The ERP response to the amount of information conveyed by words in sentences.
• **Information Theory**
  
  • Each word conveys a certain amount of information
    • should be predictive of the amount of effort required to process the word
  
  • The amount of information
    • Computed from probabilistic models of the language
  
  • The amount of cognitive effort
    • Observed by measuring word reading times
    • More informative words \( \leftrightarrow \) longer reading time

• The effect of word information on the ERP response?
Introduction – Quantifying word information

• Surprisal

\[
\text{surprisal } (w_{t+1}) = - \log P(w_{t+1}|w_1...t)
\]

• A measure of the extent to which its occurrence was unexpected
  • w: actual \textit{words} or \textit{PoSs}

• Entropy Reduction

\[
H(W_{t+1...k}) = - \sum_{w_{t+1...k}} P(w_{t+1...k}|w_1...t) \log P(w_{t+1...k}|w_1...t)
\]

• Decrease when the next word or PoS is encountered
  • Quantify how much ambiguity is resolved by the current \textit{word} or \textit{PoS}

• Use 4 definitions of word information
Introduction – The present study

• Objectives (twofold)
  • Investigate whether a relation between word information and ERP amplitude exists
    • Look at 6 ERP components
      • N400
      • Post-N400 Positivity
      • Early Post-N400 Positivity
      • P600
      • Left Anterior Negativity
      • Early Left Anterior Negativity
    • Compare the surprisal and entropy reduction measures
      • Expect the effect of word surprisal on the size of N400
• Objectives (twofold)
  • Compare the explanatory value of different probabilistic language models assumptions closer to cognitive reality more predictive
  • Compare 3 model types:
    • n-gram models
      • do not embody any cognitive or linguistic theory
    • recurrent neural networks (RNN)
      • domain-general temporal learning and processing systems
    • phrase-structure grammars (PSG)
      • capture hierarchical syntactic structure
Methods – EEG data collection

• Participants
  • 24 right-handed native English speaker (10 female, mean age 28.0 years)

• Materials
  • 205 sentences from the UCL corpus of reading times
    • 54% with yes/no comprehension questions

• Procedure
  • Word by word presentation
    • Duration: 190+20m ms (m: number of characters in the word)
    • Interval between words: 390 ms
Methods – Estimating word information

• Training Corpus
  • 1.06 million sentences from the British National Corpus (BNC)

• Language Models
  • N-gram models
    • taking the previous n-1 words into account
    • 3 models (n = 2,3,4) & 3 additional models training on full BNC obtained
  • Recurrent neural networks (RNN)
    • 9 training corpora to obtain a range of increasingly accurate models
    • 10 RNN models trained on words and 10 on PoS yielded
  • Phrase-structure grammars (PSG)
    • 9 PSGs defined over words/PoS-strings obtained
Methods – Estimating word information

• Linguistic accuracy
  • Average log-transformed word probability over the experimental sentences
  
  ![Model](image1)
  Fit to ERP amplitudes

  higher linguistic accuracy better

• Entropy Reduction
  • Only RNN used to estimate entropy
  
  ![Model](image2)
  Fit to ERP amplitudes

  higher linguistic accuracy better

  \[ H(W_{t+1...k}) = - \sum_{w_{t+1...k}} P(w_{t+1...k}|w_{1...t}) \log P(w_{t+1...k}|w_{1...t}) \]

  • Chain rule:
  
  \[ P(w_{t+1...k}|w_{1...t}) = \prod_{i=1}^{k} P(w_{t+i}|w_{1...t+i-1}) \]

  • Only RNN computes \( P(w_{t+1}|w_{1...t}) \) over all word types in parallel
Methods – Data analysis

• Linear mixed-effects regression models
  • Variance among subjects and among items is taken into account
  • Include a factor of word surprisal

• Exploratory and confirmatory analyses
  • The current study is mostly exploratory

• Divide the full data set into two subsets
  • Exploratory Data
    • comprising the 12 odd-numbered subjects
  • Confirmatory Data
    • comprising the 12 even-numbered subjects

Generate hypotheses  Test hypotheses
Results

• Exploratory analysis
  • Identified 4 potential effects:
    • Word surprisal – amplitude of N400
    • Word surprisal – amplitude of LAN
    • Word entropy reduction – EPNP
    • Word entropy reduction – PNP

• PoS information measures
  • No potential effects
Results

• Confirmatory analysis
  • Only one survives ...

  • Word surprisal – amplitude of N400
    • Reliable evidence for an effect of word surprisal on the N400

  • Not any other relation between word (or PoS) information and any ERP component.
Results

• Exploratory and confirmatory analyses
  • Investigate whether this effect can indeed be considered an N400
    • Take the full set of data
    • Showed the strongest overall effect on the N400

• The high-surprisal words result in more negative deflection
  • Word surprisal indeed affects N400 amplitude
Results

- Comparing word classes
  - Content words (open-class) vs. function words (closed-class)
    - Function words: no reliable N400 effect
    - Content words: weaker effect
    - Most likely because ...
      - Function words – lower surprisal & elicit a smaller N400
  - In other words ...
    - part of the effect over all words is due to the word class difference
Results

- **Model Comparison**
  - Effects on all words ...
    - The n-gram model
      - explains variance over and above the others
    - The RNN
      - explains variance that the PSG does not account for (not reverse)

<table>
<thead>
<tr>
<th>Words</th>
<th>Model</th>
<th>n-gram</th>
<th>RNN</th>
<th>PSG</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>n-gram</td>
<td>$\chi^2 = 3.25, \ p &lt; .08$</td>
<td>$\chi^2 = 2.44, \ p &gt; .1$</td>
<td>$\chi^2 = 8.65, \ p &lt; .01$</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>$\chi^2 = 1.80, \ p &gt; .15$</td>
<td>$\chi^2 = 2.44, \ p &gt; .1$</td>
<td>$\chi^2 = 4.44, \ p &lt; .04$</td>
</tr>
<tr>
<td>Content</td>
<td>n-gram</td>
<td>$\chi^2 = 6.98, \ p &lt; .01$</td>
<td>$\chi^2 = 2.34, \ p &gt; .1$</td>
<td>$\chi^2 = 3.68, \ p &lt; .06$</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>$\chi^2 = 2.55, \ p &gt; .1$</td>
<td>$\chi^2 = 0.52, \ p &gt; .4$</td>
<td>$\chi^2 = 6.29, \ p &lt; .02$</td>
</tr>
<tr>
<td>Function</td>
<td>n-gram</td>
<td>$\chi^2 = 0.45, \ p &gt; .5$</td>
<td>$\chi^2 = 1.97, \ p &gt; .15$</td>
<td>$\chi^2 = 0.80, \ p &gt; .3$</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>$\chi^2 = 1.65, \ p &gt; .15$</td>
<td>$\chi^2 = 2.59, \ p &gt; .1$</td>
<td>$\chi^2 = 0.22, \ p &gt; .6$</td>
</tr>
</tbody>
</table>
Results

• Model Comparison
  • Effects on content words ...
    • Similar, except that the RNN now outperforms the n-gram model.
  • Effects on function words ...
    • Very weak in general
    • No one model type accounts for variance over and above any other

<table>
<thead>
<tr>
<th>Words</th>
<th>Model</th>
<th>n-gram</th>
<th>RNN</th>
<th>PSG</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>n-gram</td>
<td></td>
<td>$\chi^2 = 6.01, \ p &lt; .02$</td>
<td>$\chi^2 = 8.65, \ p &lt; .01$</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>$\chi^2 = 3.25, \ p &lt; .08$</td>
<td></td>
<td>$\chi^2 = 4.44, \ p &lt; .04$</td>
</tr>
<tr>
<td></td>
<td>PSG</td>
<td>$\chi^2 = 1.80, \ p &gt; .15$</td>
<td>$\chi^2 = 2.44, \ p &gt; .1$</td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>n-gram</td>
<td></td>
<td>$\chi^2 = 2.34, \ p &gt; .1$</td>
<td>$\chi^2 = 3.68, \ p &lt; .06$</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>$\chi^2 = 6.98, \ p &lt; .01$</td>
<td></td>
<td>$\chi^2 = 6.29, \ p &lt; .02$</td>
</tr>
<tr>
<td></td>
<td>PSG</td>
<td>$\chi^2 = 2.55, \ p &gt; .1$</td>
<td>$\chi^2 = 0.52, \ p &gt; .4$</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>n-gram</td>
<td></td>
<td>$\chi^2 = 1.97, \ p &gt; .15$</td>
<td>$\chi^2 = 0.80, \ p &gt; .3$</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>$\chi^2 = 0.45, \ p &gt; .5$</td>
<td></td>
<td>$\chi^2 = 0.22, \ p &gt; .6$</td>
</tr>
<tr>
<td></td>
<td>PSG</td>
<td>$\chi^2 = 1.65, \ p &gt; .15$</td>
<td>$\chi^2 = 2.59, \ p &gt; .1$</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

• Objectives (twofold)
  • Investigate whether ERP amplitudes depend on word & PoS information
    • Expectation: N400 related to word surprisal
      • Proved
  • Identify the model type whose information measures best predict the ERP data
    • The n-gram and RNN models outperformed the PSG in general
Discussion

• The N400 effect of word surprisal
  • Reading a word with higher surprisal value $\rightarrow$ increased N400 amplitude
    • ERP component sensitive to word predictability confirmed

• Across the full range of surprisal values, average N400 amplitudes differed by about 1 $\mu$V
  • Also found with cloze probability
  • Greater when only sentence-final words are varied
  • Most likely because ...
    • Effects are more pronounced on sentence-final words
    • Cloze differences tend to be larger in handcrafted experimental sentences than in naturalistic materials
Discussion

• The N400 effect of word surprisal
  • The strength of the surprisal effect grows nearly monotonically with linguistic accuracy
    • Confounding variable: very unlikely
      • Need to explain not only the effect of surprisal but also the effect of linguistic accuracy

• Previous studies: less predictable function words $\rightarrow$ increased N400 size
  • Not in this study
    • Natural language does not display much variance in function word surprisal
Discussion

• The N400 effect of word surprisal
  • Why surprisal would be predictive of N400 size?
    • Previous studies: two functional interpretations of the N400
      • Semantic integration
        • Increased integration -> larger N400
      • Retrieval of lexical information from memory
        • Retrieval difficulty -> larger N400
  • Memory-retrieval account supported
    • Surprisal estimated by language models
      • minimally sensitive to semantics
Discussion

• Other ERP components and information measures
  • No reliable ERP effects of entropy reduction found, nor for PoS measures
    • Previous studies:
      • Surprisal and ER may not correspond to cognitively distinct processes
        • Alternative quantifications of the same cognitive factor
    • Present study:
      • Only word surprisal showed an effect
        • Two information measures quantify neutrally different processes
• Left anterior negativities
  • ELAN effect
    • Previous studies: elicited by the mismatch between the structural prediction
      • based on the syntactic category of the word currently being processed

• Reasons why an ELAN effect was unlikely to rise
  • An ELAN only appears in cases of outright syntactic violations
    • All the experimental sentences are grammatically correct
  • An ELAN is more often absent than present in experiments that use visually presented sentences
Discussion

• Left anterior negativities
  • LAN effect
    • Previous studies: elicited by a range of syntactic violations beyond the local phrase structure
      • Number / case error
    • Not restricted to syntactic violations
    • LAN effect could have been observed in the data
      • to the extent that syntactic difficulty is captured by word information
Discussion

• Late Positivities
  • (E)PNP effect
    • Previous studies
      • Anterior post-N400 positivity in response to *syntactic disambiguation*
        • much like (E)PNP
      • Entropy reduction – the amount of *ambiguity* resolved by a word or PoS
        • Entropy reduction might predict the (E)PNP
    • Present study
      • Exploratory Analysis
        • A potential (E)PNP effect of word entropy reduction
      • Confirmatory analysis
        • No such effect remained
Discussion

• Late Positivities
  • P600 effect
    • Occurred by a syntactic garden path
      • Triggered by the appearance of a word with unexpected syntactic category
      • Reflect syntactic reanalysis
    • Also found in cases without increased syntactic processing difficulty
      • Alternative interpretations of the P600 effect
        • Syntactic processing plays no central role
        • No reason to expect any effect of information quantities
Discussion

• Implications for models of sentence comprehensions
  • The n-gram and RNN model accounted for variance in N400 size over and above the PSG
    • The more parsimonious models ...
      • do not rely on assumptions specific to language
      • ... outperform the hierarchical grammar based system
  • The assumptions underlying the PSG model ...
    • not efficacious for generating expectations about the upcoming word
Conclusion

• A strong relation between ...
  • the surprisal of a word
  • the amplitude of the N400 component in response to reading that word

• Probabilistic language model can be used to estimate word information values
  • Allowing for a very flexible approach to model evaluation and comparison
    • instrumental in uncovering the representations and processes that underlie human sentence processing