**Decomposing Semantic Decomposition Reveals Origin of Thematic Roles in LLMs**

*Lucas Y. Li, Zander Lynch, Marten van Schijndel (Cornell University)*

In traditional theories of event semantics, predicates project thematic relations onto their arguments [1], entailing Dowty's proto-role properties [2]. The alternative Neo-Davidsonian framework argues for thematic separation, whereby semantic roles may be introduced by non-predicate event elements such as NPs [3, 4]. Previous psycholinguistic research argues for separation using an incremental processing model, as NPs appearing before the verb must assign themselves thematic roles [4-6]. However, the influence of post-verbal NPs is not yet known. While existing studies rely on controlled stimuli, a corpus analysis using computational methods allows the comparison of event elements in a large dataset of naturalistic sentences.

In this work, we probe LLM event representations using the Semantic Proto-Role Labeling (SPRL) task on the English Universal Decompositional Semantics dataset (Decomp) [7, 8]. While we do not assert that LLMs model human cognition, they learn *statistical regularities* over their training corpora and thus can reveal what information is important for accurate thematic interpretation. To evaluate the effect of incremental processing, we experiment with two Transformer LLMs [9] of the same size, RoBERTa-Large [10] (bidirectional) and GPT2-Medium [11] (incremental). As >96% of sentences in Decomp are active, we generate additional passive sentences by passivizing verbs and switching arguments to preserve semantic roles but balance syntactic positions. We fine-tune the LLMs for proto-role property prediction on the augmented and combined Decomp V1 and V2 datasets. Evaluation shows both RoBERTa and GPT-2 meet or exceed average human inter-annotator agreement on all properties (**Table 1**).

To quantify the degree to which the models use NPs and verbs during SPRL, we implement Generalized Contextual Decomposition (GCD), an interpretability method originally proposed for LSTMs [12, 13], which we adapt for attention in Transformers. Given an input sequence, GCD applies masking to partition the tokens into in-focus ($\beta$) and out-of-focus ($\gamma$) components, which are propagated through all layers of the model. For any layer *i*, its hidden state representation $z_i$ is linearly decomposed into $z_i = \beta_i^z + \gamma_i^z + \delta_i^z$, where $\beta_i^z$ contains information received from the in-focus input, $\gamma_i^z$ contains information from the out-of-focus input, and $\delta_i^z$ contains information from bias parameters of layers $\leq i$. Thus, GCD reveals how much, and whether positively or negatively, each component contributes to each classifier logit[1] in SPRL. **Fig. 1** shows an example decomposition of the property *instigation*; GPT-2 correctly identifies *dogs* (being non-human; 1a) and *followed* entities (1b) as typically non-instigative and assigns them negative contributions, while modifiers *police* (1a), *tirelessly* (1a), and *ruthless* (1b) increase instigation.

**Fig. 2:** For proto-agent properties, NPs rarely contribute negative evidence, and only contribute positive evidence in active subjects or passive objects (2a, c). NPs and verbs contribute equally to agentiveness in RoBERTa (2a), while GPT-2 disproportionately identifies agentiveness using NPs (2c). Surprisingly, NP contributions are consistent regardless of passivization (eg. pre-verb active subjects vs post-verb passive objects), even for incremental GPT-2. For proto-patient properties, NPs and verbs contribute largely negatively in RoBERTa (2b), but mostly positively in GPT-2 (2d). Verbs contribute more to GPT-2's interpretation of proto-patients (2d) than proto-agents (2c), even in passive sentences, indicating agentiveness primarily originates from the argument, while patienthood primarily originates from the verb but still uses information from the entity. This provides empirical evidence for the Neo-Davidsonian theoretical analysis of Kratzer [14], who argues for severing agents, but not patients, from verbs. Additionally, while GPT-2 identifies positive evidence for both proto-agent and proto-patient properties, RoBERTa focuses mostly on positive evidence of agentiveness. In conclusion, we find evidence for Neo-Davidsonian separation in LLMs beyond what is explainable by incremental processing, with further study of humans needed. We call for a re-examination of traditional verb-first event semantics in favor of a representation leveraging all event elements.

---

[1]As opposed to attention *weights*, which reveal what information a model has access to.

| Property | Humans | | RoBERTa | | GPT-2 | |
|---|---|---|---|---|---|---|
| | Act | Pass | Act | Pass | Act | Pass |
| awareness | 89.7 | 88.9 | **91.7** | 89.7 | 91.2 | **90.5** |
| Δ_location | 77.9 | 77.8 | 80.0 | 77.4 | **81.8** | **79.0** |
| Δ_state | 65.0 | **68.3** | 67.2 | 65.5 | **69.3** | 68.3 |
| Δ_possession | 91.3 | 90.1 | **92.8** | **91.1** | 92.5 | 90.7 |
| existed_after | 85.5 | 86.9 | **87.5** | **88.3** | 87.2 | 87.5 |
| existed_before | 85.0 | 85.7 | **86.7** | **86.9** | 86.4 | 84.1 |
| existed_during | 96.1 | 96.4 | **97.2** | **97.4** | 96.7 | 97.8 |
| instigation | 72.4 | 69.0 | **75.7** | **71.4** | 73.0 | 68.3 |
| sentient | 89.5 | 88.5 | 92.1 | 90.3 | **92.6** | **91.5** |
| volition | 86.8 | 88.1 | **88.2** | **88.5** | 86.3 | 87.7 |
| average | 83.9 | 84.0 | **85.9** | **84.6** | 85.7 | 84.5 |

**Table 1**: SPRL inter-annotator and model F1 scores (n=639 unseen sentences). Act denotes active sentences and Pass passive.

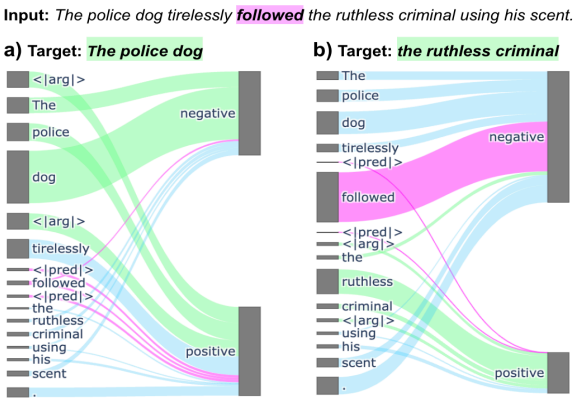


**Fig 1**: Example SPRL GCD GPT-2 contributions for the proto-agent property *instigation*.

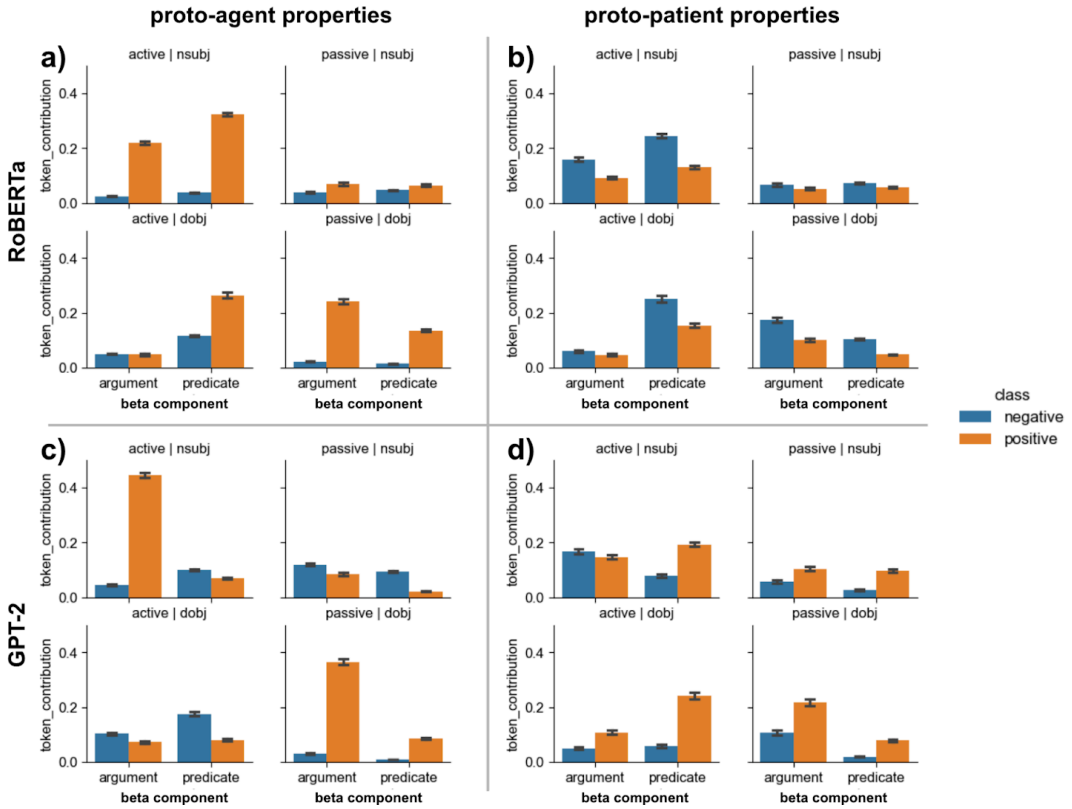


**Fig 2**: SPRL GCD token contributions of predicates (verbs) and target arguments (NPs) in the Decomp test set (n=1981 unseen sentences). Token contribution is defined as $\beta^z/(|\beta|\ (\beta^z + \gamma^z))$, where $|\beta|$ denotes the in-focus component's token length. Error bars show standard error.

## References


[1] Chomsky. 1981.
[2] Dowty. *Lang.*, 1991.
[3] Parsons. 1990.
[4] Husband. *Lang. Ling. Compass.*, 2023.
[5] Bornkessel et al. *Lang. Cogn. Proc.* 2003.
[6] Chow et al. *Lang. Cogn. Neurosci.*, 2018.
[7] Reisinger et al. *TACL*, 2015.
[8] White et al., *EMNLP*, 2016.
[9] Vaswani et al. *NeurIPS*, 2017.
[10] Liu et al. *arXiv preprint*, 2019
[11] Radford et al. *OpenAI Blog*, 2019.
[12] Murdoch et al. *ICLR*, 2018.
[13] Jumulet et al. *CoNLL*, 2019.
[14] Kratzer. 1996.