# Linzen and F. Jaeger 2015,and L. Jäger et al 2015 (readings 04, 05)

Shravan Vasishth

18 Nov 2015, Tokyo

# Introduction

- I will discuss the Linzen and Jaeger paper with slides and by referring to the paper.
- The Lena Jaeger et al paper I will only discuss superficially as I already talked about it in the early days of the course.

# Surprisal

We (somehow) compute the conditional probability of word n+1 given words 1...,n: $Prob(w_{n+1} \mid w_1, \ldots, w_n)$.

One empirical method for doing this is by computing cloze probability. A computational method is to use a corpus of text. Here's a crude method:

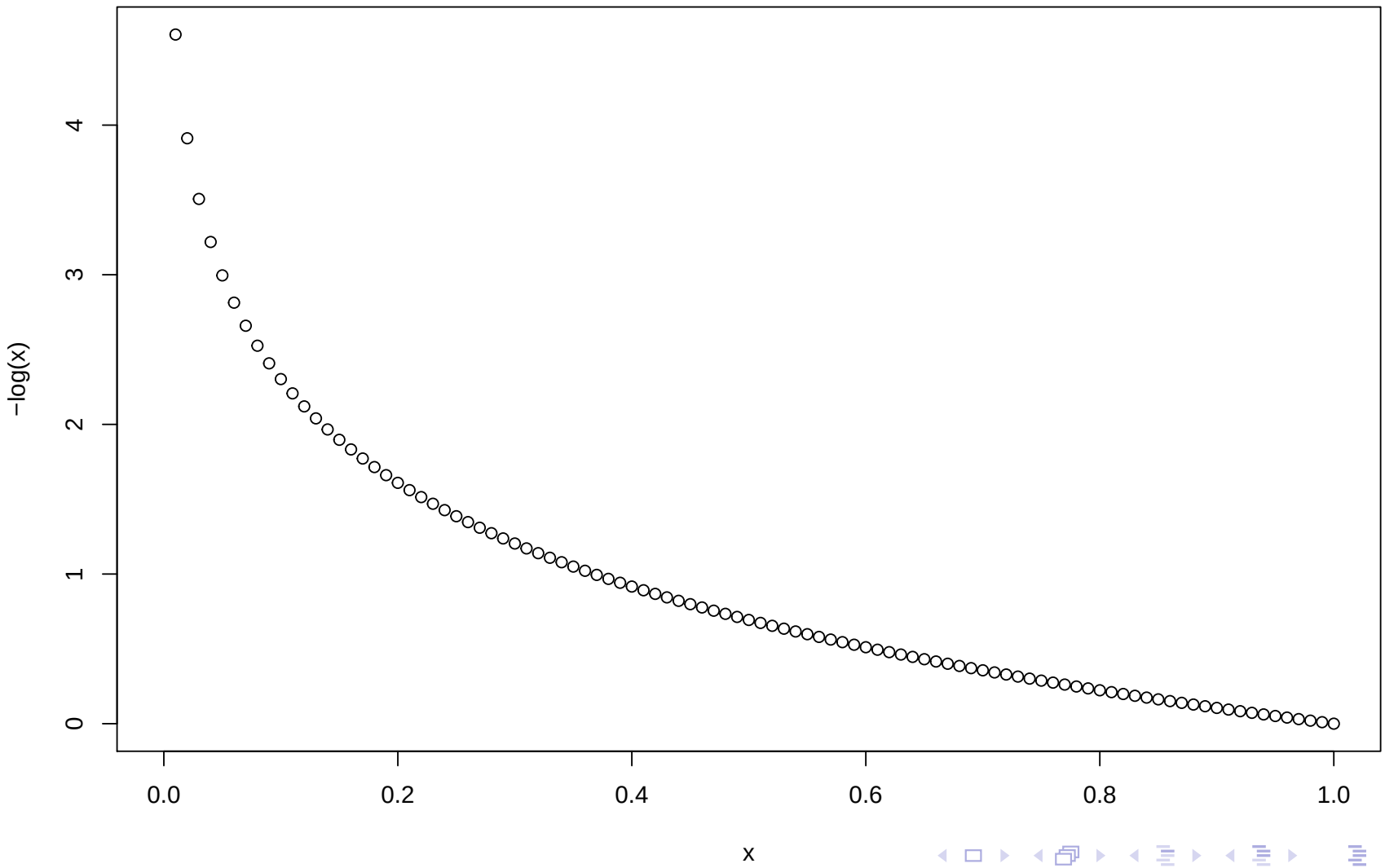
- List all sentences that begin with part of speech

Det

- List all possible continuations $i = 1, \ldots, k$, count number of cases $n_i$ and calculate relative frequencies: $n_i / \sum_{i=1}^{k} n_i$. This gives you a bigram surprisal for each possible continuation.

# Surprisal

```
x<-seq(0,1,by=0.01)
plot(x,-log(x))
```



# Entropy

Entropy is the average length of a message, measured in bits, that would have to sent to describe a sample.

- Example 1: Consider a fair coin, with Prob(Heads)=0.5

In one trial, entropy is

```
-(0.5*log2(0.5)+0.5*log2(0.5))
```

```
## [1] 1
```

You need to transmit only one bit to express the outcome.

# Entropy

Entropy is the average length of a message, measured in bits, that would have to sent to describe a sample.

- Example 2: Consider a two-headed coin, with Prob(Heads)=1

In one trial, entropy is 0:

```
-(1*log2(1))
```

```
## [1] 0
```

You need to transmit zero bits to express the outcome.

# Entropy and entropy reduction

## Applying entropy to sentence processing

Suppose that only two continuations are possible after a determiner:

- Prob(N | Det) = 0.9
- Prob(Adjective | Det) = 0.1

[Note: these have to sum to 1.]

Then, entropy at the determiner is:

```
(ent_det<- -(0.9*log2(0.9)+0.1*log2(0.1)))
```

```
## [1] 0.4689956
```

# Entropy and entropy reduction

Now, if the high probability continuation occurs, the next possible continuation might be

Prob(Verb| Det N) = 1

The new entropy is:

```
## total certainty:
(ent_detN_CERTAIN<- -(1*log2(1)))
```

```
## [1] 0
```

We get a large entropy reduction:

```
ent_detN_CERTAIN - ent_det
```

```
## [1] -0.4689956
```

# Entropy and entropy reduction

Consider an alternative scenario:

Suppose that, as before, only two continuations are possible after a determiner:

- Prob(N| Det) = 0.9
- Prob(Adjective | Det) = 0.1

# Entropy and entropy reduction

But after seeing a noun, suppose there was a very likely continuation, vs an unlikely continuation:

-Prob(Verb| Det N) = 0.999

-Prob(Relative Pronoun | Det N) = 0.001

```
## low entropy:
(ent_detN_UNCERTAIN<- -(0.999*log2(0.999)+0.001*log2(0.001)
```

```
## [1] 0.01140776
```

There is a large reduction in uncertainty at the Noun:

```
ent_detN_UNCERTAIN - ent_det
```

```
## [1] -0.4575878
```

# Surprisal vs Entropy Reduction

Surprisal is all about how surprised we are *at the current point in the sentence*

Entropy reduction is alsoabout our uncertainty at the current point but can be computed (i) with reference to what comes immediately after the current point (call it ER1) or (ii) about what the *rest* of the sentence looks like (ER2).

There is an interesting theoretical question here: *how much forward prediction do we really do*? In other words, Is ER1 or ER2 (my terms) the correct metric?

# An aside on inhibition and entropy

## This is Bruno Nicenboim's observation (p.c.)

Notice an interesting point that nobody seems to have discussed yet:

- When we have a highly probable continuation vs many low probability continuations, entropy will be lower than when we have many equiprobable continuations: implies possibly high ER (high processing cost)
- Inhibition predicts that the low entropy case will be *harder* than the high entropy case; this is because it would be harder to inhibit the highly likely candidate. So this is an alternative to the ERH.

Examples in class.

# Linzen and Jaeger 2015

Verbatim quote from the abstract:

- In a self-paced reading study, we use lexical subcategorization distributions to factorially manipulate both the strength of expectations and the uncertainty about them.
- We compare two types of uncertainty: uncertainty about the verb's complement, reflecting the next prediction step; and uncertainty about the full sentence, reflecting an unbounded number of prediction steps.
- We find that uncertainty about the full structure, but not about the next step, was a significant predictor of processing difficulty: Greater reduction in uncertainty was correlated with increased reading times (RTs).
- We additionally replicated previously observed effects of expectation violation (surprisal), orthogonal to the effect of uncertainty. This suggests that both surprisal and uncertainty affect human RTs.

# Linzen and Jaeger 2015

Example of uncertainty calculation

He accepted

possible continuations: 80% NP 20% S

Entropy:

```
-(0.8 * log2(0.8) + 0.2 * log2(0.2))
```

```
## [1] 0.7219281
```

# Linzen and Jaeger 2015

He forgot

possible continuations: 55% NP 9% S 18% PP 14% Inf

```
-(0.55 * log2(0.55) + 0.09 * log2(0.09)+
    0.18 * log2(0.18)+0.14 * log2(0.14))
```

```
## [1] 1.629445
```

Linzen and Jaeger's question: Does the difference in entropy between *accepted* and *forgot* lead to differences in processing difficulty?

# Hypothesis 1: The competition hypothesis (Elman et al)

"it may be costly to generate and maintain a larger number of predictions that compete with each other, especially if their probabilities are similar."

# Hypothesis 2: Entropy reduction hypothesis (Hale)

"it is reduction in uncertainty that is costly rather than the mere existence of uncertainty"

LJ consider two variants: - single-step prediction (prediction of the next syntactic step) - full prediction (prediction of the entire syntactic structure of the sentence)

A priori, *full prediction* in its strong form seems pretty weird and unrealistic to me. (Consider what the subject might predict when they see *The* . . . )

*Yet, full prediction is what LJ find evidence for.*

# Aside: Levy & Gibson 2013 critique of entropy reduction

"This state of affairs has led others to question the extent to which these studies provide support for the role of uncertainty-based hypothesis, in particular the entropy reduction hypothesis (Levy & Gibson, 2013)."

Actually, in Levy and Gibson, there is no discussion of ERH. The paper points to the Levy et al 2013 paper on Russian relative clauses for a discussion of ERH.

However, in the Levy et al Russian RC paper, there isn't much evidence *against* ERH either, so ERH remains a viable explanation (this is my opinion).

# Problems with previous studies (according to LJ)

Previous studies have never done a systematic comparison of ERH (in either variant) and surprisal. (see page 6)

"The goal of the current study is to assess the effects of entropy and entropy reduction in the same materials, within the same syntactic framework, while avoiding structural confounds."

# Key predictions for this study

- Entropy at word n should cause a slowdown at word n (=Competition model)
- Entropy reduction at word n should cause a slowdown at word n (Hale's ERH)

LJ formulate two versions of entropy (single-step and full). So we have two variants of entropy calculations, and then the competing accounts

| | Entropy | ER |
|---|---|---|
| Single step | NO | NO |
| Full | NO | YES |

Although they downplay this in the paper a lot, they come out in favor of Full, ER. There is no evidence in favor of other cells.

# Experiment (SPR)

The men / discovered / (that) / the island / had been invaded

The critical regions are

- the island: point of ambiguity (entropy high or low)
- had been invaded: disambiguation point

"Subcategorization frequencies were taken from Gahl, Jurafsky, and Roland's (2004) database (described in more detail below)."

# Holding surprisal constant, varying entropy

"*find* and *propose* have a similar SC subcategorization probability (0.22 and 0.25, respectively) and thus similar SC surprisal. But *propose* occurs with multiple other frames (NP: 0.57; infinitives: 0.14), whereas for *find* the NP frame is the only alternative to SC that occurs with a substantial probability (0.72). As a result, *propose* has higher subcategorization entropy than *find* (1.56 vs. 1.09)."

# The experiment design and predictions

The men / discovered / (that) / the island / had been invaded

Predictions at verb *discovered*:

- Under entropy: high entropy at verb should cause increased RTs
- Under ERH: low entropy at verb should cause increased RTs

Surprisal predictions at disambiguation point:

- Disambiguation to sentential complement (high surprisal) should be costly.

# Materials

Eight verbs in each cell:

| | Low Surprisal | High Surprisal |
|---|---|---|
| Low entropy | | |
| High entropy | | |

See Table 1.

# Primary results (referring to the a priori predictions)

The men / discovered / (that) / the island / had been invaded

Predictions vs results at verb *discovered*:

- Under entropy: high entropy at verb should cause increased RTs (No effect found)
- Under ERH: low entropy at verb should cause increased RTs (No effect found)

In addition: "Unexpectedly, low SC surprisal verbs were read more slowly, though this difference was only marginally significant ($p = .09$)."

# Primary results (referring to the a priori predictions)

Surprisal predictions at disambiguation point:

- Disambiguation to sentential complement (high surprisal) should be costly. (Validated)

"The absence of an entropy effect does not support either the competition or the entropy reduction hypotheses."

From this analysis of the a priori predictions in the paper, my conclusion would have been in favor of surprisal! Why do the authors argue for ERH? Answer: the secondary analysis (Section 5)

# Secondary analysis using full prediction (section 5)

- The secondary analysis in this paper was somewhat hard going!
- The key point is that the SC will have many possible instantiations = high entropy (say 40 bits). This has the paradoxical effect that when we are very certain that an SC is coming up, entropy will be **high**

```
## very sure of SC continuation
.9*40+.1*14
```

```
## [1] 37.4
```

```
## equally unsure of SC and NP continuation
.5*40+.5*14
```

```
## [1] 27
```

Consequence: in the ambiguous region, we will have **lower** entropy in ambiguous sentences than unambiguous sentences!

# Secondary analysis using full prediction (section 5)

Results

- At verb: effect of full entropy reduction on RTs (relates to the surprising surprisal effect in the primary analysis; surprisal and full entropy are highly correlated)
- At verb: marginal negative effect of entropy on RTs (could this happen because subjects want to move forward faster if they are more uncertain? Maybe this is not evidence against entropy! An eyetracking study might show that outgoing saccade length is longer in high entropy conditions).
- At ambiguous region: they did find some effects (see their Table summary), but I am not very clear on what is going on here (see below).
- At disambiguating region: surprisal had an effect (expected direction)

So the evidence for ERH (full entropy reduction) is at the verb.

# LJ discussion

- At verb, full ERH shows the predicted effect
- At disambiguating region, surprisal shows the predicted effect

Their conclusion: both surprisal and ERH play a role.

# Lookahead distance

- The authors discuss what the appropriate lookahead distance might be.
- For the ambiguous region, they point out that under infinite lookahead, the entropy of the SC will dominate in determining entropy before the disambiguation; hence, entropy will be *higher* for the *unambiguous* conditions. They found the opposite pattern. So this is effectively an empirical argument against full entropy, if I understand this correctly.

# Lookahead distance

I would expect the forward prediction to increase by word position; so I predict that the best predictor would be a dynamic computation of Entropy (or Entropy Reduction) by position

- At *The...*, prediction would be limited to the next upcoming word.
- But at *The man hit the...*, prediction might go all the way to the (an) end of the sentence.

It would be hard to develop an independent criterion for how much forward prediction to assume by position.

# Concluding remarks

- This is a very thought-provoking paper, and represents a nice attempt to evaluate the predictions of surprisal and entropy.
- AFAIK, there is no work on Japanese or Chinese that looks at these issues using corpus data and online reading experiments.
- This paper also shows that the predictions relating to entropy are non-trivial to derive; certainly not something you can or should work out by introspection. I certainly tend to forget this sometimes.
- The evidence for entropy in the paper is a bit weak in my opinion, the evidence for surprisal seems stronger.

# Concluding remarks

-The primary analysis is done over four regions, and the effect of surprisal has t-value 2.17. This corresponds to an approximate p-value of:

```
2*pnorm(-2.17)
```

```
## [1] 0.03000685
```

However, a Bonferroni correction for multiple comparisons would leave us with a Type I error of $0.05/4 = 0.0125$.

So the evidence for surprisal in the primary analysis is somewhat weak in the primary analysis (but note that Bonferroni is conservative).

# Concluding remarks

Note that in psycholinguistics we ignore this issue, but this will probably change in the future. See e.g.

von der Malsburg, T. and Angele, B. (2015). False positive rates in standard analyses of eye movements in reading. Technical report. Manuscript published on arXiv.

`http://arxiv.org/abs/1504.06896`

# Concluding remarks

- In the secondary analysis, the evidence for entropy has p-values barely squeezing in below 0.05 (no t-values are shown)
- At verb: p=0.047, 0.08
- At ambiguous region: very low p-values but I can't understand what is going on there. As the authors also note: "it is possible that the observed effect stems from other differences between the ambiguous and unambiguous sentences, such as the presence of temporary ambiguity."
- At the disambiguating region, surprisal p=0.01, ERH p=0.03

So, if we rely on p-values, the evidence for surprisal seems solid, not sure about ERH.

# Concluding remarks

- But really, p-values don't really tell us much about the hypothesis of interest.
- They create a false sense of accept-and-reject mentality in the reader, whereas the evidence is usually graded (e.g., there is evidence for ERH in the data, we don't need to commit to a yes/no decision about the evidence)
- The estimated mean and the confidence interval around that mean are more important for understanding what the evidence is in the data.
- One could just go all the way and just compute Bayesian 95% credible intervals and just report those. This is what I do.