Roger Levy (2008)

Expectation-Based Syntactic Comprehension

Anna Finzel  Melanie Tosik
Johannes Schneider  Sebastian Golly

May 13, 2013
Outline

Background

Surprisal Theory

Surprisal Theory in Action
  Comparison with Other Processing Theories
  Surprisal vs. Locality
  Subject Preference

Shortcomings

Conclusion
Outline

Background

Surprisal Theory

Surprisal Theory in Action
  Comparison with Other Processing Theories
  Surprisal vs. Locality
  Subject Preference

Shortcomings

Conclusion
Relative Clause Processing: Approaches

- Garden-Path Model
- Good-Enough Processing
- Unrestricted Race Model
- Constraint-Based Models
Relative Clause Processing: Approaches

Resource-limitation vs. resource-allocation
Relative Clause Processing: Approaches

- Resource-limitation
  - Late Closure
  - Minimal Attachment
  - Dependency Locality Theory
  - e.g. King and Just (1991)
Relative Clause Processing: Approaches

- Resource-allocation
  - expectation-based
  - plausibility \( \Rightarrow \) (1) competition; (2) reranking
- Sentence comprehension
  - parallel
  - incremental
  - probabilistic
Relative Clause Processing: Approaches

Levy’s proposal: *Surprisal Theory*
(cf. Hale (2001))
Outline

Background

Surprisal Theory

Surprisal Theory in Action
  Comparison with Other Processing Theories
  Surprisal vs. Locality
  Subject Preference

Shortcomings

Conclusion
Main Properties of Surprisal Theory

- **Expectation-based** theory of syntactic comprehension
- Focus on **resource-allocation**
- The parsing process is
  - parallel
  - incremental
  - probabilistic
- The difficulty of a word is proportional to its **surprisal**
Preference Distributions

• Comprehending a sentence:
  Constructing a preference ranking over all possible structures → parallel

• Preference ranking: probability distribution → probabilistic
  • consists of an allocation of resources among the structures → resource-allocation
  • is updated constantly → incremental

• Processing difficulty is proportional to the degree of update in the preference distribution → surprisal
Surprisal

- **Surprisal**: determinant of a word’s processing difficulty
  - in information theory: *negative log-probability* of the word
  - is minimized when a word *must* appear in a given context
  - approaches infinity as a word becomes less and less likely
  - can be interpreted as the *difficulty of updating* the preference distribution

- **Nothing new**
  - Term coined by *Tribus (1961)*
  - Surprisal theory: originally proposed by *Hale (2001)*
Modeling Surprisal Theory

- **Surprisal**: $- \log P(w_i|w_1...w_{i-1})$
- **Probabilistic word model**
  - statistical generative process that determines *conditional word probabilities*
  - can be used to **predict the next word** in a sequence
  - can be used to **estimate surprisal values**
- **Examples**:
  - n-Gram Models
  - Hidden Markov Models
  - Probabilistic Context-Free Grammars (PCFGs)
A Simple PCFG

\[
\begin{align*}
.5 & \quad S \rightarrow \text{NP } V_{itr} \\
.4 & \quad S \rightarrow \text{NP}_{NOM} \ V_{tr} \ \text{NP}_{ACC} \\
.1 & \quad S \rightarrow \text{NP}_{ACC} \ V_{tr} \ \text{NP}_{NOM} \\
1.0 & \quad \text{NP} \rightarrow \text{Det } N \\
1.0 & \quad V_{itr} \rightarrow \text{gackert} \\
1.0 & \quad V_{tr} \rightarrow \text{sieht} \\
.4 & \quad \text{Det} \rightarrow \text{die} \\
.4 & \quad \text{Det} \rightarrow \text{der} \\
.2 & \quad \text{Det} \rightarrow \text{den} \\
.2 & \quad N \rightarrow \text{Henne} \\
.8 & \quad N \rightarrow \text{Hahn}
\end{align*}
\]
## How it works

<table>
<thead>
<tr>
<th></th>
<th>die</th>
<th>Henne</th>
<th>sieht</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>NP $V_{itr}$</td>
<td>.5 $NP V_{itr}$</td>
<td>.5 $NP V_{itr}$</td>
</tr>
<tr>
<td>.4</td>
<td>NP$<em>{NOM}$ $V</em>{tr}$ NP$_{ACC}$</td>
<td>.4 NP$<em>{NOM}$ $V</em>{tr}$ NP$_{ACC}$</td>
<td>.8 .4 NP$<em>{NOM}$ $V</em>{tr}$ NP$_{ACC}$</td>
</tr>
<tr>
<td>.1</td>
<td>NP$<em>{ACC}$ $V</em>{tr}$ NP$_{NOM}$</td>
<td>.1 NP$<em>{ACC}$ $V</em>{tr}$ NP$_{NOM}$</td>
<td>.2 .1 NP$<em>{ACC}$ $V</em>{tr}$ NP$_{NOM}$</td>
</tr>
</tbody>
</table>

$$S = - \log P(\text{Henne}|\text{die})$$

$$= - \log 1 = 0$$

$$S = - \log P(\text{sieht}|\text{die Henne})$$

$$= - \log .5 = .3$$

<table>
<thead>
<tr>
<th></th>
<th>der</th>
<th>Hahn</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8</td>
<td>NP$<em>{NOM}$ $V</em>{tr}$ NP$_{ACC}$</td>
<td>1.0 NP$<em>{ACC}$ $V</em>{tr}$ NP$_{NOM}$</td>
</tr>
<tr>
<td>1.0</td>
<td>.2 NP$<em>{ACC}$ $V</em>{tr}$ NP$_{NOM}$</td>
<td></td>
</tr>
</tbody>
</table>

$$S = - \log P(\text{der}|\text{die Henne sieht})$$

$$= - \log .2 = .7$$

$$S = - \log P(\text{Hahn}|\text{die Henne sieht der})$$

$$= - \log 1 = 0$$
Interim Summary

- **Comprehending** a sentence: Constructing a *preference distribution* over all possible structures
- **Processing difficulty** is proportional to the *degree of update* in the preference distribution
- Difficulty incurred in processing a word can be quantified by its *surprisal value*: \(- \log P(w_i|w_1...w_{i-1})\)
- To *calculate* surprisal, we can use different kinds of *probabilistic word models* (e.g. PCFGs)
Outline

Background

Surprisal Theory

Surprisal Theory in Action
  Comparison with Other Processing Theories
  Surprisal vs. Locality
  Subject Preference

Shortcomings

Conclusion
Theories to be Compared

- Predictability
- Locality
- Competition and dynamical models
- Tuning
- Pruning and attention shift
- Prediction-based connectionist models
Theory to be Compared

- Locality
Key Idea of Locality

- Greater distance between words causes greater processing difficulty
- Preference for more local syntactic relationships directly guides disambiguation
Key Idea of Locality

- Greater distance between words causes greater processing difficulty
  → Dependency Locality Theory (DLT; Gibson, 1998)
- Preference for more local syntactic relationships directly guides disambiguation
  → Active Filler Hypothesis (AFH; Clifton & Frazier, 1989)
Key Idea of Locality

- Greater distance between words causes greater processing difficulty
  → Dependency Locality Theory (DLT; Gibson, 1998)
- Preference for more local syntactic relationships directly guides disambiguation
  → Active Filler Hypothesis (AFH; Clifton & Frazier, 1989)
Common Relative Clauses

(1)  a. The reporter who attacked the senator admitted the error.
b. The reporter who the senator attacked admitted the error. (Gibson, 1998)
Common Relative Clauses

Surprisal
Dependency Locality Theory
(Active Filler Hypothesis)

→ Similar predictions:
Object RC is more difficult than the Subject RC
Subject-Modifying Relative Clauses

(2)  
\begin{align*}
\text{a.} & \quad \text{The player [that the coach met \textbf{at 8 o’clock}] bought the house.}\ldots \\
\text{b.} & \quad \text{The player [that the coach met \textit{by the river at 8 o’clock}] bought the house.}\ldots \\
\text{c.} & \quad \text{The player [that the coach met NEAR THE GYM \textit{by the river at 8 o’clock}] bought the house.}\ldots \\
\end{align*}
(Jaeger et al., 2005)
Table 1
Surprisal and average reading times at matrix verb for (2)

<table>
<thead>
<tr>
<th>Number of PPs intervening between verbs</th>
<th>1 PP</th>
<th>2 PP</th>
<th>3 PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLT prediction</td>
<td>Easier</td>
<td>Harder</td>
<td>Hardest</td>
</tr>
<tr>
<td>Surprisal</td>
<td>13.87</td>
<td>13.54</td>
<td>13.40</td>
</tr>
<tr>
<td>Mean reading time (ms)</td>
<td>510 ± 34</td>
<td>410 ± 21</td>
<td>394 ± 16</td>
</tr>
</tbody>
</table>
(3) a. I read that the governor of the province retiring after the troubles is very rich.
b. I read that the province of the governor retiring after the troubles is very rich.
c. I read that the bodyguard of the governor retiring after the troubles is very rich.
(van Gompel et al., 2005)
(Yet Another) Interim Summary

Unlike locality, surprisal makes the right predictions for:

- Object over subject relativizations
- English subject-modifying relative clauses of varying lengths
- Local ambiguous sentences
Subject Preference

• Case syncretism in languages: “Haus” = acc/nom/(dat)
• With free word order this leads to possible ambiguities
  (4) Die Henne sieht den Bussard
  (5) Die Henne sieht der Bussard
• SVO is a “default” word order and read more quickly
• Locality explanation: movement + locality asymmetries (no frequencies)
• Other alternative: different construction-frequencies
Subject Preference

- Two experiments with wh-questions ("was" and "welches")
- No differences in construction frequencies in wh-questions
- Does the subject preference persist in this case?
- How does surprisal explain these results?
• “was”-sentences:

(6) Was erforderte *den* Einbruch in die Nationalbank? [SVO]

(7) Was erforderte *der* Einbruch in die Nationalbank? [OVS]

• Higher reading times in object-initial sentence, but at the PP, not at the NP
Explanation by Surprisal

- Surprisal in “welches”-sentences:
- all possible structural continuations that can lead to the main verb

(8) [Welches System]_{SUBJ} V.sg...
(9) [Welches System]_{OBJ} V.sg...
(10) [Welches System]_{OBJ} V.pl...
(11) *[Welches System]_{SUBJ} V.pl...

- → lower expectation for plural verb
Explanation by Surprisal

- Surprisal in “was”-questions
- Remember:
- disambiguation at post-verbal NP
- but higher RTs at PP
In German, object NPs are empirically more likely than subject NPs to be post-modified by prepositional phrases. This is shown in Table 5: it is true not only of subject versus object NPs overall, but also specifically of subject versus object NPs in the immediate post-verbal position. This means that the probability of the rule (14) is higher than the probability of (15). In the online comprehension of (12), immediately after hearing *Einbruch* the comprehender therefore has a greater expectation of seeing a PP (and hence a preposition) next in the *den* condition than in the *der* condition. Hence the surprisal at *in* is greater in the *der* condition.

![Graph showing the difference between object–initial and subject–initial reading times and surprisals of (11)](image)

**Fig. 7.** Predicted vs. actual reading time differentials for (12).

36 All subject/object differentials in Table 5 are significant by Fisher's exact test. For all corpora, postmodification also remains more frequent for object than for subject NPs when only PPs headed by the word *in* are considered, although only the figures for post-verbal NPs (not the figures for all NPs) are statistically significant.
• Explanation for higher RTs at PP:
  • \( \text{NP}_{\text{ACC}} + \text{PP} \) much more frequent than \( \text{NP}_{\text{NOM}} + \text{PP} \)
  • \( \rightarrow \) higher surprisal in OVS-condition

• Explanation for “normal” RTs at NP:
  • more frequent to put subject directly after verb in OVS than vice versa
  • this reduces surprisal between conditions
Result

- Surprisal predicts which conditions are harder to process
- In contrast to other theories, it predicts precisely WHEN the difficulty occurs
Outline

Background

Surprisal Theory

Surprisal Theory in Action
  Comparison with Other Processing Theories
  Surprisal vs. Locality
  Subject Preference

Shortcomings

Conclusion
Difficulties in Relative Clauses

- Object RCs are more difficult than subject RCs
- But WHEN does this difficulty occur?
- DLT (Locality): at the verb - here extra integration cost is paid
- Surprisal?
• RC similar to head-final clause:
• verb must occur at some point but comprehender doesn’t know when

(12) The reporter who sent the photographer to the editor hoped for a good story.

(13) The reporter who the photographer sent to the editor hoped for a good story.

• the more material in between, the easier it is for the test person (according to surprisal...)
• \[ \rightarrow \text{surprisal predicts that object RCs are read faster} \]
• plus reading times should be higher at the embedded subject in object RCs
• But this is not at all the way it is:
• increased RT at the verb in object RCs
• embedded subject is read quickly
• $\rightarrow$ surprisal fails in Relative Clauses
Difficulties with “digging-in effect”

- While multiple analyses are possible, the favored analysis becomes stronger even without evidence.
- Best example: NP/Z-ambiguities:
  
  (14) As the author wrote the book grew.
  
  (15) As the author wrote the book describing babylon grew.

- Test persons judge the second sentence ungrammatical more often.
Combining Locality and Surprisal?

- Surprisal good at predicting local effects in language processing
- "Which word comes next?"
- Locality is good in non-local environments as RCs with long distance dependencies
- For future research: a combined approach?
Outline

Background

Surprisal Theory

Surprisal Theory in Action
  Comparison with Other Processing Theories
  Surprisal vs. Locality
  Subject Preference

Shortcomings

Conclusion
Conclusion

- Expectation-based
- Probability is decisive
- Probabilistic word models cause difficulty
- Resource is allocated to input $\Rightarrow$ difficulty in understanding arises with incorrect allocation
Criticism

- No explanations of why rare structures are produced less frequently
- No predictions about competition effects (cf. e.g. Van Dyke & McElree (2006))
- Surprisal highly dependent on syntax
Any questions?
Discussion!
Questions

- English = locality, German = expectation
  - Not one-universal-theory-fits-all, but dependent on typology of the language?
  - Select the best from both approaches due to their shortcomings?
  - ACT-R?
Bibliography


