UID across linguistic levels and Limits of UID and Surprisal Theory

Information Theory Lecture

Vera Demberg and Matt Crocker

Information Theory Lecture, Universität des Saarlandes

April 20th, 2015
Uniform Information Density Hypothesis (Jaeger 2010)

Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).
Uniform Information Density Hypothesis (Jaeger 2010)

Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).

What if syntactic constraints mean that there is higher / lower than optimal ID?

▶ can this be compensated for at other linguistic levels?
Uniform Information Density Hypothesis (Jaeger 2010)

Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).

What if syntactic constraints mean that there is higher / lower than optimal ID?

- can this be compensated for at other linguistic levels?
- We’ve already seen something related in Mahowald 2013 paper: choice of lexemes was dependent on ID (math / mathematics)
Uniform Information Density Hypothesis (Jaeger 2010)

Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).

What if syntactic constraints mean that there is higher / lower than optimal ID?

- can this be compensated for at other linguistic levels?
- We’ve already seen something related in Mahowald 2013 paper: choice of lexemes was dependent on ID (math / mathematics)

If we have high syntactic/semantic surprisal, are we going to observe longer speech durations?
Effect of UID across linguistic levels

- AMI meeting corpus:
  - has fine-grained word durations.
  - “correctly” spelled transcriptions: necessary for parsing.
  - speaker turns, disfluencies annotated
- Focus group meetings for designing remote control.
- Over 1M words. Native and non-native English speakers.
Corpus data containing transcriptions of spoken dialog is noisy.

Preprocessing steps:

- Remove non-words such as “hmm” and “uhm”.
- Remove incomplete words or incorrectly transcribed words (“recogn”). (loss rate 2.9% of word types)
- Remove incomplete sentences.
We parsed the corpus to obtain syntactic surprisal
We synthesized it to obtain maximally accurate word length estimates

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Native English</th>
<th></th>
<th>Non-native</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef</td>
<td>t-value</td>
<td>Sig</td>
<td>Coef</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.2947</td>
<td>149.74</td>
<td>***</td>
<td>0.3221</td>
</tr>
<tr>
<td>DurMARY</td>
<td>0.5304</td>
<td>69.27</td>
<td>***</td>
<td>0.4699</td>
</tr>
<tr>
<td>FreqAMI</td>
<td>-0.0226</td>
<td>-18.10</td>
<td>***</td>
<td>-0.0321</td>
</tr>
<tr>
<td>FreqGiga</td>
<td>-0.0264</td>
<td>-41.19</td>
<td>***</td>
<td>-0.0248</td>
</tr>
<tr>
<td>SrprslGiga-4</td>
<td>0.0018</td>
<td>5.36</td>
<td>***</td>
<td>0.0033</td>
</tr>
<tr>
<td>DurMARY:FreqGiga</td>
<td>-0.0810</td>
<td>-27.20</td>
<td>***</td>
<td>-0.0993</td>
</tr>
<tr>
<td>SrprslSyntax</td>
<td>0.0033</td>
<td>24.21</td>
<td>***</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

no of data points 320,592 391,106

**Result:** syntactic surprisal has some explanatory power beyond frequency and n-gram predictability.
Effect size

Effect size of syntactic surprisal on speech durations

- 7msec additional duration per “unit” of surprisal
- range of surprisal: [0, 25], most values [2, 15]
- “thing”: one instance with surprisal 2.179, other with 16.277
- this corresponds to a difference in duration of 104msec

Conclusion so far:

- Syntactic surprisal affects spoken word duration.
- Compatible with the notion that longer word durations to some extent “compensates” for high ID at other linguistic levels.
- Does this also hold for content-wise surprising words?
Estimating semantic surprisal is not trivial.

What we can do so far:

- re-weigh trigram probabilities based on semantic similarity
  (Mitchell 2011; Mitchell, Lapata, Demberg & Keller 2010)
  reimplementation by Stefan Fischer (BSc UdS 2014)
- probability of semantically fitting words is increased
- probability of badly fitting words is reduced.

Example

The author wrote the next chapter.
The child listened to the next chapter.
The busdriver handed out the next chapter.

tri-gram model: $P(\text{chapter}|\text{the next})$
Semantic language model

Components:

- trigram language model \( p(w_n | w_{n-2}, w_{n-1}) \)
- vector space model (simple bag-of-words model) \( \cos(\vec{c}, \vec{w}) \)

Example: The author wrote the next chapter.

- trigram probability: \( P(\text{chapter} | \text{the next}) \)
- semantic similarity: \( \cos(\vec{\text{author}} \ast \vec{\text{wrote}}, \vec{\text{chapter}}) \)

- scale to make sure it remains a proper probability model.
- (weighing only affects prediction of content words)
Results

Modelling (on content words only):
▶ with random slopes for trigram surprisal, word length and semantic surprisal under subject.
▶ Semantic surprisal on residuals: with random slope under subj as well.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>t-value</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.024059</td>
<td>3.38 ***</td>
<td></td>
</tr>
<tr>
<td>DurMARY</td>
<td>0.419718</td>
<td>135.71 ***</td>
<td></td>
</tr>
<tr>
<td>FreqAMI</td>
<td>-0.113547</td>
<td>-53.42 ***</td>
<td></td>
</tr>
<tr>
<td>FreqGiga</td>
<td>-0.071767</td>
<td>-28.80 ***</td>
<td></td>
</tr>
<tr>
<td>SrprslGiga-3gram</td>
<td>0.042968</td>
<td>15.64 ***</td>
<td></td>
</tr>
<tr>
<td>DurMARY:FreqAMI</td>
<td>-0.027267</td>
<td>-17.69 ***</td>
<td></td>
</tr>
<tr>
<td>SrprslSemantics</td>
<td>0.003673</td>
<td>2.398 **</td>
<td></td>
</tr>
</tbody>
</table>

Results:
▶ Semantic surprisal significantly improves model fit over a model with only trigram surprisal
▶ semantic surprisal is a significant positive predictor on residuals of the baseline model.
▶ semantically scaled surprisal achieves better model fit than trigram surprisal. (according to AIC, BIC, logLik)
Syntactic and semantic surprisal together

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>t-value</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.019907</td>
<td>2.80</td>
<td>***</td>
</tr>
<tr>
<td>DurMARY</td>
<td>0.413524</td>
<td>137.21</td>
<td>***</td>
</tr>
<tr>
<td>FreqAMI</td>
<td>-0.098595</td>
<td>-45.59</td>
<td>***</td>
</tr>
<tr>
<td>FreqGiga</td>
<td>-0.072078</td>
<td>-30.00</td>
<td>***</td>
</tr>
<tr>
<td>SrprslSyntax</td>
<td>0.065747</td>
<td>27.53</td>
<td>***</td>
</tr>
<tr>
<td>SrprslSemantics</td>
<td>0.016957</td>
<td>6.13</td>
<td>***</td>
</tr>
<tr>
<td>DurMARY:FreqAMI</td>
<td>-0.034701</td>
<td>-22.33</td>
<td>***</td>
</tr>
</tbody>
</table>

Table: Syntactic and semantic surprisal on content words.

- Correlation on content words: 0.6, overall: 0.75
- Models with syntactic and semantic surprisal improve over and above one another.
What happens if we train in-domain?

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>t-value</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.025</td>
<td>3.103</td>
<td>**</td>
</tr>
<tr>
<td>DurMARY</td>
<td>0.405</td>
<td>107.389</td>
<td>***</td>
</tr>
<tr>
<td>FreqAMI</td>
<td>-0.077</td>
<td>-24.726</td>
<td>***</td>
</tr>
<tr>
<td>FreqGiga</td>
<td>-0.063</td>
<td>-22.779</td>
<td>***</td>
</tr>
<tr>
<td>SrprslGiga-4</td>
<td>0.043</td>
<td>16.396</td>
<td>***</td>
</tr>
<tr>
<td>SrprslSemantics</td>
<td>-0.036</td>
<td>-10.082</td>
<td>***</td>
</tr>
<tr>
<td>DurMARY:FreqAMI</td>
<td>-0.029</td>
<td>-12.488</td>
<td>***</td>
</tr>
</tbody>
</table>

Table: Fixed effects of the baseline model with semantic surprisal.

Is this an effect of in-domain words?
### In-domain semantic surprisal model

<table>
<thead>
<tr>
<th>Interval of SrprslSemantics</th>
<th>Predictor</th>
<th>Coefficient</th>
<th>t-value</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[0, \infty[$</td>
<td>(Intercept)</td>
<td>-0.003</td>
<td>-1.162</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>SrprslSemantics</td>
<td>-0.027</td>
<td>-8.499</td>
<td>***</td>
</tr>
<tr>
<td>$[0, 2[</td>
<td>(Intercept)</td>
<td>-0.001</td>
<td>-0.135</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>SrprslSemantics</td>
<td>-0.071</td>
<td>-13.535</td>
<td>***</td>
</tr>
<tr>
<td>$[2, \infty[$</td>
<td>(Intercept)</td>
<td>-0.000</td>
<td>-0.088</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>SrprslSemantics</td>
<td>0.008</td>
<td>2.705</td>
<td></td>
</tr>
</tbody>
</table>

Table: Three models of SrprslSemantics as a random effect over the residuals of baseline models learned from the remaining fixed effects. The first model is over the entire range.

- Correlation is positive for high surprisal words, but negative for low-surprisal words.
- Many of the low-surprisal words are in-domain.
Effect of training the semantic language model

Tables: The effect of semantic surprisal predictions on spoken word durations.
Conclusions

- Choice of training data for language model can make a huge difference.
- Choice should be made based on what should be modelled.
- AMI corpus has lots of different speakers, who don’t work on the domain (design of a remote control) all day.
- Therefore, a domain-general model may reflect the speakers’ language models more than the domain-specific model.

This raises a few questions:

- What do our models look like / how adaptive are they?
- How much domain adaptation do speakers do?
- What’s the trade-off between short-term and long-term adaptation effects?
- What’s optimal?
Conclusions

▶ Choice of training data for language model can make a huge difference.
▶ Choice should be made based on what should be modelled.
▶ AMI corpus has lots of different speakers, who don’t work on the domain (design of a remote control) all day.
▶ Therefore, a domain-general model may reflect the speakers’ language models more than the domain-specific model.

This raises a few questions:
▶ What do our models look like / how adaptive are they?
▶ How much domain adaptation do speakers do?
▶ What’s the trade-off between short-term and long-term adaptation effects? What’s optimal?
Short-term vs. long-term adaptation

How online is it all?

We know:

- Language-users reduce words in predictable contexts (as just seen in study on AMI corpus).
- Previous research indicates that reduction may be stored in lexical representation if a word is often reduced. (Piantadosi et al., Mahowald et al.)
- If representation influences production regardless of context, production should be biased by how often each word has been reduced in the speakers prior experience.
- Is this all “online”? or are (reduced) representations stored in the lexicon?

This is explored in more detail in Seyfarth (2014).
Frequency vs. average predictability

Piantadosi et al. discussed word frequency vs. word predictability, showing that word predictability is a better predictor of word length.

- Here, relate frequency to average predictability vs. local predictability.

Example from Seyfarth (2014)

nowadays (black) and current (gray) have very similar frequency. current occurs in a small range of following contexts, while nowadays occurs in a wider range of contexts.

THIS IS BACKWARD BIGRAM SURPRISAL!
Testable hypothesis

**Question:**
Are word pronunciations influenced by their typical predictabilities or fully “online” by their current local predictability?

**Prediction:**
High-informativity words – words that are usually unpredictable – should have longer durations than those words which are usually predictable, when all other factors are held equal.
Method

- word durations from Buckeye and Switchboard
- forward and backward bigrams from Fisher part 2 corpus (dialogs)
- only nouns, verbs, adverbs and adjectives were included in the analysis.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
<th>$p(\chi^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>0.0257</td>
<td>0.0057</td>
<td>4.48</td>
<td>—</td>
</tr>
<tr>
<td>Baseline duration</td>
<td>0.5879</td>
<td>0.0150</td>
<td>39.32</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Syllable count</td>
<td>0.0592</td>
<td>0.0104</td>
<td>5.71</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Speech rate</td>
<td>−0.3406</td>
<td>0.0077</td>
<td>−43.97</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Bigram prob. given previous</td>
<td>−0.0102</td>
<td>0.0007</td>
<td>−15.00</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Bigram prob. given following</td>
<td>−0.0205</td>
<td>0.0007</td>
<td>−30.55</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Orthographic length</td>
<td>0.0437</td>
<td>0.0167</td>
<td>2.62</td>
<td>0.0089</td>
</tr>
<tr>
<td>Part of Speech = Adjective</td>
<td>0.0033</td>
<td>0.0032</td>
<td>1.04</td>
<td>(&lt; 0.0001)</td>
</tr>
<tr>
<td>Part of Speech = Adverb</td>
<td>−0.0172</td>
<td>0.0042</td>
<td>−4.09</td>
<td>—</td>
</tr>
<tr>
<td>Part of Speech = Verb</td>
<td>−0.0275</td>
<td>0.0022</td>
<td>−12.54</td>
<td>—</td>
</tr>
<tr>
<td>Informativity given following</td>
<td>0.0244</td>
<td>0.0023</td>
<td>10.77</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Table 1: Fixed effects summary for model of Buckeye word durations.
Method

- word durations from Buckeye and Switchboard
- forward and backward bigrams from Fisher part 2 corpus (dialogs)
- only nouns, verbs, adverbs and adjectives were included in the analysis.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
<th>$p(\chi^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTERCEPT</strong></td>
<td>0.0287</td>
<td>0.0023</td>
<td>12.62</td>
<td>—</td>
</tr>
<tr>
<td><strong>Baseline duration</strong></td>
<td>0.5363</td>
<td>0.0102</td>
<td>52.34</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><strong>Syllable count</strong></td>
<td>0.0492</td>
<td>0.0070</td>
<td>7.01</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><strong>Speech rate</strong></td>
<td>-0.3260</td>
<td>0.0044</td>
<td>-74.79</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><strong>Bigram prob. given previous</strong></td>
<td>-0.0082</td>
<td>0.0005</td>
<td>-18.17</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><strong>Bigram prob. given following</strong></td>
<td>-0.0227</td>
<td>0.0004</td>
<td>-53.91</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><strong>Orthographic length</strong></td>
<td>0.1343</td>
<td>0.0115</td>
<td>11.69</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><strong>Part of Speech = Adjective</strong></td>
<td>-0.0051</td>
<td>0.0021</td>
<td>-2.36</td>
<td>(&lt; 0.0001)</td>
</tr>
<tr>
<td><strong>Part of Speech = Adverb</strong></td>
<td>-0.0186</td>
<td>0.0026</td>
<td>-7.18</td>
<td>—</td>
</tr>
<tr>
<td><strong>Part of Speech = Verb</strong></td>
<td>-0.0410</td>
<td>0.0017</td>
<td>-23.87</td>
<td>—</td>
</tr>
<tr>
<td><strong>Informativity given previous</strong></td>
<td>0.0040</td>
<td>0.0016</td>
<td>2.48</td>
<td>0.0131</td>
</tr>
<tr>
<td><strong>Informativity given following</strong></td>
<td>0.0142</td>
<td>0.0016</td>
<td>8.72</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Table 4: Fixed effects summary for model of Switchboard word durations.
Conclusions and Discussion

- Conclusion of Seyfarth 2014: Words that usually appear in predictable contexts are somewhat reduced in all contexts, even those in which they are unpredictable.

- At the same time, he confirms that local predictability is a strong factor affecting word durations, too.

- Will these results hold up with better language models?

- Is it an effect of collocations / compounding?

- Why is backward predictability such a good predictor?
15 min break.

then continue with limits of UID.
Potential counter-evidence: Dutch morphology
1. Potential counter-evidence: Dutch morphology

2. Relationship between surprisal and processing is not super-logarithmic (= no free lunch)
1. Potential counter-evidence: Dutch morphology

2. Relationship between surprisal and processing is not super-logarithmic (＝ no free lunch)

3. What happens to the linking theory of surprisal and processing difficulty for very high / low values of surprisal?
1. Potential counter-evidence: Dutch morphology

2. Relationship between surprisal and processing is not super-logarithmic (= no free lunch)

3. What happens to the linking theory of surprisal and processing difficulty for very high / low values of surprisal?

4. Stretching assumptions to the extreme: What if UID was the only factor affecting language evolution?
Potential counter-evidence: Dutch morphology

Relationship between surprisal and processing is not super-logarithmic (= no free lunch)

What happens to the linking theory of surprisal and processing difficulty for very high / low values of surprisal?

Stretching assumptions to the extreme:
What if UID was the only factor affecting language evolution?
Dutch interfixes

Dutch morphology has interfixes for word compounds, similar to German.

About Dutch interfixes
interfixes -s- or -e(n)-

- oorlog-s-verklaring (declaration of war)
- dier-en-arts (veterinary)
- null-interfix is most common in Dutch.
- no deterministic rules; but distribution correlates with whether the first part of the compound typically occurs with the interfix, and to a lesser extent whether the second part of the compound does so.
- hence, the interfix is predictable from the parts of the compound.

UID would predict that that the duration of the interfix in speech is shorter when the interfix is more predictable from context.
Dutch interfixes

Dutch morphology has interfixes for word compounds, similar to German.

About Dutch interfixes

- S- or e(n)-
  - oorlog-s-verklaring (declaration of war)
  - dier-en-arts (veterinary)
  - null-interfix is most common in Dutch.
  - no deterministic rules; but distribution correlates with whether the first part of the compound typically occurs with the interfix, and to a lesser extent whether the second part of the compound does so.
  - hence, the interfix is predictable from the parts of the compound.

UID would predict that that the duration of the interfix in speech is shorter when the interfix is more predictable from context.

The data says that interfixes were pronounced longer when they were more predictable!
Experiment

100h of read speech yielded 1155 -s- and 742 -e(n)- (excluding ambiguous ones where the second part of compound started with a similar sound).

- durations of interfixes annotated via ASR
- realization of interfixes were transribed by phoneticians
- automatic tool was then used to segment signal

Results:

- if the first part of the compound was usually followed by -s-, the duration of -s- was longer
- same for -e(n)-
- there was also a marginal effect for more frequent words to have longer interfix durations
- the paper doesn’t give detail regarding correlation of predictors, or report whether the effect was stable without other predictors present
Explanation for this observed interfix effect: a paradigmatic effect, where it’s spoken clearly when the speaker is certain about it.

related effects of paradigmatic enhancement also found in English suffixes (Cohen, 2014)

UID holds for other cases of Dutch spoken word durations, in particular, optional schwa insertion between liquid and obstruant in Dutch, e.g. film / fil@m, dorp / dor@p (Tily & Kuperman, 2012)

other lengthening effects or drop of syllables in Dutch consistent with UID (Pluymaekers, M., Ernestus, M., and Baayen, R., 2005b)
1. Potential counter-evidence: Dutch morphology

2. Relationship between surprisal and processing is not super-logarithmic (≠ no free lunch)

3. What happens to the linking theory of surprisal and processing difficulty for very high / low values of surprisal?

4. Stretching assumptions to the extreme:
   What if UID was the only factor affecting language evolution?
Why would the effect of surprisal on processing difficulty be super-logarithmic?

- if anything that’s more dense than channel capacity leads to more than linearly more difficulty, utterances that adhere most closely to UID have lowest total difficulty.
- a peak then cannot be “balanced out” by a similar size trough.
- therefore, successful communication needs UID (audience design account).

(but where does the difficulty occur? At the point of the peak, or when retrieving that information later at the trough?)
Surprisal vs. reading times

Proposed relationships between predictability and reading time

- Linear (guessing)
- Super-logarithmic (could explain UID effects)
- Logarithmic (optimal visual discrimination, highly incremental processing)
- Reciprocal (hypothesized by Narayanan & Jurafsky, 2004)

Reading time

Probability (log scale)
Surprisal vs. reading times

Proposed relationships between predictability and reading time

- Linear (guessing)
- Super-logarithmic (could explain UID effects)
- Logarithmic (optimal visual discrimination, highly incremental processing)
- Reciprocal (hypothesized by Narayanan & Jurafsky, 2004)

Eye-tracking

Self-paced reading

Vera Demberg and Matt Crocker (UdS) Limits of UID April 20th, 2015 26 / 41
Audience interpretation time is unaffected by uniformity.
Conclusions from lack of super-logarithmic relation

- Audience interpretation time is unaffected by uniformity.
- But overloading the channel is still going to be bad for communication.
Conclusions from lack of super-logarithmic relation

- Audience interpretation time is unaffected by uniformity.
- But overloading the channel is still going to be bad for communication.
- Therefore, we’d still end up with optimal communication being uniform close to channel capacity.
Conclusions from lack of super-logarithmic relation

- Audience interpretation time is unaffected by uniformity.
- But overloading the channel is still going to be bad for communication.
- Therefore, we’d still end up with optimal communication being uniform close to channel capacity.
- These accounts differ somewhat in how closely UID should be adhered to, latter account allows for more variation (it’s possible to catch up without paying additional price).
Conclusions from lack of super-logarithmic relation

- Audience interpretation time is unaffected by uniformity.
- But overloading the channel is still going to be bad for communication.
- Therefore, we’d still end up with optimal communication being uniform close to channel capacity.
- These accounts differ somewhat in how closely UID should be adhered to, latter account allows for more variation (it’s possible to catch up without paying additional price).
- Working memory will be needed more for decreasing uniformity, this may also come at a cost (measurable at retrieval time). Nothing about this question has been shown in the paper.
Conclusions from lack of super-logarithmic relation

- Audience interpretation time is unaffected by uniformity.
- But overloading the channel is still going to be bad for communication.
- Therefore, we’d still end up with optimal communication being uniform close to channel capacity.
- These accounts differ somewhat in how closely UID should be adhered to, latter account allows for more variation (it’s possible to catch up without paying additional price).
- Working memory will be needed more for decreasing uniformity, this may also come at a cost (measurable at retrieval time). Nothing about this question has been shown in the paper.
- Overall, I see no strong argument against UID from this result.
1. Potential counter-evidence: Dutch morphology

2. Relationship between surprisal and processing is not super-logarithmic
   (= no free lunch)

3. What happens to the linking theory of surprisal and processing
difficulty for very high / low values of surprisal?

4. Stretching assumptions to the extreme:
   What if UID was the only factor affecting language evolution?
Prediction and Prediction Error

Framing Surprisal as Prediction and Learning

- We understand the world through prediction (necessary in order to robustly draw inferences from noisy and incomplete input).
- The extent to which these predictions are violated (prediction error) is the amount of new information gained (a.k.a. Shannon information, surprisal).
- Amount of information is proportional to effort to process this information.

[Eye-tracking and Self-paced reading graphs]

[from Smith and Levy, 2013]
The experienced soldiers *warned* about the dangers of conducting the midnight raid.

(prediction error (surprisal))

- Reflected in observable processing difficulty
- Causes changes to expectations / beliefs / weights

(from Florian Jaeger)
Does this always hold? Examples from learning:

► maybe some cues are more salient to us than others, we learn from them more readily

► 2nd language learning: some distinctions present in new language but not in old language don’t get learned easily

► categorial perception in speech comprehension

→ we’ve learned what to pay attention to

► there is SO much signal all the time, our brain selects what information to act on – what’s the role of attention in surprisal?
Apparent paradoxes

In language learning

- novices pay attention to (=processing evidence, reading times) high frequency patterns
- experts pay attention to low frequency patterns

- information about morphological marking may be in the signal, but it’s not processed by some learners.

→ Always need to define information with respect not only to a comprehender model in terms of predictability, but in terms of how much it changes our beliefs about the world.
→ interpretable information
Example from visual processing

- blue and snow screen similarly unsurprising to human
- blue screen is maximally predictable, snow is maximally unpredictable
- also: Goldilocks effects in visual attention of infants

(from Itti and Baldi, 2009)
Children look when it’s the “right” amount of information
Can we map this back onto language?

- someone (English native with no second language) reading a fairy tale in English, a fairy tale in English, a fairy tale in Chinese
- language model for surprisal trained on English
- model: fairy tale < lecture < Chinese
- reader: Chinese < fairy tale < lecture
- problem in linking hypothesis.
- (can probably get similar effect by replacing Chinese with medical report / physics article)
Is user model together with channel capacity the solution?

- We should get optimal processing when good use of channel capacity.
- Same signal can have different amount of information depending on user model.
- With respect to the information processing channel, the same signal can thus lead to different amounts of processable information.
- Channel and time: when don't people just take more time to make up for lack in channel capacity?
- When can meaning not be recovered?
- Not clear whether we can really make do without GOALS.
For discussion

- How does our brain learn what to pay attention to?
For discussion

- How does our brain learn what to pay attention to?
- What information should our brain process?
For discussion

- How does our brain learn what to pay attention to?
- What information should our brain process?
  - example: some colors capture our attention more than others
  - example: some distinctions in phonetics are more noticeable to us than others
- Can we relate this to Shannon information/surprisal?
- Can we conceptualize this without the notion of goals?
  - if not, how/why is it valid to ignore goals in our models?
- What roles do features play that have no effect in time?
  - (more informative signal that takes the same time but higher speaker effort)
For discussion

- How does our brain learn what to pay attention to?
- What information should our brain process?
- Example: some colors capture our attention more than others
- Example: some distinctions in phonetics are more noticeable to us than others
- Can we relate this to Shannon information / surprisal?
  Can we conceptualize this without the notion of goals?
  If not, how / why is it valid to ignore goals in our models?
For discussion

- How does our brain learn what to pay attention to?
- What information should our brain process?
- example: some colors capture our attention more than others
- example: some distinctions in phonetics are more noticeable to us than others
- Can we relate this to shannon information / surprisal? Can we conceptualize this without the notion of goals? If not, how / why is it valid to ignore goals in our models?
- What roles do features play that have no effect in time? (more informative signal that takes same time but higher speaker effort)
Potential counter-evidence: Dutch morphology

Relationship between surprisal and processing is not super-logarithmic (= no free lunch)

What happens to the linking theory of surprisal and processing difficulty for very high / low values of surprisal?

Stretching assumptions to the extreme: What if UID was the only factor affecting language evolution?
UID predicts that the production system is set up in such a way that information density directly or indirectly affects speakers’ preferences during production. That is, as speakers incrementally encode their intended message, their preferences at choice points should be affected by the relative information density of different continuations compatible with the intended meaning. Hence, UID does not predict that every word provides the same amount of information, but rather that, where grammar permits, speakers aim to distribute information more uniformly without exceeding the channels capacity. (Jaeger, 2010)
Assumptions stretched to the extreme

UID predicts that the production system is set up in such a way that information density directly or indirectly affects speakers’ preferences during production. That is, as speakers incrementally encode their intended message, their preferences at choice points should be affected by the relative information density of different continuations compatible with the intended meaning. Hence, UID does not predict that every word provides the same amount of information, but rather that, where grammar permits, speakers aim to distribute information more uniformly without exceeding the channels capacity. (Jaeger, 2010)

Ferrer-i-Cancho et al., 2013 discuss what would happen if **UID was the one and only force** affecting language use and language evolution.
UID predicts that the production system is set up in such a way that information density directly or indirectly affects speakers’ preferences during production. That is, as speakers incrementally encode their intended message, their preferences at choice points should be affected by the relative information density of different continuations compatible with the intended meaning. Hence, UID does not predict that every word provides the same amount of information, but rather that, where grammar permits, speakers aim to distribute information more uniformly without exceeding the channels capacity. (Jaeger, 2010)

Ferrer-i-Cancho et al., 2013 discuss what would happen if UID was the one and only force affecting language use and language evolution.

Take a moment to think about what that would mean for how language should be structured.
Strong assumptions

**Full UID:** For any sequence of linguistic elements (grammatical or not), the information density of the sequence should be entirely uniform.

**Strong UID:** For any sequence in the language (only grammatical), the information density of the sequence should be entirely uniform.

→ this is clearly not the case.

Factors not considered?
▶ language is for communication!
▶ The set of possible messages is not finite.
▶ Properties of messages will change as a function of changes in the world.
Strong assumptions

**Full UID:** For any sequence of linguistic elements (grammatical or not), the information density of the sequence should be entirely uniform.

**Strong UID:** For any sequence in the language (only grammatical), the information density of the sequence should be entirely uniform.

The paper shows that for full UID, it follows that all linguistic units must be independent from one another.
→ this is clearly not the case.
Strong assumptions

**Full UID:** For any sequence of linguistic elements (grammatical or not), the information density of the sequence should be entirely uniform.

**Strong UID:** For any sequence in the language (only grammatical), the information density of the sequence should be entirely uniform.

The paper shows that for full UID, it follows that all linguistic units must be independent from one another.

→ this is clearly not the case.

Factors not considered?
Strong assumptions

**Full UID:** For any sequence of linguistic elements (grammatical or not), the information density of the sequence should be entirely uniform.

**Strong UID:** For any sequence in the language (only grammatical), the information density of the sequence should be entirely uniform.

The paper shows that for full UID, it follows that all linguistic units must be independent from one another.

→ this is clearly not the case.

Factors not considered?

- language is for communication!
- The set of possible messages is not finite.
- Properties of messages will change as a function of changes in the world.
The roles of UID and Surprisal

The role of UID
UID is a kind of selectional mechanism, it supports the selection of shorter forms for predictable events, this happens online, and can lead to lexicalization and long term affect language evolution.

- it does not contain a mechanism for introducing new structures or new words.

The role of Surprisal
There is substantial evidence of correlation between surprisal and behavioural measures such as reading times and ERPs, there is evidence that surprisal might be an error signal that drives learning.

- it does not have a mechanism for accounting for e.g. memory effects.

Relevant question for us to take home:
What are other factors affecting language evolution and language use?