAR
- **Solution:** S&M decompose word and phoneme strings into "triples"
  - FISH = _FI SH_ ISH FIS
  - Each input unit is associated with 1000 random triples
  - Active if that triple appears in the input word

Wickelfeatures

- S&M still suffer some specific effects: Information learned about a letter in one context is not easily generalized

Improving S&M Model:Plaut *et al*

- Plaut et al (1996) solution: non-position-specific + linguistic constraints
  - Monosyllabic word = onset + vowel + coda
  - Strong constraints on order within these clusters:
    - E.g, if ‘t’ and ‘s’ are together, ‘s’ always precedes ‘t’
  - Only one set of grapheme-to-phoneme units is required for the letters in each group
  - Correspondences can be pooled across different words, even when letters appear in different positions

Improving the Model: Plaut *et al* (1996)

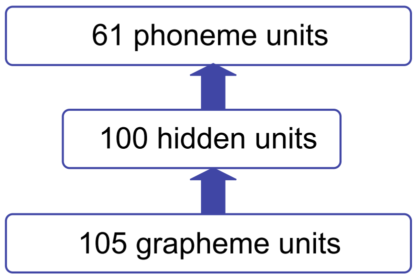
- Input representations:
  - Onset: first letter or consonant cluster (30)
    - y s p t k q c b d g f v j z l m n r w h ch gh gn ph ps rh sh th ts wh
  - Vowel (27)
    - e I o u a y ai au aw ay ea ee ei eu ew ey ie oa oe oi oo ou ow oy ue ui uy
  - Coda: final letter or consonant cluster (48)
    - h r l m n b d g cxf v j s z p t k q bb ch ck dd dg ff gg gh gn ks ll ng nn ph pp ps rr sh sl ss tch th ts tt zz u e es ed
- Monosyllabic words are spelled using one or more candidates from each of the 3 groups:
  - THROW: (‘th’ + ‘r’), (‘o’), (‘w’)

Output representations

- Output Representations
- Phonology: groups of mutually exclusive members
  - Onset (23)
    - s S C
    - z Z j f v T D p b t d k g m n h
    - l r w y
  - Vowel (14)
    - a e i o u @ ^ A E I O U W Y
  - Coda (24)
    - r s z
    - l f v p k
    - m n N t
    - b g d S Z T D C j
    - ps ks ts
- "Scratch" = 's k r a _ _ _ _ _ _ _ _ _ C'

The network architecture

- The architecture of the Plaut et al network:
  - The are a total 105 possible orthographic onsets, vowels, and codas
  - The are 61 possible phonological onsets, vowels and codas
- Performance of the Plaut et al model:
  - Succeeds in learning both regular and exception words
  - Produces the frequency x regularity interaction
  - Demonstrates the influences of frequency and neighbourhood size
- What is the performance on non-words?
  - For consistent words (HEAN/DEAN): model (98%) versus human (94%)
  - For inconsistent words (HEAF/DEAF/LEAF): model (72%), human (78%)
    - This reflects production of regular forms: both human & model produced both
- Highlights the importance of encoding … how much knowledge is implicit in the coding scheme



Summary

- Seidenberg & McClelland trained based on the log frequencies of words
  - People learn from absolute frequencies which: low frequency items too rare?
  - Plaut *et al* model, however, succeeds with absolute frequencies
- The right encoding scheme is essential for modeling the findings
  - How much linguistic knowledge is "given" to the network by Plaut's encoding?
  - They assume this knowledge could be partially acquired prior to reading
    - I.e. children learn to pronounce "talk" before they can read it
  - Doesn't scale to polysyllabic words
- Does not explain the double dissociation:
  - ✓ Surface dyslexics (can read exceptions, but not non-words)
  - ✗ Phonological dyslexics (can pronounce non-words, but not irregulars)

Connectionist models of Acquisition

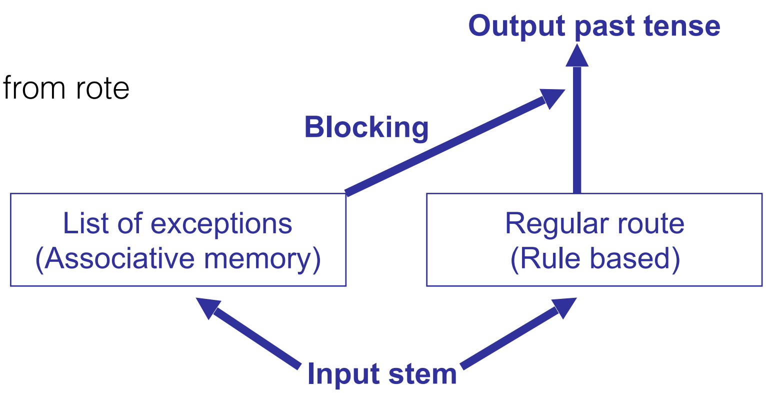
- Symbolic models emphasise the learning of rules and exceptions
- Connectionist models have no direct correlate to such mechanisms
  - Knowledge is stored in a distributed weight matrix, learned from experience
- Models of learning:
  - Start state of the cognitive system
  - Learning mechanism
  - Training environment
  - Acquired skill
- Connectionist models provide an opportunity to model the learning process itself, not just the resulting acquired skill
  - We can test models against developmental data, at various points during learning
  - Discontinuities in performance (sudden changes in behaviour) can be explained by "emergent properties" of a single, continuous mechanism

Learning the Past Tense

- The problem of past tense formation:
  - Regular formation: stem + 'ed'
  - Irregulars do show some patterns:
    - No-change: hit » hit (all end in a 't' or 'd')
    - Vowel-change: ring » rang, Sing » sang (rhymes often share vowel-change)
    - Arbitrary: go » went
- Young children often form the past tense of irregular verbs (like GO) by adding ED: <u>overregularisations</u>
  - *"go"+"ed" » "goed"*
  - Suggests incorrect application of a <u>learned rule</u>, not just rote learning or imitation
- Overregularisations often occur after the child has already succeeded in producing the correct irregular form: "***went***"
- Thus we need to explain this "U-shaped" learning curve

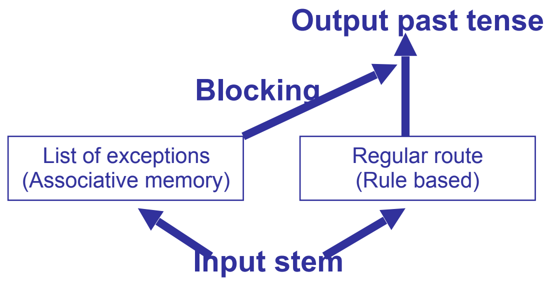
A Symbolic Account: Dual-Route Model

- General pattern of behaviour:
  - Early: children learn past tenses by rote (forms are stored in memory)
  - Later: recognise regularities, add general device to add 'ed' suffix
  - Now: no need to memorise forms, but this leads to incorrect generalisation of the regular rule to irregulars
  - Finally: distinguish which forms can be generated by the rule, and which must be stored (and accessed) as exceptions
- A Dual Route Model:
  - Errors result during the transition from rote learning to rule-governed
  - Recovery occurs after sufficient exposure to irregulars:
    - Increased "strength"
    - Frequency based
    - Faster recovery for frequent irregulars



The Dual-Route Model

- As with reading aloud, this proposal requires two qualitatively different types of mechanism



- Accounts for the observed dissociation:
  - Children make mistakes on irregulars only
- Evidence for double dissociation (Pinker 1994)
  - In some language disorders, children preserve performance on irregulars but not regulars
  - In other disorders, the opposite pattern is observed
- Accounts for the U-shaped learning curve
  - And since irregulars differ in "representational strength" it explains why overregularisation of high frequency irregulars is uncommon
- No explicit account of how the "+ed" rule is learned

Language Acquisition

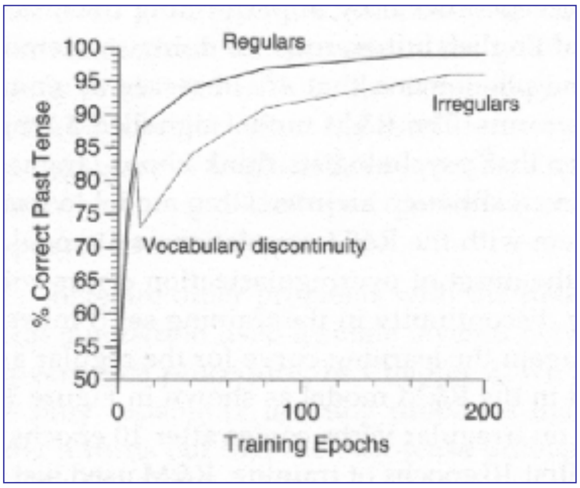
- Perhaps the notion of inflection is innately specified, and need not itself be learned:
  - The inflectional mechanism is triggered by the environment or maturation
  - Then the exact (language specific) manifestation must be learned
- Criticisms:
  - Early learning tends to be focussed on irregular verbs
  - Irregular sub-classes (hit, sing, ring) might lead to incorrect rule learning
    - Do occur, but typically late in learning
    - How are good/spurious rules distinguished and selected
  - English is unusual in possessing a large class of regular verbs
    - Only 180 irregulars
  - Only 20% of plurals in Arabic are regular
  - Norwegian has 2 regular forms for verbs: 3 route model ?

Towards a Connectionist Model

- No distinct mechanisms for regular and irregular forms
- No innately specified maturation stage, no rules to be triggered
- Parsimonious:
  - Simplifies the structural complexity of the starting state
  - Learning exploits the structure of the learning environment
- Rummelhart and McClelland (1986)
  - 1st attempt to model this problem (or any development system)
  - Modelled U-shaped learning, but heavily criticised (Pinker & Prince 1988)
- Plunkett & Marchman
  - Use a feed-forward network, one hidden layer

Rummelhart and McClelland (1986)

- A single-layer feed-forward network (perceptron)
  - Input: is a phonological representation of the stem (wickelfeatures)
  - Output: is a phonological representation of the past tense (wickelfeatures)
  - Trained using the perceptron learning rule
- Training:
  - First trained on 10 high frequency verbs (8 irregular, 2 regular), 10 epochs
  - Perfect performance
  - Then 420 (medium frequency) verbs (80% regular), 190 epochs
  - Early in training, shows tendency to overregularise, i.e. modelling stage 2
  - End of training, exhibits “adult” (near perfect) performance
  - Generalised reasonably well to 86 low frequency verbs in test set



Performance of R&M (1986)

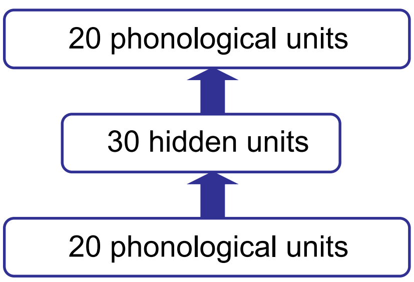
- Criticisms:
  - Problems with representation using wickelphones/wickelfeatures
  - U-shape depends on sudden change from 10-420 in the training regime
  - Rote learning of 1st 10 verbs: no generalisation to novel stems after 10 epochs
  - Most of the 410 new verbs are regular:
    - overwhelming the network and leading to overregularisation

- Justification: children do exhibit vocabulary spurt at end of year 2
  - But overregularisation errors typically occur at end of year 3
  - Vocabulary spurt is mostly due to nouns

- Single layer Perceptron only works for linearly separable problems
  - Plunkett & Marchman (1991) show residual error remains after extensive training
  - Suggests a hidden-layer network

Plunkett and Marchman (1993)

- A standard feed forward network with one hidden layer



- Maps a phonological representation of the stem to a phonological representation of the past tense

- Initially, the model is trained to learn the past tense of 10 regular and 10 irregular verbs
  - Represents currents estimates of children's early vocabulary

- Training proceeds using the standard backprop algorithm, in response to error between actual and desired output
  - Is this developmentally plausible?

- Learning must configure the network for both regulars and irregulars
  - Consider: *hit » hit*, but *pit » pitted*
  - We know multi-layer networks can do this, but considerable training may be required

Plunkett and Marchman (continued)

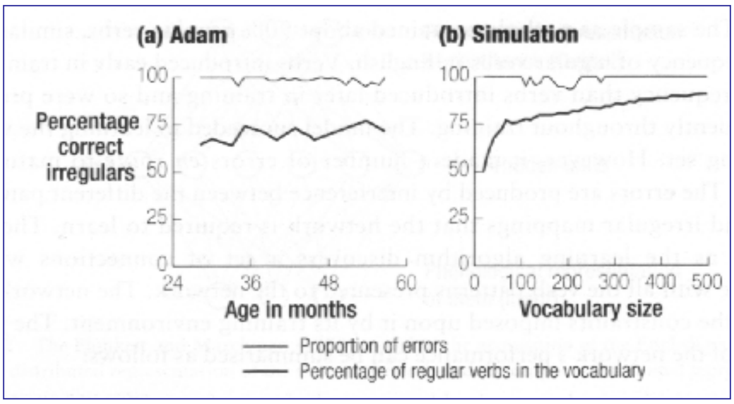
- Training:
  - Initial period of 10 regular and 10 irregular verbs
  - Then vocabulary was gradually increased, to mimic the gradual uptake of words in children
  - Total: 500 word stems, 90% regular (similar to the relative frequency of regulars in English)
  - Higher frequency verbs were introduced earlier in training, and so were also presented to the network more often
    - Irregulars are more frequent, so appear more often in training
    - This is essential, otherwise the regulars swamp the network
    - Arguably more accurately reflects the childs learning environment

- The final model successfully learned the 500 verbs in the training set
  - But errors were made *during* the learning phase
  - Caused by interference between mappings for regulars and irregulars before mature connection weights have been discovered

Performance of P&M

- Early acquisition is characterised by a period of error free performance
- Low overall rate (5-10%) of overregularisation errors
- Overregularisation is not restricted to a particular period of development
- Common irregulars do not exhibit overregularisation (e.g. ‘goed’ is rare)
- Errors are phonologically conditioned: No change verbs (hit) are robust to overregularisation (e.g ‘hitted’ is rare)
- Only a very small number of irregularisation errors are observed (e.g. where the network produces ‘bat’ for ‘bite’)

- Generally compatible with the results of studies by Marcus *et al* (1992):
  - Early performance is error free, and then low error is more or less random



Discussion

- Performance is closely tied to the training environment:
  - Onset of overregularisation is tied to a "critical mass" of regulars entering the child vocabulary
  - This subsides as the training learns the final solution for the task
- Highly sensitive to training environment:
  - Requires more training on arbitrary irregulars (go/went), which are highly frequent
  - Robust for no-change verbs (hit, put) which are more numerous (type) and less frequent (token)
- Models the frequency x regularity interaction:
  - Faster reaction time for high frequency irregulars than low frequency ones
  - No frequency advantage for regulars
- Differential behaviour for regulars and irregulars result from lesioning
- Suggests it is dangerous to infer dissociations in mechanisms due to observed dissociations in behaviour
  - Critical mass effects during learning can have the appearance of a distinct mechanism

Criticism

- We know multi-layered networks can learn such mappings in general; not proof that children use the same type of mechanism
- Pinker & Prasada argue that the (idiosyncratic) statistical properties of English help the model:
  - Regulars have low token frequency but high type frequency: facilitates the generalisation across this class of items
  - Irregulars have low type frequency but high token frequency: facilitates rote learning mechanism for these words
- They argue no connectionist model can accommodate default generalisation for a class which has both low type and token frequency
  - "Default" inflection of plural nouns in German appear to have this property
- No explanation of the double-dissociation observed by Pinker (1994)

Connectionist Language Processing – Crocker & Brouwer

Main conclusions

- Dissociations in performance, do not necessarily entail distinct mechanisms:
  - **Reading aloud:** a singe mechanism explains regular and irregular pronunciation of monosyllabic rules
  - **Past tense:** a single model of regular and irregular past tense formation

- But, explaining double dissociations is difficult
  - Has been shown to be possible in small networks, but unclear if/how larger (more plausible) networks can demonstrate double dissociations

- Connectionist models excel at finding structure and patterns in the environment: "statistical inference machines"
  - The start state for learning may be relatively simple, unspecified
  - Constraints to aid/determine learning come from the environment

- Can such models scale up? Are they successful for languages with different distributional properties?

- Reference: The English Past Tense, chapter 11 of Plunkett & Elman