Roger Levy (2008)

Expectation-Based Syntactic Comprehension

Anna Finzel Melanie Tosik
Johannes Schneider Sebastian Golly

May 13, 2013

Outline

Outline

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

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

resource-allocation

parallel

incremental

probabilistic

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:** $-\log \boldsymbol{P(w_i|w_1...w_{i-1})}$
- **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

.5	$S \rightarrow NP\ V_{itr}$
.4	$S \rightarrow NP_{NOM}\ V_{tr}\ NP_{ACC}$
.1	$S \rightarrow NP_{ACC}\ V_{tr}\ NP_{NOM}$
1.0	$NP \rightarrow Det\ N$

1.0	$V_{itr} \rightarrow$ gackert
1.0	$V_{tr} \rightarrow$ sieht
.4	Det $\rightarrow$ die
.4	Det $\rightarrow$ der
.2	Det $\rightarrow$ den
.2	N $\rightarrow$ Henne
.8	N $\rightarrow$ Hahn

How it works

die

.5	NP V_{itr}
.4	NP_{NOM} V_{tr} NP_{ACC}
.1	NP_{ACC} V_{tr} NP_{NOM}

Henne

.5	NP V_{itr}
.4	NP_{NOM} V_{tr} NP_{ACC}
.1	NP_{ACC} V_{tr} NP_{NOM}

$$S = -\log P(\text{Henne}|\text{die}) = -\log 1 = 0$$

sieht

~~.5~~	~~NP V_{itr}~~
.8 ~~.4~~	NP_{NOM} V_{tr} NP_{ACC}
.2 ~~.1~~	NP_{ACC} V_{tr} NP_{NOM}

$$S = -\log P(\text{sieht}|\text{die Henne}) = -\log .5 = .3$$

der

~~.8~~	~~NP_{NOM} V_{tr} NP_{ACC}~~
1.0 ~~.2~~	NP_{ACC} V_{tr} NP_{NOM}

$$S = -\log P(\text{der}|\text{die Henne sieht}) = -\log .2 = \mathbf{.7}$$

Hahn

1.0	NP_{ACC} V_{tr} NP_{NOM}

$$S = -\log P(\text{Hahn}|\text{die Henne sieht der}) = -\log 1 = 0$$

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

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
  $\rightarrow$ Dependency Locality Theory (DLT; Gibson, 1998)
- Preference for more local syntactic relationships directly guides disambiguation
  $\rightarrow$ 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) a. The player [that the coach met **at 8 o'clock**] bought the house...
b. The player [that the coach met *by the river* **at 8 o'clock**] bought the house...
c. The player [that the coach met NEAR THE GYM *by the river* **at 8 o'clock**] bought the house...
(Jaeger et al., 2005)

Table 1
Surprisal and average reading times at matrix verb for (2)

	Number of PPs intervening between verbs		
	1 PP	2 PP	3 PP
DLT prediction	Easier	Harder	Hardest
Surprisal	13.87	13.54	13.40
Mean reading time (ms)	510 ± 34	410 ± 21	394 ± 16

When Ambiguity Facilitates Comprehension

(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

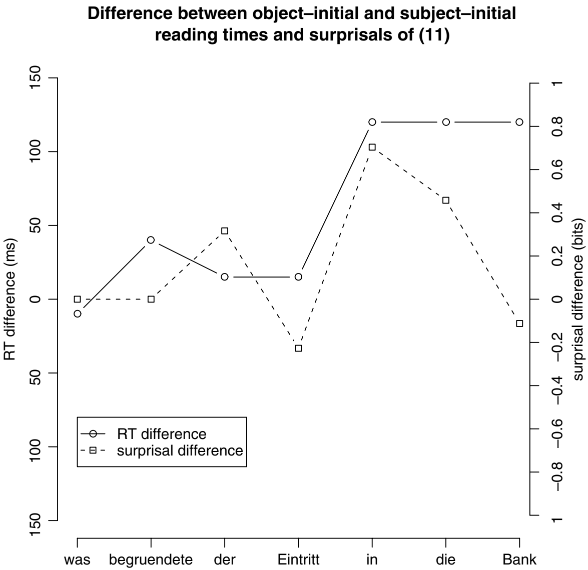

Fig. 7. Predicted vs. actual reading time differentials for (12).

- Explanation for higher RTs at PP:
- NP_{ACC} + PP much more frequent than NP_{NOM} + 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 precisly WHEN the difficulty occurs

Outline

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...)
- $\rightarrow$ 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

Conclusion

- Expectation-based
- Probability is decisive
- Probabilistic word models cause difficulty
- Resource is allocated to input ⇒ 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

- Clifton, C., & Frazier, L. (1989). Comprehending sentences with long distance dependencies. In G. Carlson & M. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 273–317). Dordrecht: Kluwer.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1–76.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In: *Proceedings of NAACL*, 2, 159–166.
- Jaeger, F., Fedorenko, E., & Gibson, E. (2005). Dissociation between production and comprehension complexity. In *Poster presentation at the 18th CUNY sentence processing conference*, University of Arizona.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580–602.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.
- Tribus, M. (1961). *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications.*
- van Dyke, J. A., & McElree, B. (2006). Retrieval Interference in Sentence Comprehension. *Journal of Memory and Language*, 55, 157-166.
- van Gompel, R. P. G., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 52, 284–307.