UCSD

Probabilistic Sentence Processing II
Probabilistic knowledge in human language comprehension and production

Roger Levy
University of California – San Diego

IPAM summer school: Probabilistic Models of Cognition
18 July 2007

I. Probability in language comprehension

- We'll look at three different kinds of effects
  - "Garden-pathing" effects
  - Expectation-versus-memory effects
  - Facilitative ambiguity effects

Garden-pathing

- Is this sentence understandable?
  The horse raced past the barn fell.
- How about this one?
  The evidence examined by the witness was forged.
- Does this help?
  The horse that was raced past the barn fell.
- These are *garden-path* sentences: they mislead you part-way through.
- The ambiguity is between *main-verb* and *reduced-relative* interpretations of the verb *raced*

Garden-pathing (2)

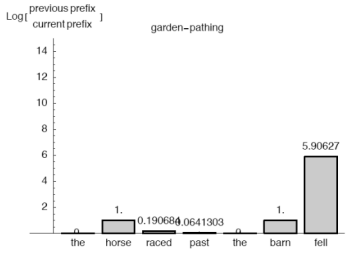
- We should see high surprisal values at the disambiguating word (*fell*)
- The surprisal of the disambiguating word is determined by marginalizing over the incremental interpretations

$$\begin{aligned} P(\text{fell}) &= \sum_{\text{I}} P(\text{fell}|I)P(I|w_{1\ldots i}) \\ &= \underbrace{P(\text{fell}|MV)}_{\textit{low likelihood}}\underbrace{P(MV|w_{1\ldots i})}_{\textit{high prior}} + P(\text{fell}|RR)P(RR|w_{1\ldots i}) \end{aligned}$$

- The high-prior interpretation has exceedingly low likelihood, leading to a high surprisal

Garden-Pathing (3)

- When a PCFG is used directly, the same effect is seen (Hale 2001)



I. Probability in language comprehension

- We'll look at three different kinds of effects
  - "Garden-pathing" effects
  - Expectation-versus-memory effects
  - Facilitative ambiguity effects

Memory retrieval-constrained processing

- On the traditional view, resource limitations, especially memory, drive processing difficulty
- Gibson 1998, 2000 (DLT): multiple and/or more distant dependencies are harder to process

Processing

the reporter who attacked the senator **Easy**

the reporter who the senator attacked **Hard**

Expt 1: Verb-final domains

- Konieczny 2000 looked at reading times at German final verbs

Er hat die Gruppe **geführt**
He has the group **led**
"He led the group"

Er hat die Gruppe auf den Berg **geführt**
He has the group to the mountain **led**
"He led the group to the mountain"

Er hat die Gruppe auf den SEHR SCHÖNEN *Berg* **geführt**
He has the group to the VERY BEAUTIFUL *mtn.* **led**
"He led the group to the very beautiful mountain"

Locality predictions and empirical results

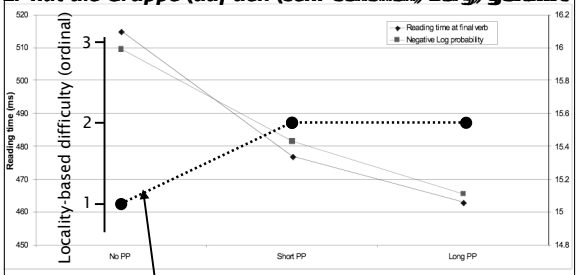
- Locality-based models (Gibson 1998) predict difficulty for longer clauses
- But Konieczny found that final verbs were read faster in longer clauses

	Prediction	Result
Er hat die Gruppe **geführt**	easy	slow
Er hat die Gruppe auf den Berg **geführt**	hard	fast
…die Gruppe auf den sehr schönen Berg **geführt**	hard	fastest

He led the group
He led the group to the mountain
He led the group to the very beautiful mountain

Predictions from surprisal model

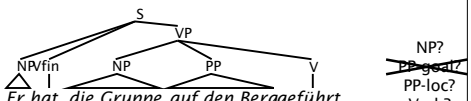
Er hat die Gruppe (auf den (sehr schönen) Berg) geführt



Locality-based models (e.g., Gibson 1998, 2000) would violate monotonicity

Deriving Konieczny's results

- Seeing more = having more information
- More information = more accurate expectations



~~NP?~~
~~PP-goal?~~
PP-loc?
Verb?
ADVP?

- Once we've seen a PP goal we're unlikely to see another
- So the expectation of seeing anything else goes up

- Rigorously tested: for $p_i(w)$, I used a PCFG derived empirically from a syntactically annotated corpus of German (the NEGRA treebank)

Disentangling verb *identity* and *location*

- General syntactic configuration of interest:
  - we know element X must appear, but we don't know exactly which X or where
- Head-final clauses satisfy this configuration
- PCFG model captures *where* the verb may appear
  - verb *location*
- But maybe it's knowledge of *which* verb may appear
  - verb *identity*; informal model proposed by Konieczny (also connectionist model of Konieczny & Döring 2003)
- Can these two be disentangled?
  - Yes!

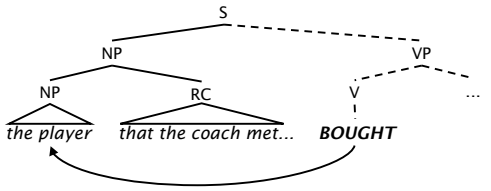
Expt 2: Verb identity vs verb position

The player (that) the coach met...

1: near the gym ***BOUGHT*** *the house*

2: near the gym by the river ***BOUGHT*** *the house*

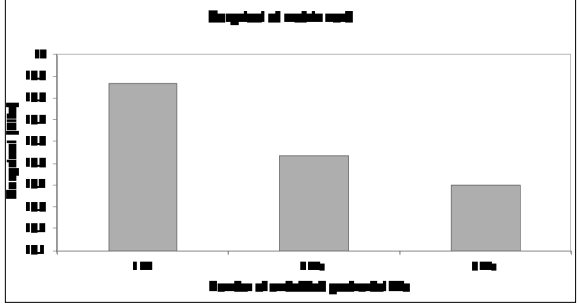
3: near the gym by the river at 8 o'clock ***BOUGHT*** *the house*



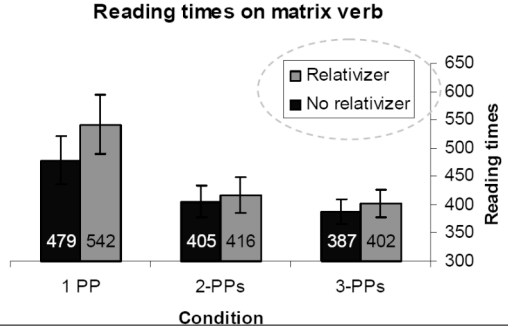
Jaeger, Fedorenko, and Gibson (CUNY 2005)

Predictions of surprisal

- PCFG derived from the parsed Brown corpus



Verb location but not identity



Upcoming-word expectations: summary

- Expectation-based processing model
  - *Surprisal* [-log $P_i(w)$] as estimate of difficulty
- How to calculate surprisal at each word using a PCFG syntactic model
- Modeling one result in German verb-final clauses, one in English matrix verbs
- Locality-based account gets it wrong
  - Expectation-based account gets it right
  - Evidence for verb *location* as well as *identity*

I. Probability in language comprehension

- We'll look at three different kinds of effects
  - "Garden-pathing" effects
  - Expectation-versus-memory effects
  - Facilitative ambiguity effects

When ambiguity facilitates comprehension

- Sometimes, ambiguity seems to *facilitate* processing:

The daughter$_i$ of the colonel$_j$ who shot himself$_{*i/j}$
The daughter$_i$ of the colonel$_j$ who shot herself$_{i/*j}$

slower ⇧

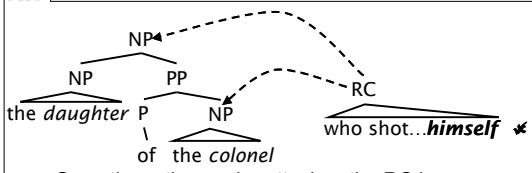
faster ⇩

The son$_i$ of the colonel$_j$ who shot himself$_{i/j}$

- Argued to be problematic for parallel constraint-based ***competition*** models (Macdonald, Pearlmutter, & Seidenberg 1994)
  - (though see rebuttal by Green & Mitchell 2006)

(Traxler et al. 1998; Van Gompel et al. 2001, 2005)

Traditional account: stochastic race model

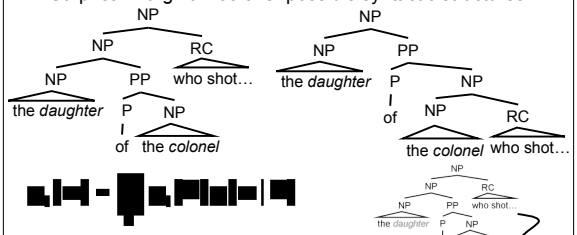


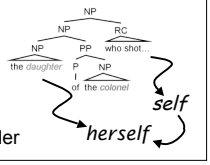
- Sometimes the reader attaches the RC low...
  - and everything's OK
- But sometimes the reader attaches the RC high…
  - and the continuation is anomalous
- So we're seeing garden-pathing 'some' of the time

(Traxler et al. 1998; Van Gompel et al. 2001, 2005)

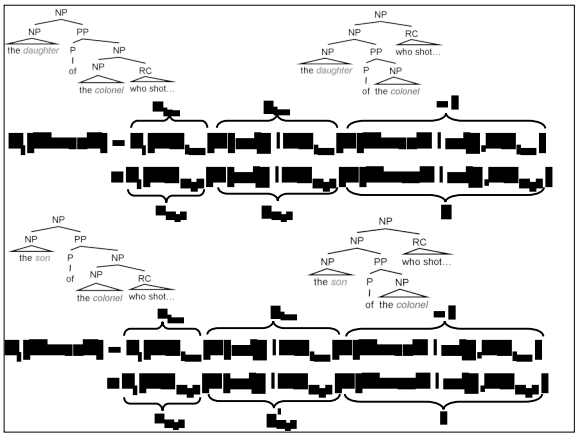
Surprisal as a parallel alternative

- Surprisal *marginalizes* over possible syntactic structures





- assume a generative model where choice between *herself* and *himself* determined only by antecedent's gender



Ambiguity reduces the surprisal

daughter…who shot… can't contribute probability mass to *himself*

But *son…who shot…* can

Surprisal and comprehension: summary

- People are sensitive to probabilistic information in language comprehension…
  - both in processing rate for upcoming events…
  - and in the management of ambiguity
- Surprisal is a unified measure of how this probabilistic information may mediate processing difficulty

II. Probability in language production

- Why do people talk the way they do?
- Linguistic communication involves transactions in uncertainty
- But it takes place under adverse conditions:
  - Auditory environment is noisy
  - People's working memory is limited
  - Environment competes for attention
  - Interlocutors have incomplete knowledge of each other
- Yet communication seems to work most of the time
- How is *redundancy* achieved?
- Micro-level study: speakers' choices in using a single, "meaningless" word

joint work with T. Florian Jaeger: Jaeger 2006, Levy & Jaeger 2006

The empirical phenomenon

- Certain types of *relative clauses* (RC) in English are optionally introduced by the word *that*

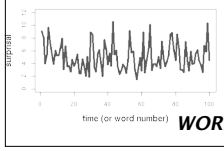
How big is the family (that) you cook for __ ?

modifies the noun *family* ← RC: *"you cook for the family"*

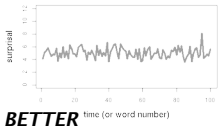
- Relative clauses are an important part of the infinite expressive capacity of human language (recursion)
- What governs use of the optional function word *that*?

Hypothesis about language use

- Language comprehension involves serial input
- Results from surprisal-based studies suggest that comprehenders find more predictable (=less informative) words and phrases easier
- Under some basic assumptions, it can be proven that spreading out information evenly in a sentence is communicatively optimal (Jensen's inequality)



WORSE



BETTER

Spreading out information in RCs

- In an RC without *that*, the first word does two things:

How big is the family you… (*phrasal predictability*)

1) It signals that a relative clause has begun
2) It signals some information about the contents of the relative clause

- Inserting *that* separates these two things:

How big is the family (1) *that* (2) *you…*

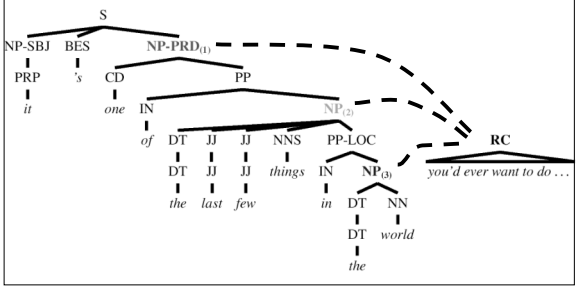
- Hypothesis: speakers should use *that* more when the RC's onset is informationally dense

Dataset

- Corpus of spontaneous telephone conversation by speakers of American English (*Switchboard* corpus)
- Roughly 1 million words of conversation have been annotated for linguistic structure
- Contains 3,452 datapoints (relative clauses for which *that* can potentially be omitted)

Probabilistic model of structural production

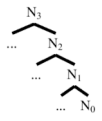
- We use tree structures to represent natural language structure and ambiguity as a sentence unfolds…



Calculating phrasal predictability

- The use of tree structure also gives us a recurrence relation expressing the predictability of an upcoming phrase in the tree:

$$P(RC_{n+1\ldots}|w_{1\ldots n}, T_{1\ldots n}) = \sum_{i=0}^{k}\left[\underline{P(RC|N_i)}\prod_{j=0}^{i-1}\underline{P(*END*|N_j)}\right]$$



we need to estimate these model parameters

The statistical problem

- There are two statistical questions to be addressed:
  1. How do we choose the phrasal predictability model $P(X|N_i)$?
  2. How do we assess whether phrasal predictability is associated with speakers' behavior in *that*-use?
- These correspond to two somewhat different types of statistical question:
  1. prediction: designing an accurate model of an outcome (machine learning)
  2. hypothesis testing: assessing a particular factor's association with an outcome (classical statistics???)

The statistical problem (2)

- In both cases, there are huge numbers of features that may potentially affect the outcome
  - e.g., each English noun may have distinctive tendencies for RC modification (*way*, *apple*)
- Problem of *model selection*: which features to put into the model?
- The answer differs for each statistical question:
  1. Prediction: a very large, overparameterized model is OK, as long as it accurately predicts outcomes
  2. Hypothesis testing: test the factor of interest in a small model with carefully developed control factors

Two-step model

P(RC | *context*) → **P(*that* | RC)**

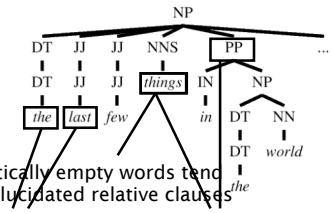
Control factors

- three outcomes (RC, *END*, other)
- regularized multinomial logistic regression (exponential model)
- large number of surface & structural features of *context* (~3.3×10^6; $n\approx10^6$)

- binary outcome
- unregularized logistic regression (bootstrapped by speaker cluster)
- phrasal predictability is a *single* covariate
- a select set of controls * constitutes another 27 parameters (n=3,452)

Feature space for prediction model

- Linguistic theory suggests many types of features that may be important:



semantically empty words tend to be elucidated relative clauses

definite articles and superlative adjectives especially together, like RCs

pertain to modifiers of the noun tend to fill this need for elucidation

Investigating control factors

- Separate studies (Jaeger 2006a,b) had investigated the role of many other factors in *that*-use:
  - Length of the relative clause and distance from "gap":
    …one of the things *that* we were just talking about __ as a matter of fact this ~~week~~ at work…
  - Disfluency (production difficulty)
    - …we certainly can, uh, force, uh, government, uh *that* we elect __…
  - Adjacent identical segments
    - …I mean no one individual *that* that's true for __…
  - Speaker gender (women seem to say *that* more than men)
- These factors & others were selected from a larger set using backward AIC optimization

Putting the two models together

- Hypothesis test: enter -log P(RC|context) as covariate with the control factors in a logistic regression
- Result: phrasal predictability is associated with *that*-omission at $p<0.0001$ (Wald statistic)
- We can also run backward model selection using AIC again on the new model
- Result: several control factors drop out of the model
  - adjacent identical segments seem not to matter
  - speaker gender effect goes away
- *Phrasal predictability helps us make sense of* that-*use*

Production study: conclusion

- Speakers seem sensitive to information density as a principle of communicative optimality
- An optional function word, like *that* acts as a "pressure valve" for speakers to regulate information flow
- Leads to a *very* unconventional view of grammar
  - conventional view: a set of categorical rules reflecting universal, innate principles
  - new view: a set of statistically-oriented tools to achieve communicative ends
- Methodology: combine different statistical modeling principles to gain insights about human language

Overall conclusion

- Probabilistic knowledge is hugely important to language users
- Language users act *rationally* on their probabilistic knowledge in:
  - processing rates in language comprehension
  - ambiguity management in language comprehension
  - information rate-sensitive choices in language production

Thank you!

http://ling.ucsd.edu/~rlevy