

## Probabilistic Models of Human Linguistic Processing

Dan Jurafsky

*Department of Linguistics, Department of Computer Science,  
Institute of Cognitive Science &  
Center for Spoken Language Research*

*University of Colorado, Boulder*

This talk summarizes joint work with Alan Bell, Eric Fosler-Lussier,  
Susanne Gahl, Daniel Gildea, Cynthia Girand, Michelle Gregory, Lise Menn,  
Srini Narayanan, William D. Raymond, Doug Roland, Patrick Schone, and  
others.

Ohio State, May 2002

1

## Suggestive Facts

- Language and speech input is noisy, ambiguous, and unsegmented
- In other fields, probability theory is standard way to deal with these problems.
- Comparison: Association for Computational Linguistics 2000: 77% of papers probabilistic
- Psycholinguistics: Of 6 in-print college psycholinguistics textbooks, 0 have the word 'probability' in index.
- Linguistics: ???

Ohio State, May 2002

2

## Probability is not really about numbers; it is about the structure of reasoning –Glenn Shafer

- Probability theory is best normative model for solving problems of decision-making under uncertainty
- But perhaps a good normative model, but bad descriptive one?
- Perhaps human language is simply non-optimal, non-rational process?

Ohio State, May 2002

3

## Emerging Consensus

- Human cognition is rational, relies on probabilistic processing
- Anderson (1990): Bayesian underpinnings to memory, categorization, causation
- Linguistics: probabilistic models in phonology (Albright, Antilla, Beckman, Boersma, Hammond, Hayes, Pierrehumbert, Zuraw)

Ohio State, May 2002

4

### What Probability is not (necessarily)

- Fodoristicly modular or non-modular (both kinds of probabilistic models exist)
- Symbolic or Connectionist (both kinds of probabilistic models exist)
- Committed to Early or Late use of information (probabilities can be independent of time course)

### So what are implications of probabilistic model?

- **Comprehension:** More probable linguistic structures are accessed with less time, effort, evidence, preferred in disambiguation, cause less processing difficulty.
- **Production:** More probable structures are accessed faster and preferred in production choice.
- **Learning:** The probabilities of linguistic structures in context play a role in grammar/lexical learning.

### Outline of Talk

- **I: Comprehension**
- **II: Production**
- **III: Challenges to the Probabilistic Model**
- **IV: Learning**

### Part I: Comprehension

Bayesian Model of Comprehension

$$P(\text{interpretation}|\text{evidence}) = \frac{P(\text{evidence}|\text{int.})P(\text{int.})}{P(\text{evidence})} \quad (1)$$

## Lexical Frequency in Comprehension

- V Howes and Solomon (1951): tachistoscopic presentation of iteratively longer duration; HF words recognized with less presentation.
- V Forster and Chambers (1973): HF word named more rapidly.
- V Rubenstein et al (1970): LD faster to HF words.
- V Howes (1957): Words masked by additive noise. High-frequency words identified better.
- A Savin (1963): recognition errors biased toward words higher in frequency than presented words.
- A Grosjean (1980) gating, HF words recognized earlier
- A Replicated crosslinguistically, for example Tyler (1984) in Dutch. Many other methods, including fixation, gaze duration, recall.

## Frequency of Syn/Sem Category in Comprehension

- Simpson and Burgess (1985): HF sense of homograph prime causes faster reponse latencies to related target than LF sense.
- Gibson (1991) Low frequency syncats cause garden path:
  - (2) The old man the boats. (*man/N* > *man/V*)
- Jurafsky (1992,1996): Not just ranking; frequencies can combine:
  - (3) The complex houses married and single students and their families. (*complex/A* > *complex/N* and *house/N* > *house/V*)
  - (4) The building houses married and single students and their families. (better)

|                |                  |     |      |    |
|----------------|------------------|-----|------|----|
| <i>house</i>   | <b>Noun</b>      | 391 | Verb | 8  |
| <i>complex</i> | <b>Adjective</b> | 60  | Noun | 30 |

## Constructional Frequencies in Comprehension

- Relative rarity of reduced relative clauses could play role in their difficulty.
- Tabossi *et al.* (1994) showed reduced relatives are rare (8% of -ed forms occur in reduced relatives)
- Jurafsky (1996), McRae *et al.* (1998), Narayanan and Jurafsky (1998), among others showed that MC versus RR frequencies help predict reading time difficulties in MC/RR sentences.
- Jurafsky (1996): SCFG probability for MC lower than RR:
  1. RR construction includes one more SCFG rule
  2. This SCFG rule for RR has very low probability.

## Syntactic Subcategorization Frequencies

- Fodor (1978), Ford *et al.* (1982), Clifton, Jr. *et al.* (1984), Tanenhaus *et al.* (1985)
  - (5) The doctor remembered [<sub>NP</sub> **the idea**].
  - (6) The doctor remembered [<sub>S</sub> that the idea had already been proposed].
  - (7) The doctor suspected [<sub>NP</sub> the idea].
  - (8) The doctor suspected [<sub>NP</sub> **that the idea would turn out not to work**].
- Trueswell *et al.* (1993): cross-modal naming latency to noun *him* longer after S-bias verbs (*The old man suspected...him*) than after NP-bias verbs (*The old man remembered...him*).

## Summary: Converging Evidence for a Probabilistic Model of Comprehension

- Lexeme frequencies (Tyler 1984; Salasoo and Pisoni 1985; inter alia)
- Lemma frequencies (Hogaboam and Perfetti 1975; Ahrens 1998;
- Idiom frequencies (d'Arcais 1993)
- Phonological probabilities (Pierrehumbert 1994, Hay, Pierrehumbert and Beckman (in press), Pitt and McQueen (1998)
- Word transition probabilities (MacDonald (1993), Bod (2001), McDonald, Shillock and Brew (2001)
- Lexical category frequencies (MacDonald 1993, Jurafsky 1996)
- Constructional frequencies (Croft 1995; Mitchell *et al.* 1995; Jurafsky 1996)
- Subcategorization probabilities (Ford, Bresnan, Kaplan (1982); Clifton *et al.* (1984) Trueswell *et al.* (1993); Jurafsky (1996)
- Thematic role probabilities (Trueswell *et al.* 1994; Garnsey *et al.* 1997, McRae *et al.* (1998))

Ohio State, May 2002

13

## Jurafsky (1996) early model: Syntactic probabilities

- Build multiple interpretations of input in parallel
- Rank interpretations by their probabilities
- Probabilities computed from:
  - Stochastic context-free grammar probability
  - Subcategorization probability of predicates
- Limited memory causes low-probability interpretations to be pruned.
- Accounts for various types of garden-path effects (MV/RR, lexical category)
- Minus: only makes very broad reading-time predictions, only tested on handful of examples, only handles syntactic garden-paths

Ohio State, May 2002

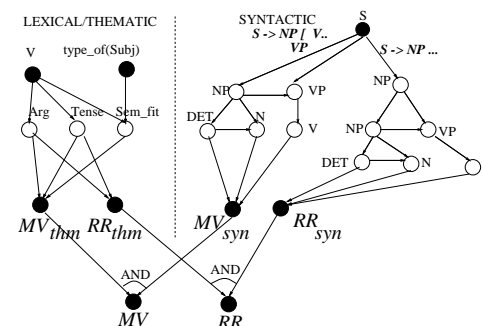
14

## More Sophisticated Model: Bayesian belief networks including semantics

- Narayanan and Jurafsky (1998): Bayesian belief network model of sentence processing
- Model of what probability to assign to a particular belief, how probability is updated on-line in light of new evidence.
- Why Bayes net? Allows representation of structured linguistic knowledge: SCFG probabilities with subcat, thematic, other lexical probabilities (potentially discourse, prosodic)

Ohio State, May 2002

15



- Predicts reading time increase whenever best interpretation is pruned
- Models McRae *et al.* (1998) results on reading time with MV/RR ambiguities.

Ohio State, May 2002

16

## Most Recent Model: More fine-grained predictions

Narayanan and Jurafsky (2001), (2002)

- Reaction time
  - $RT \propto \frac{1}{P(\text{input}|\text{context})}$
  - RT increases due to limited memory/attention
    - \* Beam search
    - \* Swap between best and other interpretations
- Preference
  - Rank of interpretation  $\propto P(\text{interpretation})$

## Part I: Comprehension: Conclusion

- Language comprehension is probabilistic
- Bayesian evidence-combination is one possible model
- **Key open problem:** building probabilistic models of linguistic knowledge.

## Part II: Probability and Production

Hypothesis: Speakers compute probability of linguistic structure in production as well!

## Previous Research: Frequency, predictability, and lexical production

- More frequent words more likely to have schwa vowels (Fidelholz 1975)
- More 'predictable' words are shorter (Lieberman 1963; Jespersen 1923)
- High-frequency collocations are more likely to have internal lenition (Bush 1999; Bybee 1995/2000)
- Methodological Conclusion: word duration reflects probabilistic effects in production.

## The Probabilistic Reduction Hypothesis

Jurafsky, Bell, Gregory, and Raymond (2000) The higher the probability of a word, the more it is reduced/shortened/lenited in lexical production.

### Explains:

- Phrase-frequency effect in word reduction (Bybee 1996, Krug 1998, Bush 1999)
- Word-repetition effect on word reduction (Fowler and Housum 1987)
- Unpredictable words should be *unreduced* or *longer* (Lieberman 1963, Bolinger 1981, Jespersen 1922)
- Frequency effect in word reduction (Fidelholz, Hooper)
- **Predicts:** any factor that increases probability of word also increases phonological reduction.

## 3 experiments on types of probability

Jurafsky *et al.* (1998), Gregory *et al.* (1999); Jurafsky *et al.* (2001); Bell *et al.* (2002)

1. Word frequency
2. Probabilistic relations between words
3. Lemma (word sense) frequency

## Local Probability

Jurafsky *et al.* (1998); Bell *et al.* (1999); Gregory *et al.* (1999); Jurafsky *et al.* (2001)

Relative Frequency (Prior Probability):

$$P(w_i) = \frac{\text{Count}(w_i)}{\sum_j \text{Count}(w_j)} = \frac{\text{Count}(w_i)}{\text{Total}} \quad (9)$$

Conditional Probability (Transitional Probability)

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \quad (10)$$

Conditional Probability from Next Word

$$P(w_i|w_{i+1}) = \frac{C(w_iw_{i+1})}{C(w_{i+1})} \quad (11)$$

## Examples of highly predictable and unpredictable words

| Highest Probability Given Previous Word<br>$P(w_i w_{i-1})$ | Lowest Probability Given Previous Word<br>$P(w_i w_{i+1})$ |
|---|--|
| rid <b>of</b>   | you're <b>to</b>   |
| supposed <b>to</b>  | have <b>of</b>   |
| tends <b>to</b>   | at <b>to</b>   |
| ought <b>to</b>   | was <b>of</b>  |
| kind <b>of</b>  | done <b>of</b>   |
| able <b>to</b>  | well <b>to</b>   |
| sort <b>of</b>  | because <b>to</b>  |
| compared <b>to</b>  | feel <b>to</b>   |

## The Corpora

- Switchboard corpus of 2430 American English (1991) telephone conversations
- 3.5 hours (38K words) phonetically labeled (Greenberg et al. 1996)
- Two Datasets:
  - **function words:** 5618 tokens of 10 most frequent English function word in Switchboard *I, and, the, that, a, you, to, of, it*, and *in* (of  $\approx 9000$  total)
  - **content words:** 2042 tokens of words ending in (underlying) t or d (of  $\approx 3000$  total)
- We coded multiple measures of reduction/lenition/shortening:
  - duration in milliseconds
  - (for function words only): vowel reduction: full or reduced
  - (for content words only): deletion of final t/d

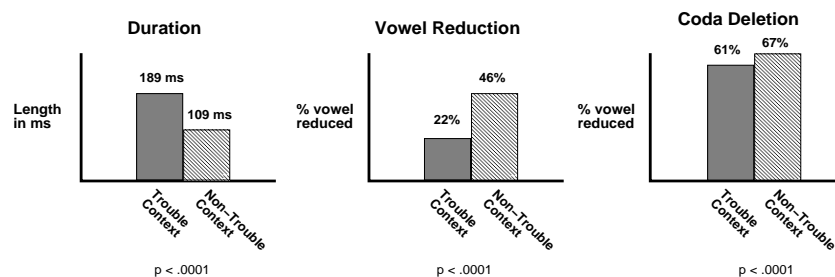
## But wait!!!

We can't just look at raw durations of words!!!

- words tend to be longer in the context of disfluencies (pauses, repetitions)
- length and reduction are sensitive to identity neighboring phonemes
- accent/stress affect word length/reduction
- words are longer at phrase/turn/clause boundaries
- words are shorter if speaker is speaking faster!
- plus many other factors!!!!

## Example of a confound: Words near disfluencies are longer

Bell, Jurafsky, Fosler-Lussier, Girand, Gildea, Gregory (2001)



## Methodology

- Test if factor causes reduction by running **regression**
- First we excluded word tokens based on
  - phrase boundary positions (Keating and Fougeron)
  - special forms (cliticized *you've*), *an*)
  - polysyllables (for duration results, only used monosyllables)
  - accent (for one experiment)
- We then control for 'base model':
  - rate of speech (syl/sec)
  - planning problem (preceding or following disfluency) (Tree and Clark 1997; Jurafsky et al. 1998; Bell et al. 1999)
  - preceding and following segments (C or V, clusters) (Guy 1980, etc)
  - syllable type (open or closed) (e.g. *it* vs. *a*)
  - reduction of following vowel
  - (for content words) inflectional status (Fasold 1972, Labov 1972, Bybee 2000) (t more likely deleted in *mist* than *misted*)
  - (for content words) Identity of underlying segment (t or d)
  - Number of syllables (for (content word) deletion)
  - Number of phones (for content word)

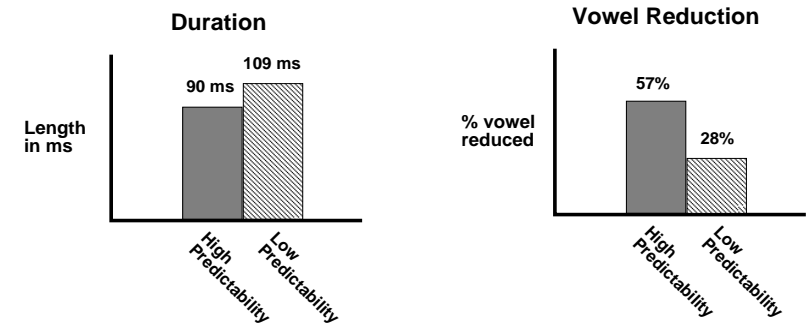
## Results: More frequent words are shorter

Gregory *et al.* (1999); article in progress

- Word frequency plays a significant role in t/d-deletion and duration.
- Highest frequency words are 22% shorter than lowest frequency words. ( $p < .0001$ )
- Highest frequency words are 11 times more likely to have t/d-deletion than lowest frequency words.

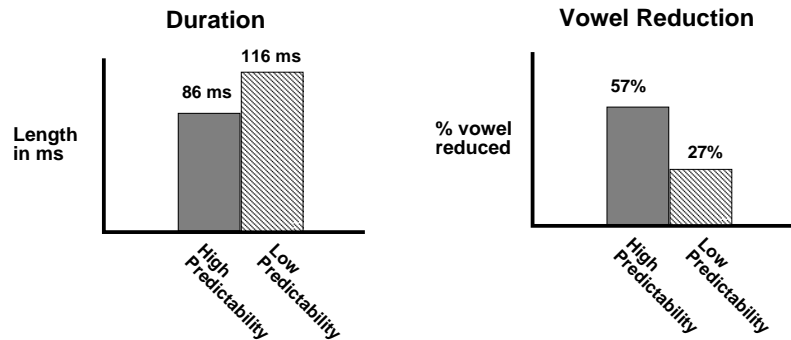
## Results: Predictable (from left) function words are shorter and more reduced

Jurafsky *et al.* (2001), Jurafsky, Bell, Fosler-Lussier, Girand, Gildea, Gregory (2001)



## Results: Predictable (from right) function words are shorter and more reduced

Jurafsky *et al.* (2001), Jurafsky, Bell, Fosler-Lussier, Girand, Gildea, Gregory (2001)



## Role of Lemma Frequency

Jurafsky, Bell and Girand (2002, to appear) (Labphon-7)

- Does the frequency of word sense (lemma) play a role in production?
- Evidence for word sense frequency in comprehension Hogaboam and Perfetti (1975); Li and Yip (1996); Ahrens (1998)
- Initial evidence for role for word sense in production (Berkenfield 2000 on *that*, Veilleux and Shattuck-Hufnagel (p.c.) on *to*)
- Initial evidence against word sense frequency in production (Jescheniak and Levelt 1994).
- Our results: no effect of lemma frequency on reduction after controlling for probability

## Summary from Part II: Lexical Production

- More frequent words are reduced
- Words which are predictable from neighboring words are reduced.
- Possible asymmetry (Bill Raymond's (2001) and ongoing work)
- The locus of frequency effects seems to be the lexeme rather than the lemma.
- Implications:
  - **Representation:** Mental grammar stores word-to-word predictability.
  - **Processing:** Speakers are computing probability of words given context.
  - **Speech synthesis:** Better models of function word synthesis

## Shattuck-Hufnagel's question: locus of probability effects?

Michelle Gregory's Dissertation (Gregory 2001)

- Where in production process do reduction effects happen (phonetic or phonological)?
- Does probability have similar effect on pitch accent placement?
  - 8757 switchboard words had been coded for pitch accent
  - Taylor TILT, Shattuck-Hufnagel POSH (variant of ETOBI (Beckman and Ayers 1997))
  - Do frequency and conditional probability predict pitch accent?

## Michelle's Experiment

- Use regression to control for factors
  - Surrounding disfluencies
  - Rate of speech
  - Number of phones
  - Surrounding phrase break
  - POS (N, V, A, function)
- Results
  - Frequent words are 25% less likely to have accent than infrequent words.
  - High Prob words are 25% less likely to have accent than infrequent words.

## Unsolved issue: what is cause of predictability effect?

Michelle Gregory's Dissertation (Gregory 2001)

1. What is the locus of the predictability effect?
  - Predictability is known to affect both *word duration* and also *likelihood of pitch accents*.
  - Are these the same phenomenon or different ones?
2. What is the cause of the predictability effect?
  - Speaker is modeling what the hearer can "figure out" (Jespersen 1923)?
  - Lexical priming for speaker? (Bard et al. 1999, Horton and Keysar 1996)
  - Both?

### Previous Work Inconclusive

- Fowler and Housum (1987): Words get shorter every time they are repeated
- Fowler suggests that speakers are modeling hearer knowledge
- Bard et al (2000), Brown and Dell, Horton and Keysar 1996 suggests not

### Michelle's Experiment

- A redesign of a cute idea due originally to Ellen Bard
- Initial condition: Speaker tells a story to hearer 1
- Followed by one of two conditions:
  - A: Speaker now tells story to hearer 2
  - B: Speaker retells story to hearer 1 again
- Michelle found that condition B is shorter than condition A
- Thus speakers shorten the words *more* if the hearer has heard them before
- Thus speakers **do** use model of hearer knowledge in articulation.

### Part III: Potential Challenges to Probabilistic Models

**Challenge 1: Surely you don't believe that people compute little symbolic Bayes equations in their heads?**

## No, I don't

- Probability theory is at Marr's 'computational level'
- Characterizes one aspect ('evidence combination') of the input-output properties of the computations that the mind must somehow be doing.
- How could this be realized at lower levels?
  - Most common assumption: activation level of some mental structure or a distributed pattern of activation.
  - Prior probabilities encoded as resting activation levels, or as weights on connections.
  - But also other possible realizations, such as exemplar models of phonology of Pierrehumbert (2001b) and others.

## Challenge 2: Subcategorization frequency from corpora doesn't match psychological experiments

- Subcategorization counts for individual verbs
- Counts from Sentence production/completion and corpora are correlated (Merlo 1994, Lapata et al. 2001)
- But correlation not that high: Merlo (1994):

|                       | NP         | S          |
|-----------------------|------------|------------|
| Trueswell vs. Garnsey | $r = .935$ | $r = .916$ |
| Trueswell vs. Corpus  | $r = .739$ | $r = .444$ |
| Garnsey vs. Corpus    | $r = .727$ | $r = .585$ |

## 2 major causes of differences in subcategorization frequencies

Roland and Jurafsky 1998, 2000, Roland *et al.* 2001, Roland 2001

- **Context-based Variation:** 'Isolated sentences' differ from 'connected discourse'
  - Null complementation in connected discourse
  - Passives in connected discourse
  - Full NPs in isolated sentences
- **Word-sense Variation:** Different *lemmas* have different subcategorization probabilities
  - Controlling for word sense seems to eliminate subcategorization differences between corpora

## Moral: True Probabilistic Model is not just Raw Frequency

Doug Roland's Dissertation (2001)

- Corpus observations produced by true probability model
- $p(\text{subcat} \mid \text{verb, verb sense, subject, object, previous sentence, conversation topic, speaker's model of hearer, ...})$
- If the conditioning events are different, the probability will be different
- In practice: can bin, remove outliers, etc (as in using Brown numbers for comprehension experiments)

### Challenge 3: Maybe frequency is just an epiphenomenon

#### Production of unaccusative verbs in English

- Unergatives      NP [<sub>VP</sub> V]              external argument, no internal
- Unaccusatives    — [<sub>VP</sub> V NP/CP]      internal argument, no external
- bloom, melt, blush, etc
- Kegl (1995): **unaccusatives are hard for agrammatic aphasics**
  - Production study of ‘agrammatic’ aphasic subject vs. control
  - Production showed significant absence of unaccusatives
  - Explanation: unaccusatives are like passives in involving traces
  - Agrammatic aphasics have difficulty with traces

### Alternative: Lexical subcategorization frequency

Susanne Gahl, Lise Menn, Gail Ramsberger, Dan Jurafsky, Beth Elder, Molly Rewega, Audrey L. Holland (2001), drawing from Menn et al 1998, Gahl 2000

- Hypothesis: *comprehension difficulty should vary with subcategorization frequency bias of word*
- Methodology: plausibility judgments in comprehension
- Sentences were transitive or intransitive, hence matching or not matching verb biases.
- Prediction: sentences should be easier if structure matches verb bias.
- Prediction: no reason to expect unaccusatives to act like passives
- 8 subjects in Tucson and Boulder, including 3 Broca's, 3 conduction, 2 anomics

### Preliminary results from Gahl et al (2001) study

- Unaccusatives as a whole are much easier than passives ( $p < .00001$ )
- Unaccusatives as a whole are not harder than unergatives ( $p = 0.691$ , n.s.)
- Sentences are easier when structures match subcat frequency bias of verb ( $p < .001$ )
- Thus Kegl's subject's difficulties with unaccusatives probably not due to traces

### Challenge 3b: Maybe frequency is just an epiphenomenon (II)

Stevenson and Merlo (1997) on comprehension in normals:

#### Unergatives are hard in reduced relatives:

The students advanced to the next grade had to study very hard.  
 The clipper sailed to Portugal carried a crew of eight.  
 The ship glided past the harbor guards was laden with treasure.

#### Unaccusatives are easier:

The witch melted in the Wizard of Oz was played by a famous actress.  
 The oil poured across the road made driving treacherous.

### Stevenson and Merlo Explanation

- An extension of Hale and Keyser (1993)
- Verbs project their phrasal syntax in the lexicon
- Causativized (transitive) forms of unergatives are more complex than causativized (transitive) forms of unaccusative verbs.
- More complex in terms of number of nodes and number of binding relations
- Stevenson (1994) parser cannot activate structure needed to parse transitivized unergatives because of limitations on creating and binding empty nodes.

### Alternative: Frequency Explanation for Unaccusative-Unergative Difference

*Stevenson and Merlo 1997, 1998, Gahl 1998, Gahl and Jurafsky 2000*

|              | Transitive |            | Intransitive |            |
|--------------|------------|------------|--------------|------------|
| Unergative   | 2869       | 13%        | 19194        | <b>87%</b> |
| Unaccusative | 17352      | <b>54%</b> | 14817        | 46%        |

Unergatives (like *race*) have a huge bias toward intransitive.

Frequency account also explains gradient effects (some unergatives easier than others) (Filip, Tanenhaus, Carlson, Allopenna, Blatt 2001)

### Part IV: Probability, Statistics, and Language Learning

#### Fundamental Problem of Learning Theory:

How to combine empirical knowledge (experience in the world) with rational knowledge (pre-existing learning biases)

- Purely nativist approaches: *Parameter Setting*
- Purely empirical approaches: *Error Back-Propagation, Instance-Based Generalization, Minimum Description Length*
- Our focus: *learning biases plus statistical induction*

### Nature + Nurture: Representing 'Learning Bias'

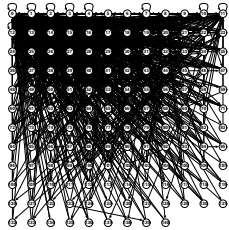
- **Minimum Description Length** (Brent and Cartright 1996, deMarcken 1997, Stolcke 1994, Stabler 2001)
- **Word Learning Biases** (Woodward and Markman 1991)
- **Vector Space Assumptions** (Landauer and Dumais 1997, Li, Burgess, and Lund 1998)
- **Faithfulness** Gildea and Jurafsky (1996)
- **Non-linguistic Biases** (Visual System: Regier 1996)

## Faithfulness Bias in Phonological Rule Induction

Gildea and Jurafsky (1996)

- Idea: machine learning to induce flapping rule from I/O pairs
- Two-level phonology: SPE rules (or OT rules) are equivalent to finite-state automata
- Using the OSTIA Automata-Induction Algorithm (Oncina93)
- Given 50,000 examples of flapping:

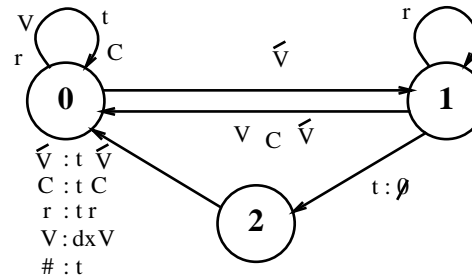
latter: l áe t axr → l áe D axr



- Failed to learn even the simple flapping rule

## After adding Faithfulness Bias

Told the system that, ceteris paribus, assume input and output are similar



Induced rule:  $t \rightarrow D / \hat{V} r^* \_ V$

## Inducing simple morphological structure

Pat Schone's dissertation (Schone 2001)

- How does a system (or human) learn that *-ed* or *-ing* are suffixes?
- Current models: based on one of two probabilistic approaches: MDL and transition probability

## Probabilistic Models of Word Segmentation

Saffran *et al.* (1996b); Saffran *et al.* (1996a)

- 8-month-old infants exposed to synthesized speech from artificial language
- Only cues to word boundaries: transition probability between syllables
- Probability was higher within words (1.0) than between words (0.33).
- Test items included words and concatenated part-words
- Infants reliably able to distinguish words from non-word sequences which had both occurred in training.

## Minimum-Description-Length (MDL) models of morphological segmentation

Brent *et al.* (1995), Goldsmith (2000)

- Split input into words and affixes with shortest resulting grammar:
- 1. talk, walk, wipe, -s, -ed
- 2. talk, talks, talked, walk, walks, walked, wipe, wipes, wiped
- Algorithm: try every possible split
- Search strategy: focus on regions of low trans prob

## Corpus-based induction of morphology: Semantics

Schone and Jurafsky (2000), Schone and Jurafsky (2001), Schone (2001)

- MDL uses no semantics: thinks *ally* is *all+y*, doesn't see relationship between *dirty* and *dirty*
- Solution: Add semantic knowledge, but not innate semantic knowledge: induced empiricist semantics.
- First find a metric for word semantic similarity
- Then split input only if stem and stem+affix are semantically close

## Induction of Lexical Semantics

Latent Semantic Analysis (LSA), Hyperspace Analog to Language (HAL) Landauer and Dumais (1997); Lund *et al.* (1995)

- Create a matrix from a large body of text
- One row for each word type
- One column for each passage
- Each cell contains frequency that word occurred in passage.
- Perform SVD on matrix, retaining only 100-300 high-order dimensions
- Result: a vector for each word:  
break 0.135 0.000 0.000 0.000 0.006 0.098 0.004...

## LSA modeling of Word Similarity

Landauer and Dumais (1997); Li *et al.* (1998)

- Distance between words is cosine between vectors:
- Words grouped by cosines:
  - break: break, spill, hit, fall, throw, hurt, pull, know, bang
  - breakfast: breakfast, supper, lunch, dinner, dessert, juice
  - duck: duck, cow, pig, bear, dog, doggie, horse, frog
- Sentences grouped by combining cosines of words:
  - cos ("The radius of spheres.", "A circle's diameter") = .55
  - cos ("The radius of spheres.", "The music of spheres") = .01.

## Pat's application of LSA to Morphology Induction

- LSA correctly identifies semantically-related pairs

| Pairs      | Distance | Pairs       | Distance |
|------------|----------|-------------|----------|
| car/cars   | 5.6      | ally/allies | 6.5      |
| car/caring | -0.71    | ally/all    | -1.3     |
| car/cares  | -0.14    | dirty/dirt  | 2.4      |
| car/cared  | -0.96    | rating/rate | 0.97     |

- Significantly better at finding morphologically (suffix) related words in CELEX (F-score):

| Algorithm       | English | German | Dutch |
|-----------------|---------|--------|-------|
| Goldsmith       | 62.8    | 75.8   | 74.2  |
| Schone+Jurafsky | 88.1    | 92.3   | 85.8  |

## Conclusion for Part IV: Probabilities and Learning

- Statistical learning from corpora
- Empiricist induction plus rationalist biases
- Able to learn phonological rules, simple concatenative morphology
- Others have shown similar approaches to grammar induction (Stolcke 1994, deMarcken 1997)
- Open questions:
  - Relation to phonological learning (Albright and Hayes; Beckman and Edwards; Boersma; Edwards, Beckman, Munson)
  - Relation to exemplar models (Pierrehumbert)

## Conclusion

- Probabilistic models offer a clean, motivated solution to evidence combination, disambiguation, and choice.
- Probability is not a *replacement* for structure, but an *augmentation*
- Key Future Work:
  - Current work mainly in phonology (Albright, Antilla, Beckman, Boersma, Hammond, Hayes, Zuraw) and syntax (Manning). What about semantics, pragmatics?
  - Testing different probabilistic models against each other
  - Relationship between probabilities and activation-based algorithms

## Other research Areas in my Lab

- Speech Recognition and Understanding
  - Pronunciation Modeling
  - Language Modeling
  - Dialog Modeling
  - Punctuation Detection
- Computational Psycholinguistics
  - Sentence Processing
  - Lexical Access in Production
  - Probabilistic Models of Human Language Processing
- Computational Linguistics and Natural Language Processing:
  - Semantic Parsing in English and Chinese
  - Verb-Argument Probabilities
  - Automatic Question Answering
  - Machine Learning of Natural Language
  - Computational Phonology