More and less complex

Linguistic complexity and cognitive workload: measurement and management, day 2, ESSLLI 2014

Asad Sayeed

Uni-Saarland

I told you that I try to start with a quote.

Here the Red Queen began again. 'Can you answer useful questions?' she said. 'How is bread made?'

'I know that!' Alice cried eagerly. 'You take some flour – '

'Where do you pick the flower?' the White Queen asked. 'In a garden, or in the hedges?'

'Well, it isn't picked at all,' Alice explained: 'it's ground – '
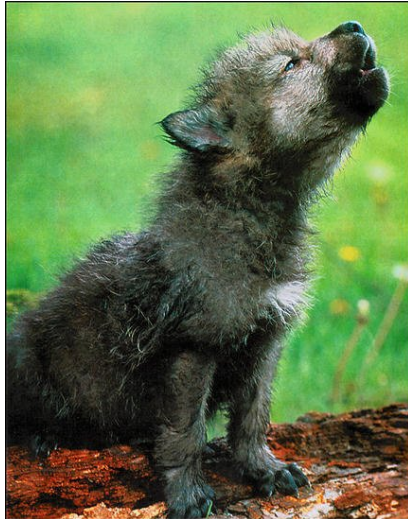
'How many acres of ground?' said the White Queen. 'You mustn't leave out so many things.'

'Fan her head!' the Red Queen anxiously interrupted. 'She'll be feverish after so much thinking.'
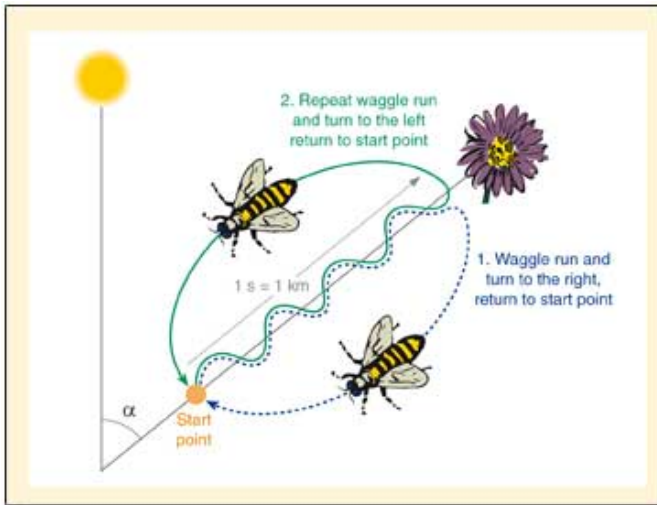
— Lewis Carroll, *Through the Looking Glass*

What makes an utterance complex?

Probably more *content* in a human sentence than a wolf howl.

Or maybe a bee waggle dance?

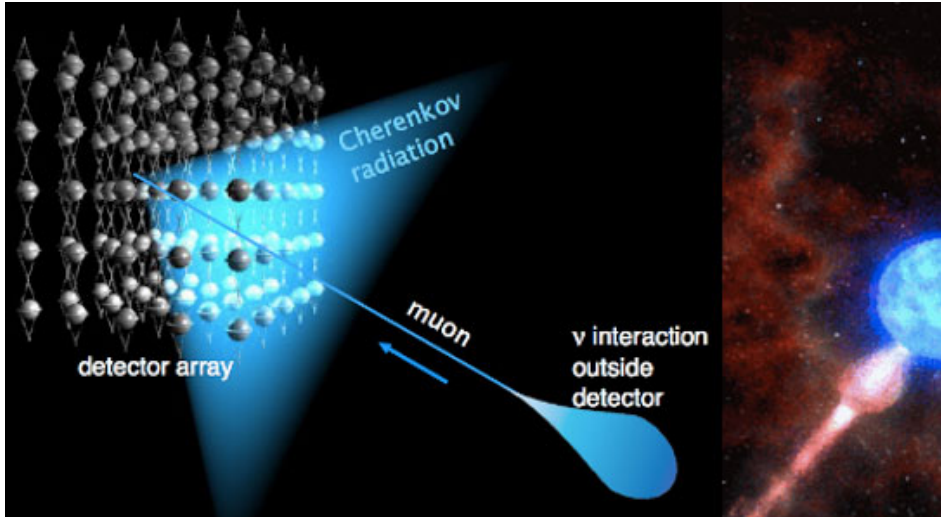# There's a clear difference between a sentence of Shakespeare. . .

. . . and a one-word interjection.

But what is that difference?

Let's look at it another way: a more physical way.

What we're calling language . . .

...is (partly?) about the serialization and deserialization of strings

Then it's about computation.

But computation takes ENERGY.

More complex computations take more energy than others.

So using language is WORK.

The thermodynamics of language?

$$\mathbf{F} = \mathbf{S}\lambda \text{ ???}$$

Not quite

- We don't really know what the "quanta" of language use actually are.
- But we know that the "energy" available for language is limited.
- Therefore,
    - Need to consider "candidates" for energy use measurement.
    - Assumption: approaching energy throughput limit $\Rightarrow$ "diminishing returns in performance.
    - Sort of like (heh) a car engine.

So what are good candidates?

Depends on your theory of linguistic behaviour.

- "Number" of "steps" it takes to "build" "structure".
- Difficulty of fitting new structure into old structure (incremental).
- Difficulty of recognizing information content.

Can we think of any others?

Back in the day...

A working hypothesis: the Derivational Theory of Complexity (DTC).

- Arose from early Transformational Generative Grammar.
- Basic idea: "effort" related to number of transformations.
  - That's our energy unit.
- Some early promising experiments.
- But you know where I'm heading with this...

IT DIDN'T WORK!
(Or at least, it was too early.)

(We're going to embark for a bit on a history of a certain kind of psycholinguistics. Apologies if it's a repeat performance from the actual psycholing class here.)
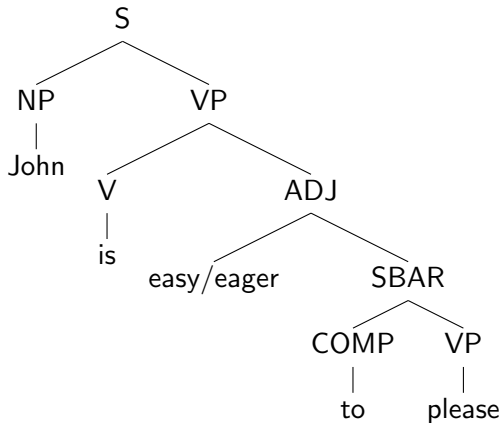
Sentences come from somewhere.

The classic distinction between "deep" structure and "surface" structure. Fodor and Garrett (1967):

1. John is easy to please.
2. John is eager to please.

These (are deemed to) have the same surface structure, but different deep structure.

The surface structure might look a bit like this.
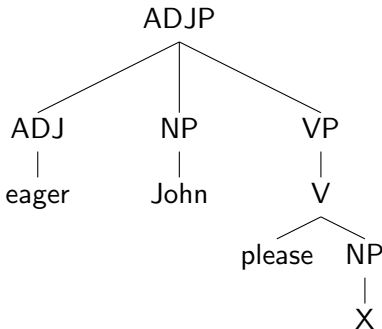
But they're fundamentally different!

"Easy" and "eager" have entirely different relationships to what they modify.

1. It is easy to please John.
2. *It is eager to please John. (where "it" is non-referring)

So even if they can appear in the same syntactic context, they "come" from somewhere else $\Rightarrow$ Deep Structure.

But if they come from somewhere else. . .

. . . how did they get to where they are? Via transformations.

ADJP
- ADJ — easy
- VP
  - V — please
  - NP — John

ADJP
- ADJ — eager
- NP — John
- VP
  - V — please
    - NP
      - X
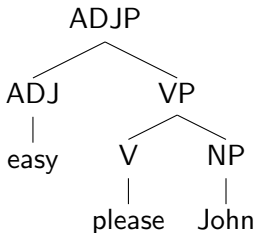
(Just for illustration, this isn't "officially" how they were supposed to look.)

What does this have to do with workload?

Fodor and Garrett construct an experiment based on relative pronouns. Hypothesis over incremental processing:

- "The man the dog bit died" $\Leftarrow$ transformations make this difficult.
- "The man whom the dog bit died" $\Leftarrow$ less difficult, because the pronoun gives the processor a clue that a transformation was performed to achieve this order.

They did quite a lot of experiments.

But I'll just summarize the highlights.

- Used sentences like "The pen (which) the author (whom) the editor liked used was new." $\Rightarrow$ ie, conditions with or without relative pronouns.
- Recordings read to 20 student subjects.
- Subjects had to explain content of sentence as soon as the recording was finished.
- Measured: response delay, recovery of subject-object relationships.

And they found that derivational distance matters.

Some major results:

- Consistently fewer subject-object relations recovered by subjects with longer delays $\Leftarrow$ without relative pronouns.
- Controlled for prosody $\Rightarrow$ helps, but not enough to overcome rel. pronoun absence.
  - (Prosody might have been affected by pronoun presence, potential confound.)

But *how* does derivation distance matter?

- Sentences like "The pen (which) the author (whom) the editor liked used was new" stretch the processor to the limit.
- If relative pronouns have a facilitating effect, then we confirm:
  - Clause boundaries are grammatically significant.
  - Parser must at some point figure out how to put things "back" where they "came" from.
  - Parser is in fact on the lookout for evidence that something must be de-transformed.
- But alas . . .

. . . not all transformations are equal.

Fodor and Garrett tested adjectives.

- "The first shot (that) the tired soldier (whom) the mosquito bit fired missed."
- According to DTC-type theories of the time, adjectives require an extra transformation.
- But they found no evidence of extra effort, quite the opposite for auditory presentation!
- (Visual presentation found no difference with or without adjectives, so the *improvement* under auditory form might just be a fluke.)

What do they conclude from this?

Derivations matter, but the number of steps is not quite the right measure.

- Facilitating effect *does* show effort to recover underlying structure.
  - But making it *easier* suggests that there isn't "re-derivation" going on.
- Lack of effect of adjectives puts the nail in the coffin.

This was one of a number of experiments that (perhaps) put the final kibosh on the DTC.

We want a looser connection to grammar.

What went wrong with the DTC?

- Too tied to the details of a theory: transformational grammar was much newer back then.
- "Derivations" were still just structural descriptions of the path between
    - An assumed "syntactic-semantic" structural representation.
    - The linear string.
    - But nothing "extrinsic" to justify that connection.
- Begging the question: want to *find* the "quantum" of grammar ⇐ assuming a kind of derivational step approximates that quantum.

So how do we find a good theory of linguistic "effort"?

The "energy units" of language effort ought to be at least one of:

- Connected to some kind of observable performative consideration, broadly construed.
    - Examples: ease of pronunciation, information transmission, etc.
- Connected to some kind of computational efficiency or mathematical constraint.
- Connected to some kind of biological limit: neural, etc.

That doesn't mean that derivations can't matter, particularly for learnability.

- But even generativists now tend to argue comp. efficiency or biological limit.

Let's go back to relative pronouns.

They seem to be a good test bed. (Why? Maybe fitting in complex information?)

- Lots of experimentation on subject relative clauses (SRC) vs. object relative clauses (ORC).
  1. The lawyer that irritated the banker filed a hefty lawsuit. (SRC)
  2. The lawyer that the banker irritated ϵ filed a hefty lawsuit. (ORC)
- The question: is one of them "harder" than the other?

(Psst, the answer is YES! Massive evidence from lots of directions!)

Separation seems to matter.

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

Object relative clauses:

- There is a bigger separation between verb and object.
- What is doing the separating? Another noun!

A clue!

So maybe the problem is the intervener ("the photographer"). Why would it have such a powerful effect?

- A memory effect: problem of holding on to "the reporter" when it could be overwritten.
- Problem of overwriting/integrating structure when you've already started building something else.

Now we have expenses we can start to quantify.

Back to energy units.

Gibson (2000 and much before), Dependency Locality Theory (DLT)

- Account for two costs, Integration Cost (IC) and Storage Cost.
  - We'll just focus on IC here.
- IC seems to correlate well experimentally with reading time data.
- Gibson provides an algorithm to calculate it based on "discourse referents".

The accounting stuff.

It's a two-step definition, literally defined in terms of energy units (EUs).

- Every *new* discourse referent costs 1 EU.
- Every time a new discourse referent must be connected to a previous head (it's not called "D(ependency)LT" for nothing), count all the intervening discourse referents as 1 EU each.

What counts as a "new discourse referent"? In a nutshell:

- Nominal referring expressions.
- Verbs that refer to event occurrences.

Table 5.1
Word-by-word predictions of the DLT for the object-extracted RC structure in (13)

Cost type	Input word													
	The	reporter	who	the	photographer	sent	to	the	editor	hoped	for	a	good	story
New discourse referent	0	1	0	0	1	1	0	0	1	1	0	0	0	1
Structural integration	0	0	0	0	0	2	0	0	0	3	0	0	0	0
Total	0	1	0	0	1	3	0	0	1	4	0	0	0	1

Table 5.2
Word-by-word predictions of the DLT for the subject-extracted RC structure in (12)

Cost type	Input word													
	The	reporter	who	sent	the	photographer	to	the	editor	hoped	for	a	good	story
New discourse referent	0	1	0	1	0	1	0	0	1	1	0	0	0	1
Structural integration	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Total	0	1	0	1	0	1	0	0	1	4	0	0	0	1

How closely does it match reading times?

Reading times measured by hitting spacebar between words.
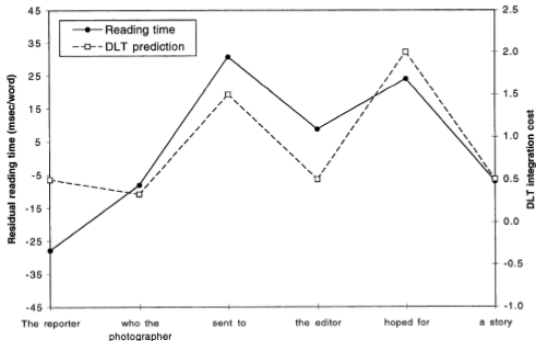For object extraction:

Figure 5.3
A comparison between residual reading times and locality-based integration costs in an object-extracted RC structure.

How closely does it match reading times?

Reading times measured by hitting spacebar between words.
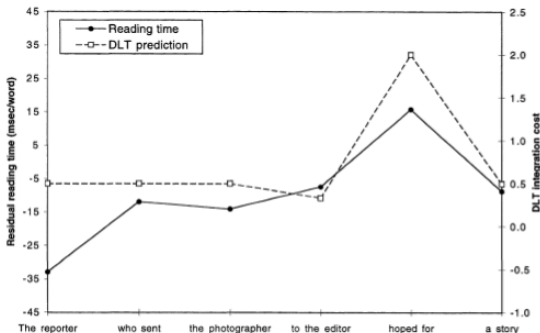For subject extraction:

Figure 5.4
A comparison between residual reading times and locality-based integration costs in a subject-extracted RC structure.

Hurrah!

Reading times (after accounting for other individual differences) seem to track IC for relative clause attachment!

- "Residual" reading time – calculated after normalising for length and doing a linear regression across all sentences.
- People speak at different rates and have other quirks, but we can "fix" these statistically.
- Combining SRC and ORC data yields a p-value of 0.005, significant.
- Evidence that this works across other languages (e.g. Japanese).

But is this all there is?

(You can probably tell I like digressions.)

Just some caveats before I go too much farther.

- This is a highly-biased, cherry-picked path through a sample of the literature.
- But I forgive myself: we've gotta start somewhere.
- It's still not 100% clear that there are *no* phenomena that can be accounted for by DTC-like constraints.
- There are psycholinguists who still work on models of structural constraint, but particularly for acquisition.
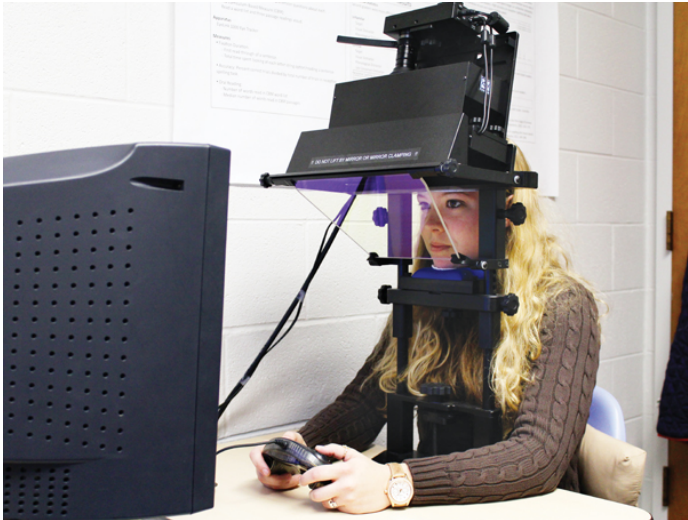
But that said. . .

Can we settle on memory/integration accounts?

No, not quite! We've just looked at syntax vs. discourse objects.

- Does it matter **what** is being referred to?
- Does it matter **in what order** it's being referred to?

What about semantic content (and just about everything else language has in it)?

(Just a moment: do y'all know what eye-tracking is?)

Naturally, someone looked at other factors.

Looking at more semantic constraints. Traxler et al. (2002):

- Eye-tracking study.
  - Subjects shown sentences on screen.
  - Position of gaze tracked across sentence, measured for: fixations, "regressions", total time.
  - Subjects quizzed for comprehension after each sentence.

  (Fairly standard.)

Testing some conditions.

Traxler et al. did three experiments:

1. SRC vs. ORC where the subjects and objects are "confusable." (e.g. "lawyer" vs. "banker" can appear in similar contexts.)
2. SRC vs. ORC where subject-verb plausibility differs across nouns, but is strictly maintained.
   - "The policeman that arrested the thief was known to carry a knife."
   - "The thief that the policeman arrested was known to carry a knife."
   - But "thief" never appears as subject of "arrested".
3. Manipulation of animacy.
   - "The director that watched the movie received a prize. . . "

You can see the goal: to figure out how much of the difficulty is semantic.

And what did they find?

A lot of stuff. But in very high-level summary:

- In "confusable" experiment, SRC vs. ORC work out as expected $\Rightarrow$ ORC more difficult (e.g. more regressions, slower).
- In "plausibility" experiment, also works out as expected, but recovery from ORC faster.
- In "animacy" experiment, ORC greatly improved by inanimate subjects.

What can we conclude from this?

The parser makes distinctions in multiple dimensions.

It could easily get out of hand!

So let's go back to what we really want out of this.

- We want to figure out/whether how complexity co-varies with task performance.
- Thus: we would ideally like a kind of "score" that takes multiple factors into account.

Fortunately, someone DID think of a way to get a single score.

It uses statistics, surprise surprise.

Actually, it's called "surprisal."

- What is surprisal? Represents (un)predictability.

Surprisal (Hale 2001)

$$- \log P(w_i | w_{1 \ldots i-1})$$

- w words, can generalize to other linguistic events.
- Highly productive concept when applied to psycholinguistic research.

Can we generalize to syntactic structures?

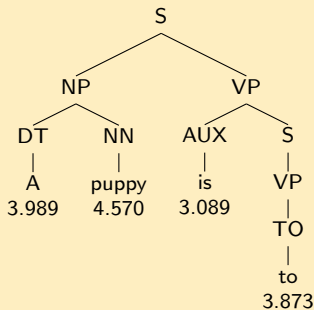Unattended Children will be given an espresso and a free puppy.

YES! Roark (2001), Roark et al. (2009) do exactly this using probabilistic context free grammars (PCFGs).

- Implemented a full-incremental parser.
- At each word, assign a surprisal score, reflecting the change in predictability brought about by the integration of that word.
- Based on *prefix probability*.

So what does it look like?

Sentence: *A puppy is to ‖ a dog what a kitten is to a cat.*

```
                           S
              ┌────────────┴────────────┐
             NP                         VP
         ┌────┴────┐              ┌──────┴──────┐
        DT        NN            AUX             S
         │         │             │              │
         A       puppy          is             VP
       3.989     4.570         3.089            │
                                               TO
                                                │
                                               to
                                             3.873
```
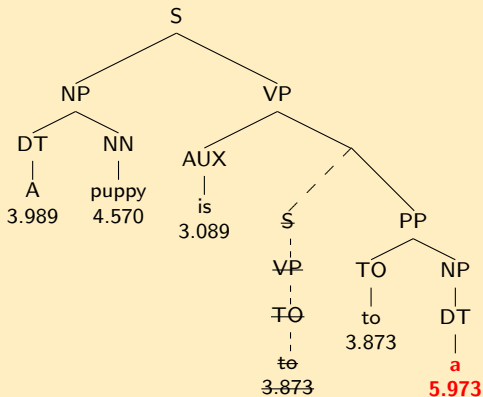
So what does it look like?

Top-ranked partial parse.

Sentence: *A puppy is to a ‖ dog what a kitten is to a cat.*

We mentioned prefix probability.

This is where those numbers come from.
How to calculate syntactic surprisal:

- S_{w_i} is syntactic surprisal at word w_i.
- T: rooted tree spanning words $w_1..w_i$.
- Then:

Prefix probability at w_i

$$\sum_T P(T, w_1..w_i)$$

- And so:

Syntactic surprisal at w_i

$$S_{w_i} = \log \sum_T P(T, w_1..w_{i-1}) - \log \sum_T P(T, w_1..w_i)$$

And the probability of a tree. . .

. . . is derived from the learned probabilities of the PCFG rules used to construct it.

Attend the Levy and Bicknell psycholinguistics course in the late afternoon to understand this better. You can tell them we sent you.

This leads to a very popular line of theorizing these days.

Uniform information density (UID) hypothesis

Speakers try to distribute info uniformly across utterances (Frank & Jaeger, 2008).

But it's backed up by a rapidly expanding body of evidence.

Just a sample. As surprisal increases:

- n-gram surprisal affects lexical/syntactic choices:
  - UID account can explain use of reduced forms in English.
    Levy and Jaeger (2007), Frank and Jaeger (2008)
  - (This is the *that*-dropping thing we do in English that allows things like "Bob said Mary believed Bill hated Bob.")
- n-gram surprisal affects syllable and word duration:
  - *Jurafsky et al. (2001), Aylett and Turk (2006)*
- Demberg et al. (2012; the "et al" includes me, heh) syntactic surprisal vs. word duration.
  - Effect can be measured in noisy conversational contexts.

And nowadays many more.

But can we do this for relative clause attachment?

Yes! Just for fun, let's talk about Chinese (just so that we're not all about English). Chen et al. (2012):

a. Subject-modifying SR (S-SR)

$$[\; e_i \; yaoqing \; fuhao \; de \;] \; guanyuan_i \; da\text{-}le \; jizhe$$
$$[\; e_i \; \text{invite} \; \text{tycoon} \; \text{DE} \;] \; \text{official} \; \text{hit} \; \text{reporter}$$

'The official who invited the tycoon hit the reporter.'

b. Subject-modifying OR (S-OR)

$$[\; fuhao \; yaoqing \; e_i \; de \;] \; guanyuan_i \; da\text{-}le \; jizhe$$
$$[\; \text{tycoon} \; \text{invite} \; e_i \; \text{DE} \;] \; \text{official} \; \text{hit} \; \text{reporter}$$

'The official who the tycoon invited hit the reporter.'

A bit about Chinese RCs.

- (*e* is just a trace/empty string.)
- Note that the complementizer appears on the *right*, *precedes* modified subject.
- Has been claimed (fits with DLT, some experimental evidence) that ORC in Chinese has the advantage! (Hsiao and Gibson, 2003)
- But SRCs are more common in all languages, so all experience-based accounts *would* be ruled out *if* this were true.

Surprisal is an experience-based account.

So is it doomed? It turns out that (Lin and Bever, 2006) when the RCs modify the object of the main clause, prediction holds:

    a. Object-modifying SR (O-SR)

        jizhe   *da-le* [e_i *yaoqing* *fuhao* *de*] *guanyuan$_i$*
         reporter   hit   [e_i invite   tycoon   DE] official
         'The reporter hit the official who invited the tycoon.'

    b. Object-modifying OR (O-OR)

        jizhe   *da-le* [*fuhao* *yaoqing* e_i *de*] *guanyuan$_i$*
         reporter   hit   [tycoon   invite e_i DE] official
         'The reporter hit the official who the tycoon invited.'

Garden paths to the rescue!

Y'all know what a garden path sentence is, right? If not:

Example of a garden path sentence in English

The horse raced past the barn fell.

- Grammatical English. Read it a couple of times until you've figured it out.
- Explainable via high surprisal.
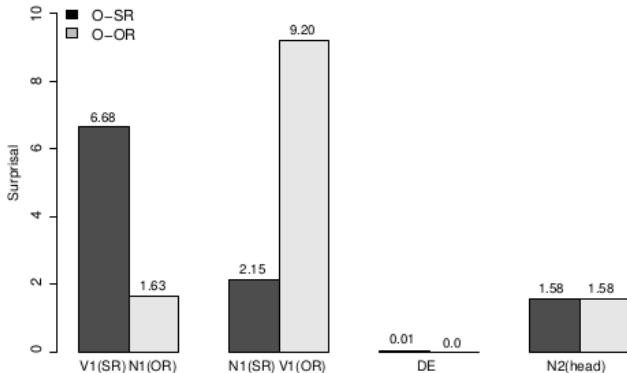- Possible that subject-modifying Chinese RCs form a garden path!

Do they?

Chen et al. build and train a simple PCFG for Chinese RCs. Then calculate surprisal as above.
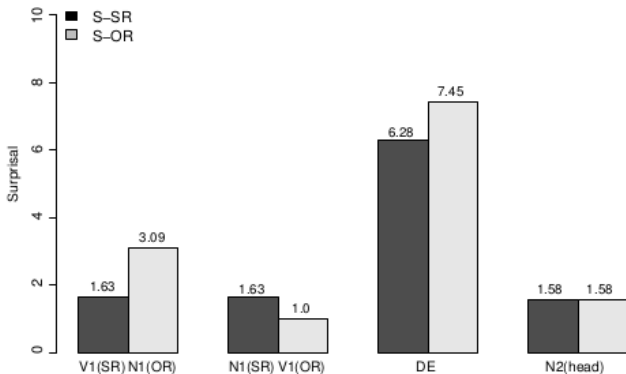
(a) Total: S-SR 11.12 vs. S-OR 13.12

Do they?

Chen et al. build and train a simple PCFG for Chinese RCs. Then calculate surprisal as above.

(b) Total: O-SR 10.42 vs. O-OR 12.41

So the point is, Chinese ORCs actually look like garden paths in the subject-modifying condition.

And surprisal might be sufficiently explanatory. (Or is it?)

Now a final quote. . .

'I'm sure I didn't mean' Alice was beginning, but the Red Queen interrupted her impatiently.

'That's just what I complain of! You should have meant! What do you suppose is the use of a child without any meaning? Even a joke should have some meaningand a child's more important than a joke, I hope. You couldn't deny that, even if you tried with both hands.'

'I don't deny things with my hands,' Alice objected.

'Nobody said you did,' said the Red Queen. 'I said you couldn't if you tried.'

'She's in that state of mind', said the White Queen 'that she wants to deny somethingonly she doesn't know what to deny!'

'A nasty, vicious temper,' the Red Queen remarked. . .

— Lewis Caroll, *Through the Looking Glass*