Data Issues In

Cognitive Studies of Syntactic Processing and Language Comprehension

By:
Rajaswa Patil
(Cognitive Neuroscience Lab, BITS Goa)

Introduction: Overview

- Language Production
  - **Multiple ways available** to express the same message!
  - What are the available **cognitive resources**?
  - What are the **environmental settings and constraints**?
  - How does a speaker **optimize communication**?

- Language Comprehension
  - **Incremental Processing** of linguistic content
  - **Predictability** of semantic, syntactic and pragmatic structures
  - **Language Proficiency** of the reader/listener
  - How do we model and monitor the process of **L1/L2 Language Acquisition?**

Introduction: Syntactic Processing

- Directly related to **language production:**
  - **Syntactic Reduction:** *"How big is the family <u>that</u> you cook for?"* [1]
  - **L1 vs L2:** Syntactic Complexity in produced language [2]
  - .....
- Directly related to **language comprehension:**
  - **Garden Path Disambiguation:** "*The old <u>man</u> the boat.*" [3]
  - **Syntactic Priming:** Contributes significantly to language comprehension [4]
  - .....

1. Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06). MIT Press, Cambridge, MA, USA, 849–856.
2. Kuiken, F, Vedder, I. Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. Int J Appl Linguist. 2019; 29: 192– 210. https://doi.org/10.1111/ijal.12256
3. Osterhout, L., McLaughlin, J., Kim, A., Greenwald, R., & Inoue, K. (2004). Sentences in the brain: Event-related potentials as real-time reflects of sentence comprehension and language learning. In M. Carreiras & C. Clifton, Jr. (Eds.), The on-line study of sentence comprehension: Eyetracking ERP and beyond. Psychology Press.
4. Traxler, M. J., Tooley, K. M., & Pickering, M. J. (2014). Syntactic priming during sentence comprehension: Evidence for the lexical boost. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(4), 905–918. https://doi.org/10.1037/a0036377

Data-driven Psycholinguistic Studies: Corpus-based Methods

The *flourishing* paradigm of ***Computational Psycholinguistics***:

- Availability of large-scale corpuses
  - Available metadata and interactive content
  - Public efforts to develop, curate and clean corpuses
  - Diversity (Ex: Multilingual data)
- Advances in computational language modeling
  - Distributed learning models
  - Scalable computation
- Tools from Information Theory
  - Allows large scale studies of language complexity and sophistication
  - Information Density metrics: Surprisal and Perplexity

Data-driven Psycholinguistic Studies: The Surprisal Metric

$$surprisal(w_i) = -\log(P(w_i | w_{i-1,i-2,\ldots n}))$$

- **Surprisal** is the Shannon Information of a lexical-unit derived from its conditional probability with respect to a given lexical-context
- The probabilities are obtained from computational language models (usually parameterized)
  - **Lexical Surprisal:** n-gram models, SRNs, LSTMs, Transformers, etc.
  - **Syntactic Surprisal:** PSG parser, Arc-eager parser, PCFG parsers, etc.
- Surprisal can be used as metric of sophistication, predictability and uncertainty

Data-driven Psycholinguistic Studies: The Surprisal Metric

- **Language Production:**
  - **Syntactic Reduction:** Uniform Information Density (UID) [1]
  - **L1 vs L2:** Helps in predicting non-native language proficiency [2]
  - ……..
- **Language Comprehension:**
  - **Garden Path Disambiguation:** Helps in modelling syntactic disambiguities [3]
  - **Syntactic Priming:** Helps in modelling the error signal associated with priming [4]
  - ……..

1. Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06). MIT Press, Cambridge, MA, USA, 849–856.
2. Yevgeni Berzak, Boris Katz, Roger Levy Assessing Language Proficiency from Eye Movements in Reading Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Pages 1986--1996
3. Schijndel, M.V., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. Cognitive Science.
4. Fine, A.B. and Florian Jaeger, T. (2013), Evidence for Implicit Learning in Syntactic Comprehension. Cogn Sci, 37: 578-591. https://doi.org/10.1111/cogs.12022

Data-driven Psycholinguistic Studies: ERPs & Eye-tracking

- **ERPs from EEG data** and **gaze-analysis from Eye-tracking data** are quite insightful in terms of the underlying neural correlates and physiological responses of syntactic processing and language comprehension related tasks
- Bridging the gaps between surprisal and ERPs & Eye-tracking:
  - **ERPs:** N400 amplitude from reading tasks is proportional to the lexical surprisal [1]
  - **Eye-tracking:** Surprisal follows a logarithmic relationship with the reading times [2]
  - **Spoken-word duration:** Syntactic surprisal is a good predictor of the spoken word durations in conversational settings [3]

1. Frank, Stefan L., et al. "Word surprisal predicts N400 amplitude during reading." (2013).
2. Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302-319.
3. Demberg, V., Sayeed, A., Gorinski, P., & Engonopoulos, N. (2012, July). Syntactic surprisal affects spoken word duration in conversational contexts. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 356-367).

Issues: Corpus-based Methods

- Cannot directly infer physiological responses and neural correlates & architectures of syntactic processing and language comprehension
- Methods are mostly black-boxed in nature, with very less interpretability being offered
- Data and models are prone to biases

Overall, the issues & limitations with corpus-based methods are quite few due to recent advancements in data science and data mining. But the issues are quite significant from the perspective of cognitive studies.

Issues: ERPs & Eye-tracking

- Due to the clinical nature of data, the logistics of data collection are challenging
- Lack of large-scale datasets
- Lack of diversity in data (hence prone to subjectivity, artifacts, etc.)
- Quality of data: High signal-to-noise ratio (SNR)
- Lack of early open-access sharing of data
- .....

Even though there have been some public initiatives to address these issues, it seems improbable to catch-up with the progress being made towards corpus-based methods

Probable Solutions: Bridging the Gaps

Is it possible to **accelerate the process of tackling issues** in ERP and Eye-tracking data for syntactic processing and language comprehension through simulated experiments?

- **Simulations**:
  - Progress in developing Cognitive Models of Syntax and Language Comprehension
  - Extracting relevant metrics (Ex: surprisal)
- **Mapping Metrics**:
  - Develop methods to map the metrics obtained from simulated cognitive models to clinical data like ERP and Eye-tracking
  - Obtain metrics by feeding stimulus to the cognitive models
  - Study the mapped ERP and Eye-tracking data for underlying physiological responses and neural correlates

Thoughts and Questions?

CONTACT:

- **E-mail:** f20170334@goa.bits-pilani.ac.in
- **Website:** https://rajaswa.github.io/