Processing difficulty in verb-final clauses matches syntactic expectations

Roger Levy
Stanford University

79th Annual Meeting of the LSA
Oakland, CA – Jan 7, 2005

Big picture

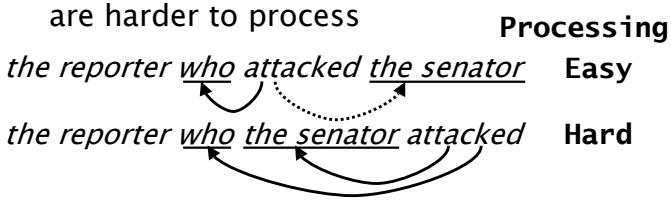
- Realistic models of human sentence processing must account for
  - Robustness to arbitrary input
  - Accurate disambiguation
  - Inference on basis of incomplete input (Tanenhaus et al 1995, Altmann and Kamide 1999)

Today

- Expectation-based models of syntactic processing
  - Works better than traditional memory-based models for verb-final clauses
  - The model of Hale 2001 fits closely to established experimental results
  - A new information-theoretical derivation of Hale's model

Memory-based processing

- On the traditional view, resource limitations, especially memory, drive processing difficulty
- One incarnation: Gibson 1998: multiple and/or distant dependencies are harder to process



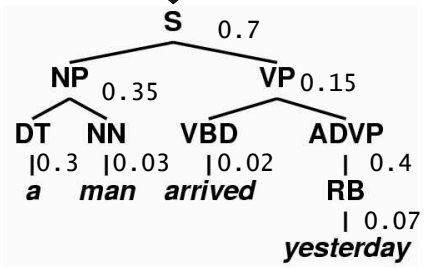
Parallel, expectation-based syntactic processing

- Several researchers (Jurafsky 1996, Narayanan & Jurafsky 1998, Crocker & Brants 2000) propose ranked-parallel syntactic processing models
- Hale 2001: a distinctive ranked-parallel model
  - The more a word is expected, the easier it is to process: ***difficulty ~ SURPRISAL(w_i)***

    SURPRISAL(w_i) ≡ -log P(w_i|CONTEXT)
  - Parallel parsing with a probabilistic context-free grammar (PCFG) determines the expectation of a word

a man arrived yesterday

0.3 S → S CC S
0.7 S → NP VP
0.35 NP → DT NN
0.15 VP → VBD ADVP
0.4 ADVP → RB
...



Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10^{-7}

Algorithms by Lafferty and Jelinek (1992), Stolcke (1995) give us $p_i(w)$ from a PCFG

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 very beautiful the mountain"

Memory-based models for final verbs

- Memory-based models (Gibson 1998) predict difficulty for longer clauses

	Prediction
Er hat die Gruppe **geführt**	easy
He led the group	
Er hat die Gruppe auf den Berg **geführt**	hard
He led the group to the mountain	
...die Gruppe auf den sehr schönen Berg **geführt**	hard
He led the group to the very beautiful mountain	

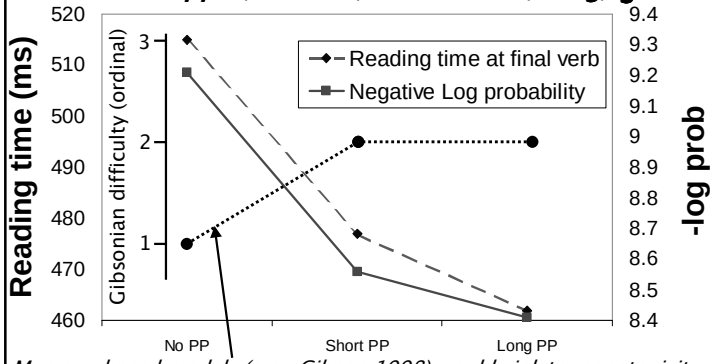
Empirical results

- But Konieczny found that final verbs were read faster in longer clauses

	Prediction	Result
Er hat die Gruppe **geführt**	easy	slow
He led the group		
Er hat die Gruppe auf den Berg **geführt**	hard	fast
He led the group to the mountain		
...die Gruppe auf den sehr schönen Berg **geführt**	hard	fastest
He led the group to the very beautiful mountain		

Predictions from Hale 2001 model

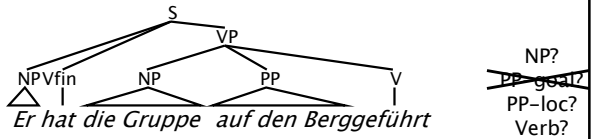
***Er hat die Gruppe (auf den (sehr schönen) Berg)* geführt**



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

Deriving Konieczny's results

- Seeing more = having more information
- More information = more accurate expectations
  - Once we've seen a PP goal we're unlikely to see another



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

  - So once we've seen a PP goal, 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)

Comprehension as disambiguation

- *sentence comprehension:* choosing the "best" [most probable] syntactic/semantic structure from among possible structures {T}
- Partial input $w_{1...i}$ induces a *preference distribution* [probability distribution] **D** over possible T
- **D** must be constantly updated (for inference!)
- Suppose *greater changes in* **D** *incur greater cost*
  - Operationalize as the *relative entropy* between distributions **BEFORE** and **AFTER** w_i

Crucial result

preference distribution after seeing w_i

preference ratio for T, before vs. after w_i (**constant!**)

Relative Entropy $\sum_T P_i(T) \log \frac{P_i(T)}{P_{i-1}(T)} = \ldots = -\log P_{i-1}(w_i)$

preference distribution before seeing w_i

surprisal at w_i

Relative entropy over *trees* comes out as surprisal over *words!*

[for details, see last slide on handout]

Implications

- Connects disambiguation with processing difficulty
- Representation-independent
  - Contrasts with most memory- and expectation-based models
- A processing model now consists solely of a conditional word probability model $p_i(w)$
- We are free to use the best available techniques for ***estimating*** $p_i(w)$
- But the techniques we use (finite-state model, PCFG, ...) are no longer a strong commitment as to the structures built by the human parser

$$\sum_T p_i(T) \log \frac{p_i(T)}{p_{i-1}(T)} \Rightarrow -\log p_{i-1}(w_i)$$

Contributions

- Established extremely good fit of the Hale 2001 model to verb-final data
  - Memory-based models do very badly!
- New information-theoretical derivation of Hale's model
- Processing difficulty in verb-final clauses *does* match syntactic expectations

Surprisal from relative entropy

$P_i(w)$ and $P_i(T)$ are just $P(w|w_{1\ldots i})$ and $P(T|w_{1\ldots i})$

$$D(P_i(T) \| P_{i-1}(T)) \equiv \sum_T P_i(T) \log \frac{P_i(T)}{P_{i-1}(T)}$$

definition of relative entropy from $P_{i-1}(T)$ to $P_i(T)$

$$= \sum_T P_i(T) \log \frac{P(T \mid w_{1\ldots i})}{P(T \mid w_{1\ldots i-1})}$$

$$= \sum_T P_i(T) \log \frac{P(T \mid w_{1\ldots i})}{P(T \mid w_{1\ldots i}) P(w_i \mid w_{1\ldots i-1})}$$

$$= -\log P(w_i \mid w_{1\ldots i-1}) \sum_T P_i(T)$$

$\sum_T P_i(T) = 1$

$$= -\log P_{i-1}(w_i)$$

a.k.a surprisal, proposed by Hale 2001, but without information-theoretic basis