Levy 2015: Memory and surprisal in human sentence comprehension

Julius Nguyen

Universität Potsdam

February, 2nd 2022

Structure

1. Nature of language
    - Formal approach
    - Cognitive approach
2. Multiple centre-embedded sentences
3. Relative clauses, complement clauses
4. Dependency Locality Theory
5. Cue-based recall theory
6. Subject-, object-extracted relative clauses
7. Agreement attraction
8. Local coherence effects
9. Expectation, cloze completion task
10. Surprisal
11. Garden-path sentences
12. PCFGs
13. Theoretical justifications for Surprisal
14. Constrained syntactic contexts
15. Summary

Nature of language: Formal approach

'Language makes infinite use of finite means' (Humboldt 1836)

Grammar of a language L can be defined as a tuple $\langle S, T, N, R\rangle$ consisting of

- a finite set of terminal symbols T
- a finite set of non-terminal symbols N
- a finite set of rules R
- a start symbol S

A language L then can be defined as a tuple $\langle G, W\rangle$ where W are all the words w_1,...,w_n with $n \in \mathbb{N}$ created by the *recursive* application of rules $\in$ R.

Nature of language: Formal approach

A minimal example: Grammar G_1 of L_1
Let G_1 be a grammar of language L_1. Then $G_1 = \langle T, N, R, S \rangle$ with:
$T =$ that, the, cat, dog, worried, ate, killed (i.e. our lexical words)[1]
$N = CP, DP, V$ (i.e. our lexical categories)
$R = \{$

S → that DP S V | that DP V
DP → the cat | the dog | the rat
V → worried | ate | killed

$\}$ (i.e. our rules)
S (i.e. our start symbol)

[1] Usually Ts are single string items. For the sake of readibility the grammar has been simplified.

Nature of language: Formal approach

Given our rules R, how do we derive words of L_1?

S → that DP S V | that DP V
DP → the cat | the dog | the rat
V → worried | ate | killed

starting point: S (by definition of G_1)

S ⇒ that DP V ⇒
that the cat V ⇒
that the cat worried

Our grammar G_1 can derive relative clauses by using the **rewritten rules** of R_1.

Nature of language: Cognitive approach

- language has a grammar
- grammar allows us to create an infinite number of sentences → recursion
- but: language is bound to the limited ressources of the brain, such as the memory, to create and process sentences created by grammar

What is the relationship between language and memory?
What are the reasons that some sentences are easier to process than others?

Multiple centre-embedded sentences

(1) a. This is the malt that the rat that the cat that the dog worried killed ate.

b. This is the malt that was eaten by the rat that was killed by the cat that was worried by the dog.

Sentences of (1) repeated with brackets:

(2) a. 1a This is the malt [that the rat [that the cat [that the dog worried] killed] ate]

b. 1b This is the malt [that was eaten by the rat] [that was killed by the cat] [that was worried by the dog].

Multiple centre-embedded sentences

- centre-recursive sentences are incredibly difficult to process despite their syntactical structure being simple

Using rules from G_1:

S → that DP S V | that DP V
DP → the cat | the dog | the rat
V → worried | ate | killed

Top-down derivation:

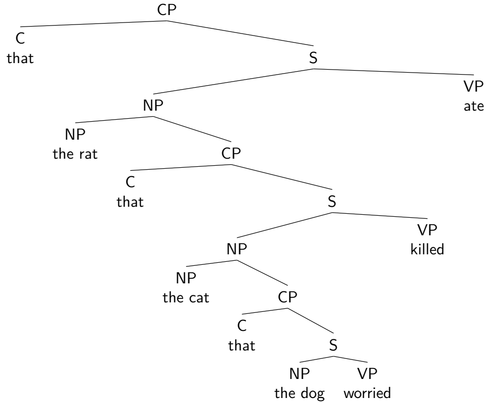
S ⇒ that DP S V ⇒ that the rat S ate ⇒
that the rat that DP S V ate ⇒ that the rat that the cat S killed ate ⇒
that the rat that the cat that DP V killed ate ⇒
that the rat that the cat that the dog worried killed ate

Multiple centre-embedded sentences

Explanations for processing difficulty:

- nested structures are difficult to process, since dependencies needs to be stored before they can be completed
- Yngve (1960): number of stacks that can be kept in the memory is $n = 3$
- level of embedding for (1a) $n = 4$

(3)



Multiple centre-embedded sentences

Issues with Yngve's model (1960)

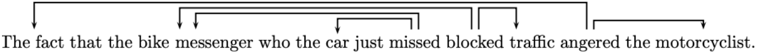
- Yngve (1960) predicts that breakdown occurs at *the* in *that the dog worried*, when deepest level n=4 is reached
- Gibson & Thomas (1999) found out that processing breakdown happens before deepest level
- cannot explain how frequently occuring left recursive structures in OV-languages (e.g. Japanese, Hindi) are parsed

Relative clauses, complement clauses

(4) a. The fact $[_{CC}$that the bike messenger $[_{ORC}$ who the car just missed] blocked traffic] angered the motorcyclist].

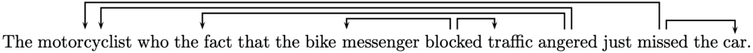
b. The motorcyclist $[_{ORC}$ who the fact $[_{CC}$ that the bike messenger blocked traffic] angered] just missed the car.

(5) a.



The fact that the bike messenger who the car just missed blocked traffic angered the motorcyclist.

b.



The motorcyclist who the fact that the bike messenger blocked traffic angered just missed the car.

- complement clauses could stand by their own
- object relative clauses, however, require reconstruction of head-trace relationship

Dependency Locality Theory (DLT)

Gibson (1998, 2000) and Grodner & Gibson (2005)

- DLT postulates two types of costs in sentence processing:
  1. **integration costs:** costs for establishing syntactical relationships
  2. **storage costs:** costs for maintaining unresolved dependencies
- distance between the head noun and its trace is greater in (4b) than in (4a)
- the greater the distance, the more intervening NPs
- (in 4b), integration at *angered*→*motorcyclist*, *fact* with *messenger*, *bike*, *traffic* intervening
- (in 4b), storage cost at *angered* $= 1$
- DLT predicts that (4b) is more difficult to process than (4a), since dependencies in the former are longer and therefore having more potentially intervening heads

Cue-based recall theory

Lewis & Vasihth (2005); Lewis et al. (2006)

- assumes that parser has no access to linear word order → cannot tell which word linearly precedes another word
- instead, tree structure is stored in memory
- retrieval is a competition-led operation
- parsing is cue-based i.e. the more favourable cues there are, the easier the retrieval (=facilitation)
- cue can also cause interference and hence lead to retrieval errors

Subject-, object-extracted relative clauses

(6) a. The reporter_i who t_i attacked the senator left the room. SRC

b. The reporter_j who the senator attacked t_j left the room. ORC

- DLT predicts that ORCs are more difficult than SRCs
- cue-based recall theory makes the same predictions for ORCs and SRCs
- divergent explanations:
  - DLT: integration cost at *attacked* higher in (6b) compared to (6a)
  - CBT: *senator* is already in memory in (6a) unlike in (6b) but semantic similarity of *reporter* and *senator* leads to high retrieval interference

Agreement attraction

Wagers et al. (2009)

(7) a. * The key to the **cabinets** were rusty from many years of disuse.
+ attractor

b. * The key to the **cabinet** were rusty from many years of disuse.
- attractor

- mismatch in subject-verb agreement
- number value of NP *cabinet* percolates to upper NP *key* leading to a favourable judgement of (7a)

Agreement attraction

Counter-evidence for percolation:

(8) a. * The **musician** who the reviewer praise so highly will probably win a Grammy.

b. * The **musicians** who the reviewer praise so highly will probably win a Grammy.

- misleading cue is outside of the relative clause → cannot percolate number value to the relative clause

Local coherence effects

- analysis which would be only taken if the sentence were to be uttered in isolation (=global analysis)

Tabor et al. (2004):

(9) a. The coach smiled at the player **tossed** the frisbee.

b. The coach smiled at the player **thrown** the frisbee.

- higher reading time in (9a) than in (9b)
- *tossed* can be a participle or a finite verb, *thrown can be only a participle*
- *the player* might be picked as the subject of *tossed* due to its retrieval cues (3SG, AGENT)

Expectation, cloze completion task

- based on Cloze completion task, Ehrlich & Rayner (1981) hypothesized that fulfilled expectations lead to higher processing rates of new input

(10) The boat passed easily under the ...

(11) Rita slowly walked down the shaky ...

- (12) shows how context-dependent sentence analyses are
- interpretation depends on how many apples are present in the context

Tanenhaus et al. (1995)

(12) Put the apple on the towel ...

Surprisal

Hale (2001)

- Hale (2001) took a quantitative approach to expectation of words
- surprisal S is defined as the log of the inverse of the probability of an event given a context C
- event is defined as a word under the condition that it is following a string of words $w_i, ..., w_{i-1}$

Hence the formula:

Surprisal S

$S = log \frac{1}{P(w_i|w_i,...,w_{i-1},C)}$

Surprisal allows us to calculate the cognitive load at each word in a sentence.

Garden-path sentences

Staub (2007)

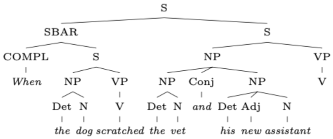
(13) a. When the dog scratched the vet and his new assistant removed the muzzle.

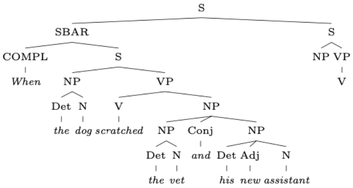
b. When the dog scratched, the vet and his new assistant removed the muzzle.

- two parses are available one of which interprets *the vet and his new assistant* as the object-NP of the subordinate clause verb *scratched* and the other as the subject-NP of the main clause

Garden-path sentences

Possible parses of (13a):


S
SBAR S
COMPL S NP VP
When NP VP NP Conj NP V
Det N V Det N and Det Adj N
the dog scratched the vet his new assistant



- difficulty of garden path sentences can be explained by probabilistic context free grammars (PCFGs, Manning & Schütze 1999)

PCFGs

The probability of a sentence to occur can be expressed by

(Simplified) Bayes' Rule

$P(T|w_{1..i}) = \frac{P(T)}{P(w_{1..i})}$ where

w are words processed as far as w$_i$, the current word, and
T are parses

The probability of a word can be calculated using

The Law of Total Probability

$$^{a}\ P(w_i|w_{1..i-1}) = \frac{P(w_i \cap w_{1..i-1})}{P(w_{1..i-1})} = \frac{\sum_T P(w_i \cap T \cap w_{1..i-1})}{P(w_1..i-1)} = \frac{\sum_T P(w_i|T \cap w_{1..i-1}) \times P(T|w_{1..i-1}) \times P(w_{1..i-1})}{P(w_{1..i-1})} = \sum_T P(w_i|T \cap w_1..i-1) \times P(T|w_{1..i-1})$$

[a] By the same Law of Total Probability $P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n)P(B_n)$

PCFGs

Rules for our examples (13a) and (13b):

Rule		Prob.	Rule		Prob.	Rule		Prob.
S	→ SBAR S	0.3	Conj	→ and	1	Adj	→ new	1
S	→ NP VP	0.7	Det	→ the	0.8	VP	→ V NP	0.5
SBAR	→ COMPL S	0.3	Det	→ its	0.1	VP	→ V	0.5
SBAR	→ COMPL S COMMA	0.7	Det	→ his	0.1	V	→ scratched	0.25
COMPL	→ When	1	N	→ dog	0.2	V	→ removed	0.25
NP	→ Det N	0.6	N	→ vet	0.2	V	→ arrived	0.5
NP	→ Det Adj N	0.2	N	→ assistant	0.2	COMMA	→ ,	1
NP	→ NP Conj NP	0.2	N	→ muzzle	0.2			
			N	→ owner	0.2			

Hence or otherwise,

$P(w_{11} = removed|w_{1..10}) = \sum_T P(w_{11=removed|w_{1..10}}, T)P(T|w_1..10)$

So $P(w_{11} = removed|w_{1..10}) = 0.2 \times 0.174 + 0 \times 0.826 = 0.0348$

where:

probability for DO-interpretation $P(T_{DO}|w_{1..10}) = 0.826$
probability for matrix clause subject interpretation $P(T_{MCLS}|w_{1..10}) = 0.174$

For (13b) grammar predicts that only matrix clause subject interpretation is available ($P = 1$).

PCFGs

Using the formula for Surprise:

$S = log\frac{1}{P(w_i|w_i..w_{i-1},T)}$

For the DO-interpretation we get $S_{removed} = log_2 0.0348 = 4.85 bits$.

For the matrix clause subject interpretation we get $S_{removed} = log_2 0.2 = 2.32 bits$

Thus, Surprisal Theory is able to predict the relative difficulty of sentence parses.

See Levy (2008) for detailed proof.

Theoretical justifications for Surprisal

Relative entropy and surprisal

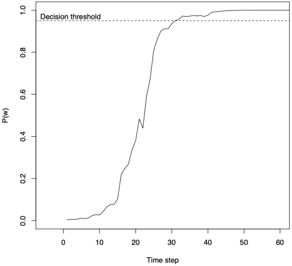
- surprisal can be represented as the costs of allocating limited memory ressources when being confronted with multiple parses
- Cover & Thomas (1991) showed that surprisal corresponds to relative entropy (i.e. the size of shift induced by a word *w*)

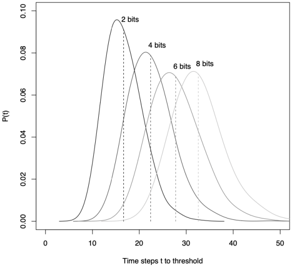
Theoretical justifications for Surprisal

Optimal perceptual discrimination

- relies on the assumption of humans as rational agents, that is individuals strive to maximise effectiveness of their actions
- before we encounter a word w, we yield a set of expectations what the word w may be
- with each input of parts of a word w, we enlarge the noisy perceptual sample
- more information available helps towards better recognition

Theoretical justifications for Surprisal





- mathematically expressed as a directed random walk in space of possible words
- average step size goes towards the correct word
- starting position of that walk corresponds to surprisal
- hence, expected time to decision threshold is linear in log-probability (i.e. longer decision time correlates with higher surprisal value)
- Smith & Levy (2008) hypothesize that rational agent can choose how many ressources they allocate to a certain task
- more ressources is reflected by a shorter processing time

Constrained syntactic contexts

- CSC refer to contexts where we know that some grammatical event X will occur but time at which that happens is unknown and we don't know anything about the exact form of X
- theories make different claims for CSCs:
  - **memory-based theories**: processing X are more difficult when there are more dependents
  - **expectation-based theories**: X is easier to process, since more information available that helps to parse X

Constrained syntactic contexts

Evidence from OV-languages (Konieczny & Döring 2003)

(14) … dass [der Freund] [**dem** Kunden] [das Auto] [aus Freude] verkaufte …

(15) … dass [der Freund] [**des** Kunden] [das Auto] [aus Plastik] verkaufte …

- (14) is an instance of CSC, (15) is not
- dative in (14) is dependent of a verb *X*
- regresion path durations in genetive-modified NP sentence in (15) were considerably longer → evidence in favour of expectation-based theories

Constrained syntactic contexts

(16) *Russian*

a. ... ofitsant, kotoryj **zabyl prinesti** [$_{DO}$ bljudo iz teljatiny]
waiter, who.NOM forget.PAST bring.INF dish.ACC of veal.GEN
[$_{IO}$ posetitelju v chernom kostjume]...
customer.DAT in black suit

'...the waiter, who forgot to bring the veal dish to the customer in the black suit '

b. ...ofitsant, kotoryj [bljudo iz teljatiny] **zabyl prinesti** [posetitelju v chernom] kostjume...

c. ... ofitsant, kotoryj [bljudo iz teljatiny] [posetitelju v chernom kostjume] **zabyl prinesti**

- higher reading time found when verbal complex is preceded by DO and IO than when it precedes the objects → evidence in favour of memory-based theory

Summary

Main sentence processing theories:

1. **memory-based theories:** limitations of memory storage explain processing difficulty.
2. **expectation-based theories:** cues or statistical probabilities explain processing difficulty.

No theory has been shown to entirely make correct predictions. There is too many cross-linguistic variation .

Expectation-based theories have the advantage that they can be mathematically expressed. That allows us to make numerical (and therefore very precise) predictions.

Summary

Range of phenomena covered were:

- Multiple centre-embedded clauses
- Relative clauses in complement clauses and *vice versa*
- subject and object-extracted relative clauses
- agreement attraction
- Local coherence effects
- Garden-path sentences

Discussion

Questions

- Which model is globally more accurate? → broader evaluation is necessary!
- What features of language do induce which type of pattern? How can we explain alternating patterns?
- How rational is the human being? How can agent control in sentence processing be explained?

References

Cover, T. & Thomas, J. (1991). Elements of Information Theory. New York: Wiley.

Ehrlich, S. F. & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. Journal of Verbal Learning and Verbal Behavior 20, 641–655.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition 68, 1–76.

Gibson (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, and W. O'Neil (Eds.), Image, Language, Brain, pp. 95–126. Cambridge, MA: MIT Press.

Grodner, D. & Gibson, E. (2005). Reading relative clauses in English. Language & Cognitive Processes 16 (2), 313–353.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, Pennsylvania, pp. 159–166.

Konieczny & Döring (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In Proceedings of the ICCS/ASCS Joint International Conference on Cognitive Science, Sydney, Australia.

Levy, R. (2015): Memory and surprisal in human sentence comprehension

Levy, R. (2008) Expectation-based syntactic comprehension. In: Cognition, vol. 106, pp. 1126–1177

References

Lewis, R. L. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science 29, 1–45.

Lewis, R. L. et al. (2006). Computational principles of working memory in sentence comprehension. Trends in Cognitive Science 10 (10), 447–454.

Manning, C. D. & Schütze, H (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Smith, N. J. & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In Proceedings of the 30th Annual Meeting of the Cognitive Science Society, Washington, DC.

Staub, A. (2007). The parser doesn't ignore intransitivity, after all. Journal of Experimental Psychology: Learning, Memory, & Cognition 33 (3), 550–569.

Tanenhaus, M. K. et al. (1995). Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632– 1634.

Wagers, M. W. et al. (2009). Agreement attraction in comprehension: Representations and processes. Journal of Memory and Language 61, 206–237.

Yngve, V. (1960): A model and an hypothesis for language structure. Proceedings of the American Philosophical Society 104, 444–466.