ACL2025読み会@名大（2025.9.26）　論文: https://arxiv.org/abs/2412.11940

ACL 2025

# The Impact of Token Granularity on the Predictive Power of Language Model Surprisal

**Byung-Doh Oh**
Center for Data Science
New York University
oh.b@nyu.edu

**William Schuler**
Department of Linguistics
The Ohio State University
schuler.77@osu.edu

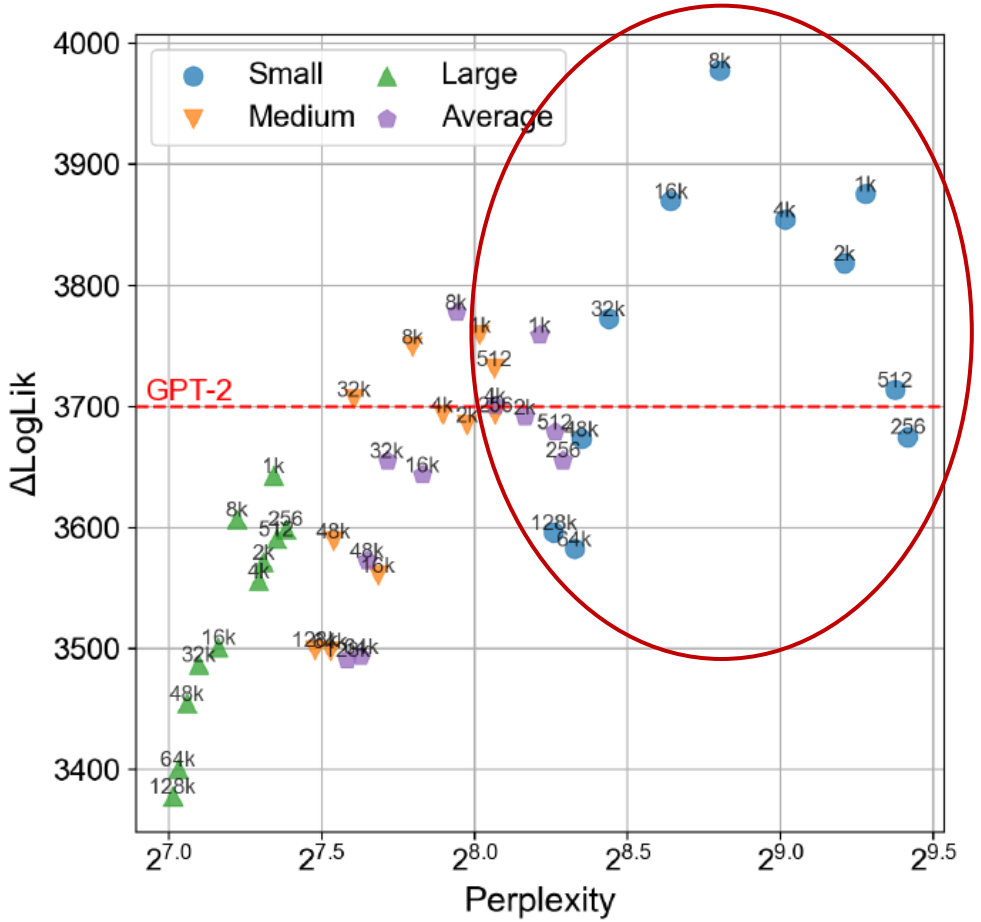
紹介者: 笹野遼平（名大）

一部の図は、論文または
著者の発表スライドより引用

# 論文の概要

- 「言語処理時の認知負荷をサプライザルがどの程度説明できるか」を分析する際の「言語モデルのトークン粒度の影響」を調査

  - 読み時間推定実験においては、Smallモデルでは、中間的粒度（|V|≈8,000）が最も予測精度が高くなるという結果
  - ガーデンパス文を対象とした実験では
    “LMs trained on coarser-grained tokens generally assigned higher surprisal to critical regions, suggesting a greater sensitivity to garden-path effects than previously reported”
    と主張



# 前提知識１：サプライザル理論

- 人間が文を処理するときに経験する「処理の難しさ」を情報理論の枠組みを用いて説明しようとする理論 (Hale 2001, Levy 2008)
- $-\log P(w_t|w_{1\ldots t-1})$で定義されるサプライザルは読み時間と相関

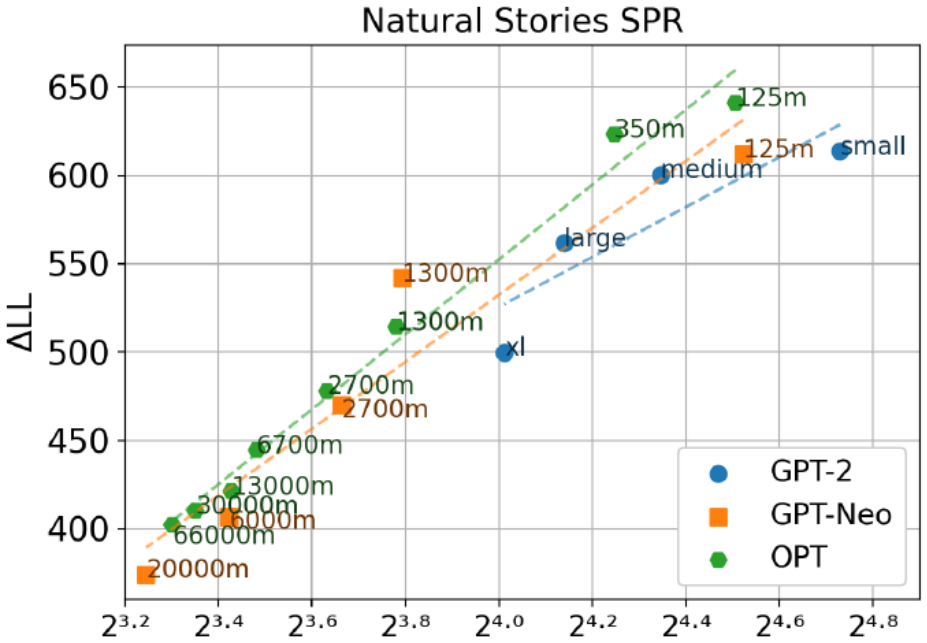
$$\text{Processing difficulty of } w_t \propto \underbrace{-\log P(w_t|w_{1\ldots t-1})}_{\text{サプライザル}}$$

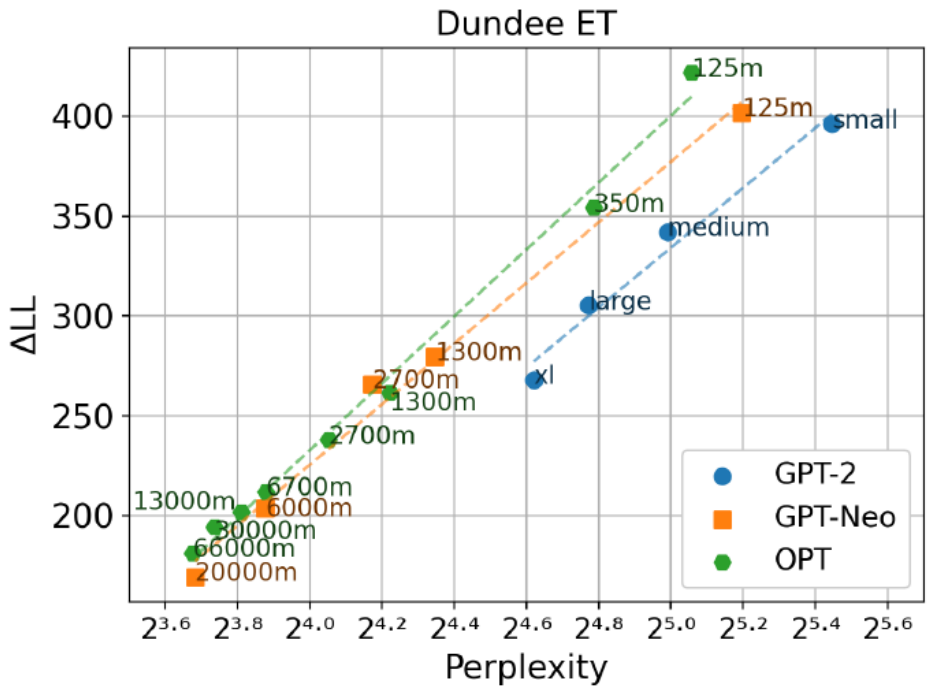
| Word | *If* | *you* | *were* | *to* | *journey* |
|---|---|---|---|---|---|
| Reading Time | 571 ms | 354 ms | 386 ms | 383 ms | 457 ms |
| LM1 Surprisal | 7.76 | 0.81 | 5.42 | 2.09 | 14.62 |
| LM2 Surprisal | 6.71 | 0.78 | 5.22 | 2.30 | 13.93 |
| LM3 Surprisal | 7.10 | 0.56 | 5.15 | 2.39 | 15.02 |

# 前提知識２：LMの性能とサプライザル

- 小さなperplexityが得られるような大規模なTransformer ベースのモデルが読み時間推定において高い性能とならない (Kuribayashi et al. 2021, Oh&Schuler 2023b)

本論文と同一著者




縦軸は読み時間推定への効果ΔLL（↑）、横軸はPerplexity（↓）

(Oh&Schuler 2023b)

# サプライザル計算におけるトークン粒度の影響

Finer granularity, more character-like ($|V| = 256$)

␣ I f ␣ y o u ␣w er e ␣to ␣ j o ur n e y

Coarser granularity, more word-like ($|V| = 128000$)

␣If ␣you ␣were ␣to ␣journey

- サプライザルの値はトークン粒度による影響を大きく受ける可能性
- $|V| = 256$の場合、1 トークンはほぼ 1 文字 ⇒ to と journey の生成確率はおよそ 6 乗分異なる（$p^1 \Leftrightarrow p'^7$ ）
- $|V| = 128\text{k}$の場合、to と journey の生成確率は同等になる

# 実験設定

- トークナイザー
  - トークナイザー: SentencePiece w/ Unigram言語モデル (Kudo 2018)
    - SentencePieceはデフォルトでは空白で必ず単語を分割
  - 語彙サイズ: {256, 512, 1k, 2k, 4k, 8k, 16k, 32k, 48k, 64k, 128k}
  - 訓練データ: English Wiki-40B trainから100万記事
- 言語モデル
  - アーキテクチャ: Mamba-2 (Dao&Gu 2024)（長い系列でも比較的高速）
  - 訓練データ: English Wiki-40B trainから520万記事(～1.5B words)
  - モデルサイズ:
    - Small, Medium, Large
- 語彙サイズ:11 ×
  モデルサイズ:3 = 33種類

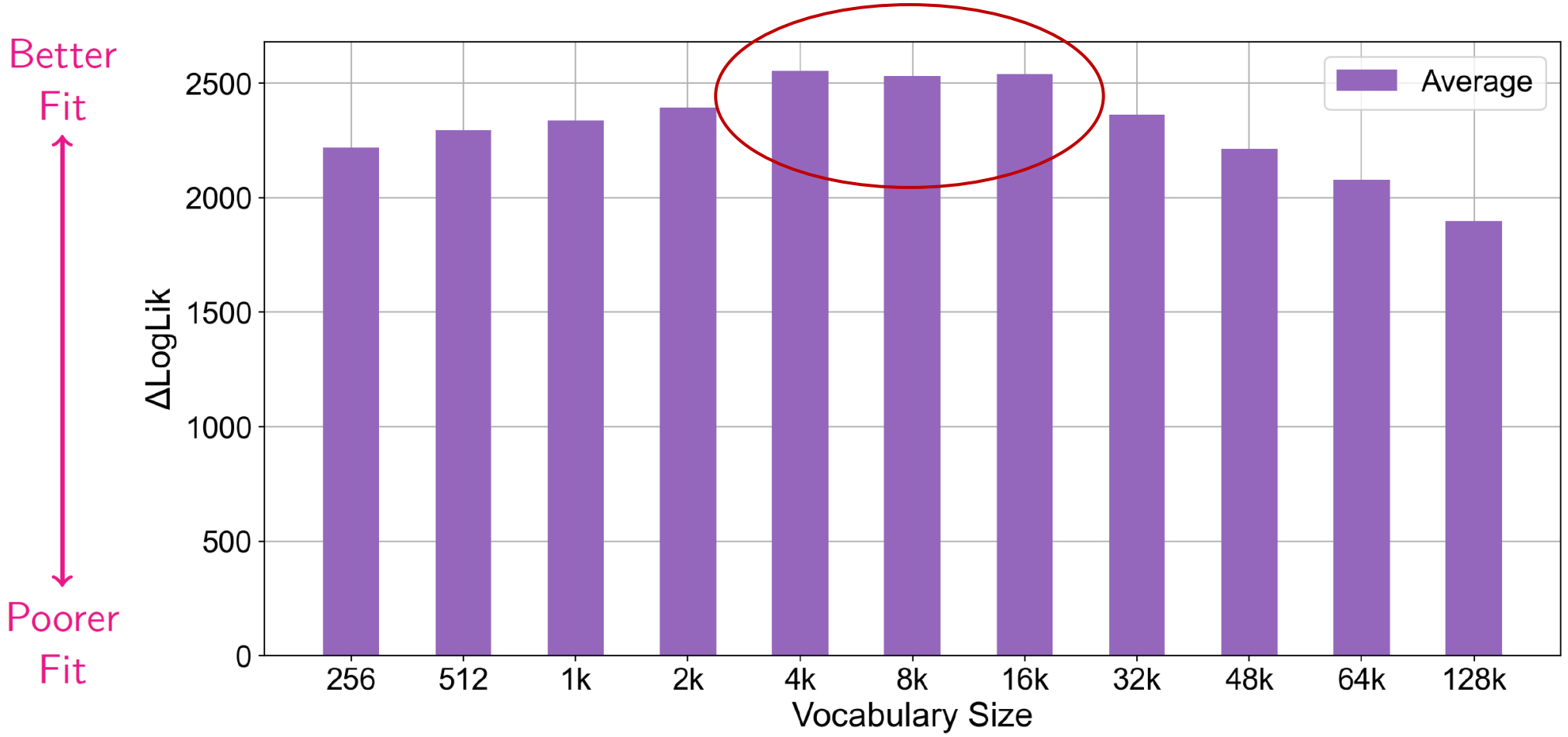
| Model | #L | #H | $d_{\text{model}}$ | #Parameters |
|---|---|---|---|---|
| *Small* | 6 | 8 | 256 | ~2.6M |
| *Medium* | 12 | 16 | 512 | ~19.8M |
| *Large* | 24 | 24 | 768 | ~88.0M |

# 実験１：読み時間データを使った実験

- 手法
  - コーパス５種類: Natural Stories, Brown, GECO, Dundee, Provo
  - 訓練前と訓練後の各言語モデルからサプライザル値を算出
  - 読み時間を線形混合効果(LME)回帰モデルによりモデリング

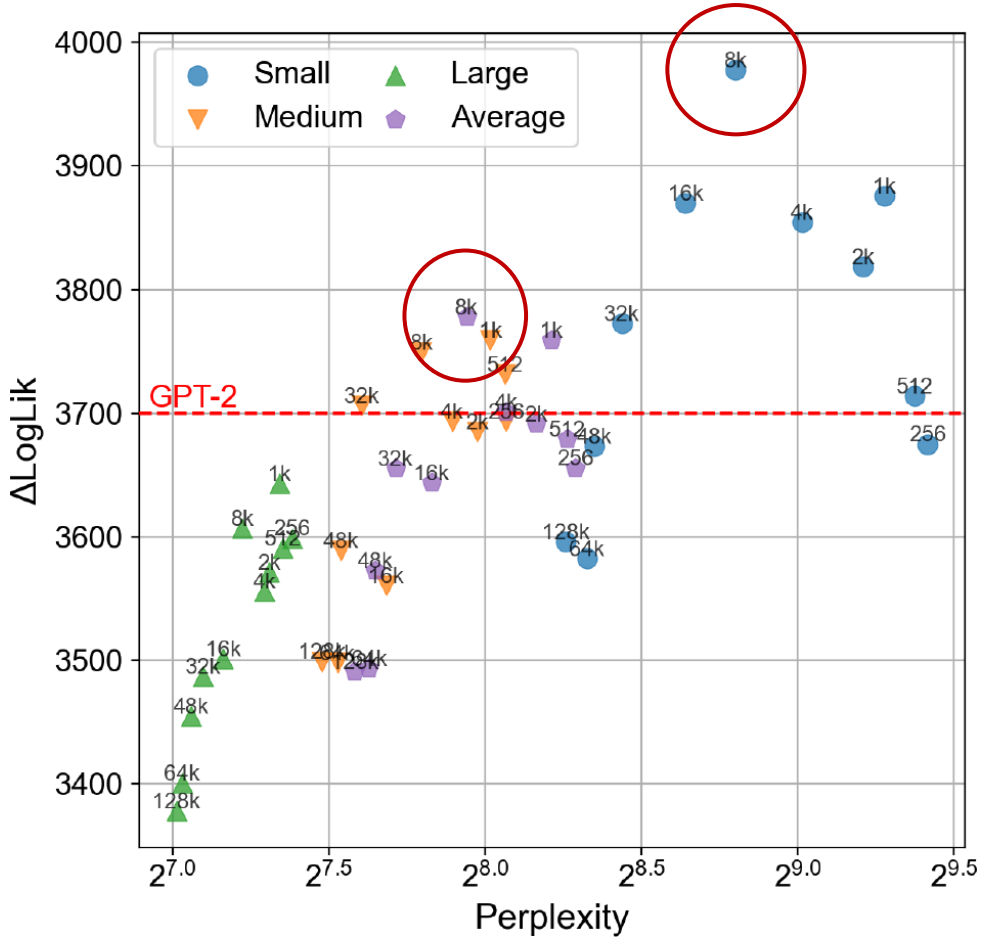
- 評価
  - ベースとする回帰モデルのサプライザル値を、各LMを用いて算出したサプライザル値に置換したときの対数尤度の増加量ΔLogLikで評価
    - ベースモデルでは単語の文字数、その単語の文中の位置などの情報に加えて、KenLMを使って算出したunigramサプライザルを使用
    - LMEモデルはガウス分布に基づくので与えられた読み時間に対する尤度が算出可能

# 結果１：訓練前のモデルを使った実験




※ 言語モデルのサイズによる違いは限定的なのでAverageのみを掲載

# 結果２：訓練後のモデルを使った実験




- Smallモデルにより算出されたサプライザルは訓練前のピークを再現
- 一方、Medium, Largeモデルについてはピークは観測できず
  - Perplexityが小さいほどΔLogLikが小さくなる傾向を再現
- 著者らは３モデルの平均について語彙サイズを 8k とした場合にGPT-2 Smallを超えることを強調しているが平均をとる妥当性は不明

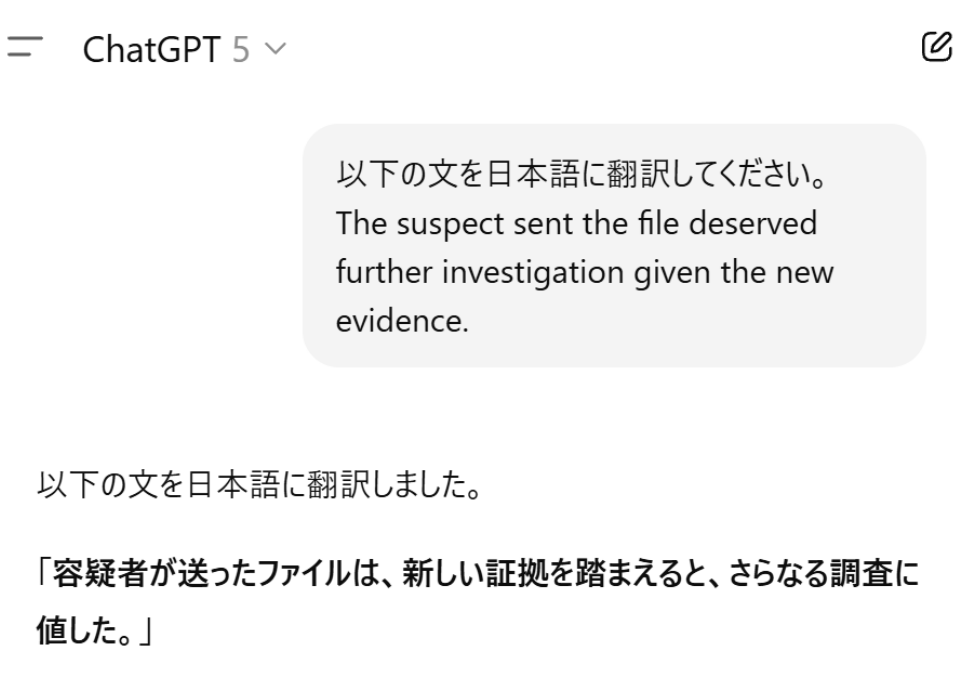
# 前提知識３：ガーデンパス文

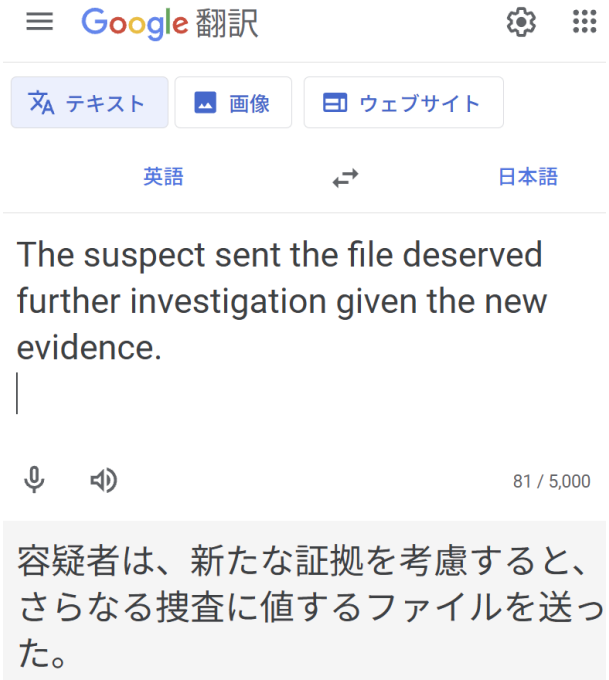
Garden path effect (GPE)

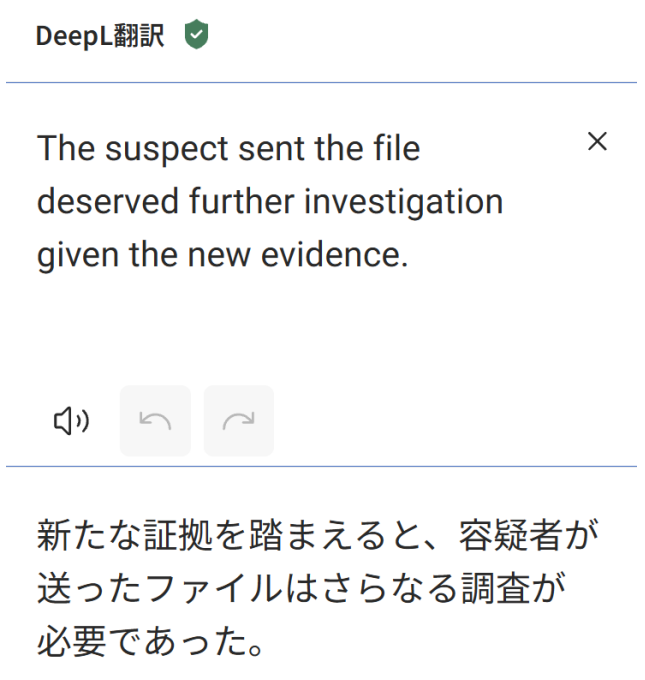
“who was”がないとdeservedを読む時間が大幅に遅くなる

The suspect sent the file deserved further investigation given the new evidence.

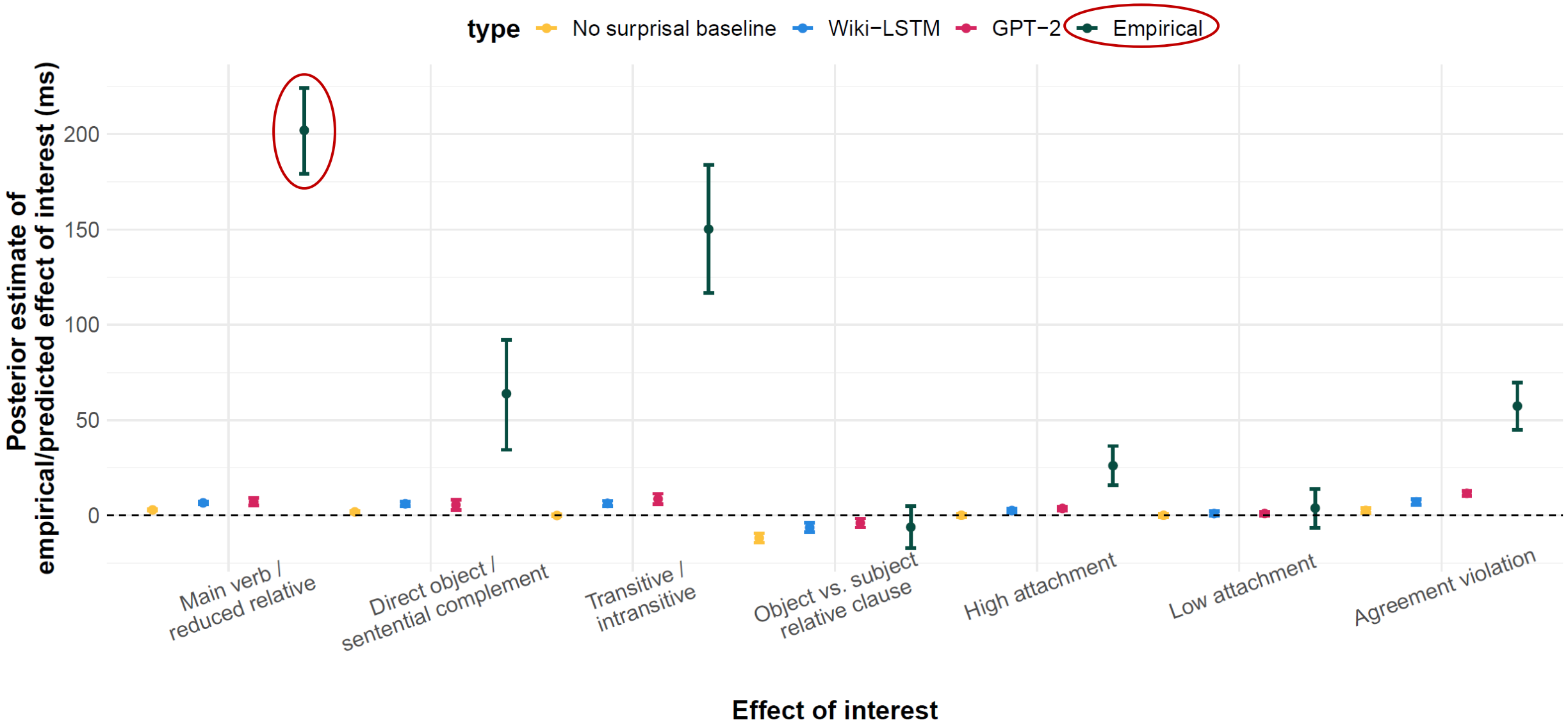
who was



# 前提知識4：人間の読み時間への影響 (Huang et al. 2024)

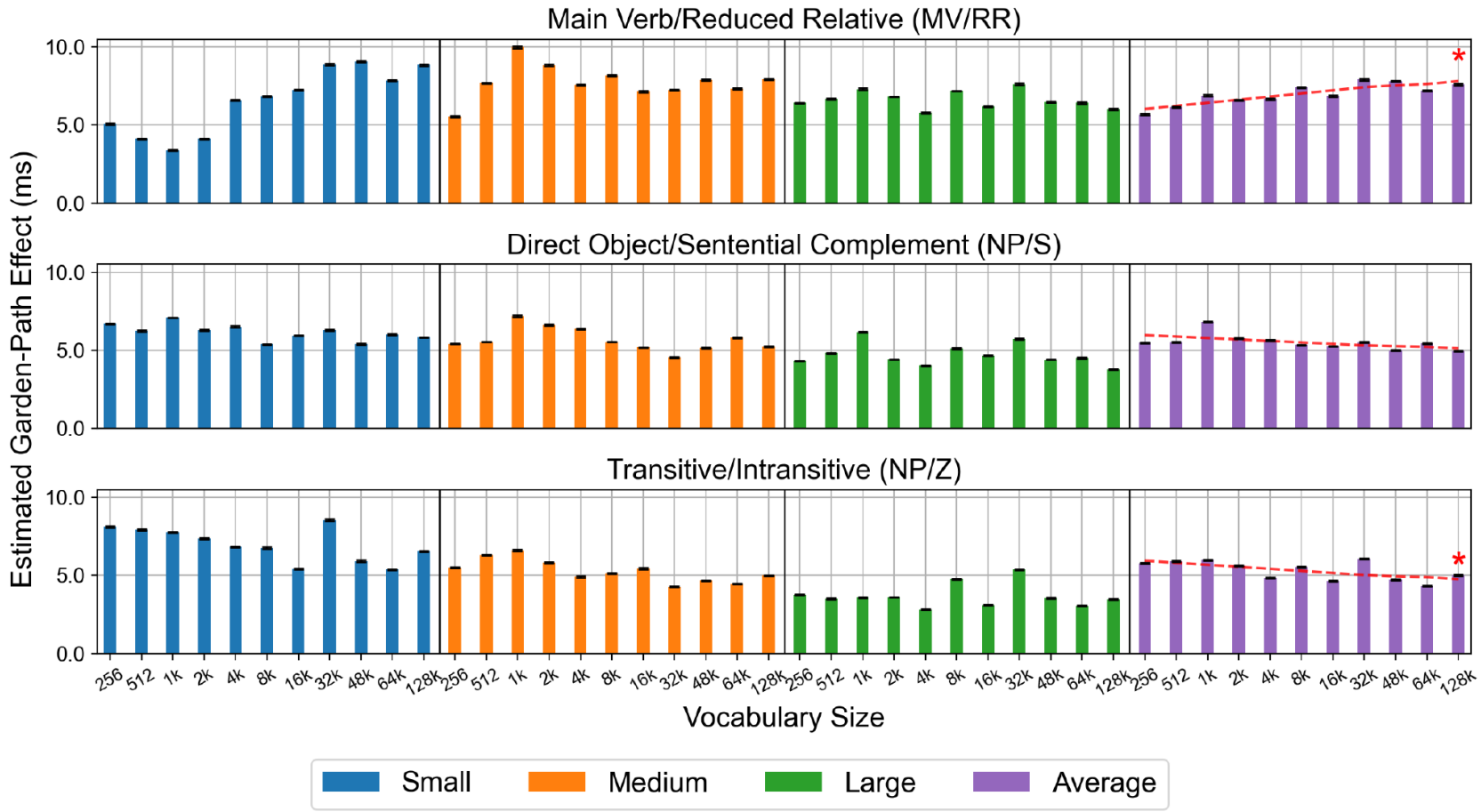


# 実験２：ガーデンパス文を使った実験

| Construction/Condition | Example |
|---|---|
| MV/RR Ambiguous | The suspect sent the file *deserved further investigation* given the new evidence. |
| MV/RR Unambiguous | The suspect who was sent the file *deserved further investigation* given the new evidence. |
| NP/S Ambiguous | The suspect showed the file *deserved further investigation* during the murder trial. |
| NP/S Unambiguous | The suspect showed that the file *deserved further investigation* during the murder trial. |
| NP/Z Ambiguous | Because the suspect changed the file *deserved further investigation* during the jury discussions. |
| NP/Z Unambiguous | Because the suspect changed, the file *deserved further investigation* during the jury discussions. |

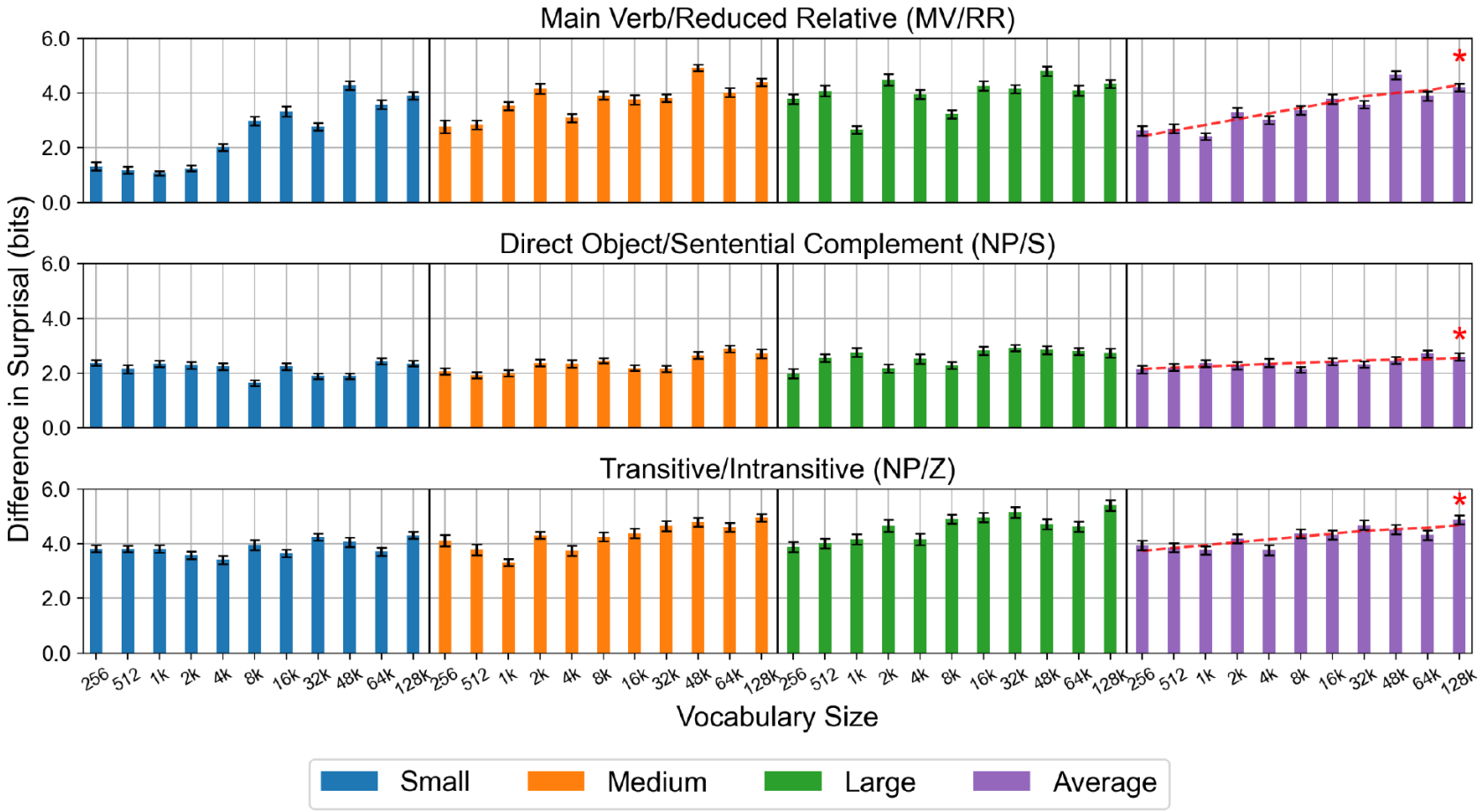
- ３タイプ、計２４ペアのデータを使用（Huang et al. 2024）
- 構文的曖昧性による処理困難さを持たない文を対象にself-paced reading (SPR)にfitするようにLMEモデルを調整
- “*critical word*”と後続する２語(spillover words)におけるサプライザルの増加による読み時間の予測値の増加量を算出

# 1つ目のspillover wordのGPE予測値




- “no clear trend in estimated garden-path effects” (著者スライドより)
- 人におけるGPEと比べ1～2桁小さい効果しか観測できない問題は改善されず

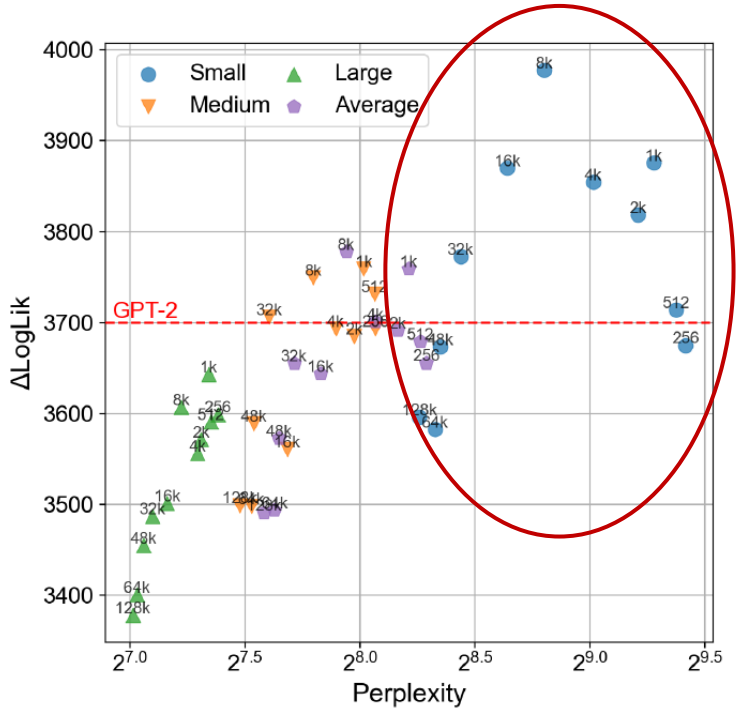
# critical wordにおけるサプライザルの増加




- “Coarser-grained tokens lead to larger differences in raw surprisal” (著者スライドより)
- この結果から何らかの結論を得るのは危険な印象

# 雑感

- サプライザル算出におけるトークン粒度は重要な問題
- 読み時間の実験において、もともと性能が高いSmallモデルに対し中間的なトークン粒度でΔLogLikが最大となることは有用な知見
  - 今後はSmall × 8k で実験すると良い？ 一方で意外性は限定的？
- ガーデンパス(GP)文の実験から何らかの知見を主張するのは厳しそう
  - 著者の主張も一貫していない
  - そもそもLLMはGP文を解析できていない？
- なぜmain (oral)に採択されたのか？
  - ドメイン知識を前提とし説明がやや不親切な印象
  - Mamba2を採用するなど技術的な面が評価？
  - 心理言語学分野の査読者の興味に合致？



# 参考文献


- John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. NAACL
- Levy. 2008. Expectation-based syntactic comprehension. Cognition
- Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. ACL-IJCNLP
- Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? TACL
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. ACL
- Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. ICML
- Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. Journal of Memory and Language