TO WHAT EXTENT CAN THE HUMAN AGENT BIAS BE CAPTURED BY LINGUISTIC EXPERIENCE? EVIDENCE FROM MACHINE-LEARNING OF CORPORA

Eva Huber[*1,2], Paola Merlo[3], and Balthasar Bickel[1,2]

[*]Corresponding Author: eva.huber@uzh.ch
[1]Department of Comparative Language Science, University of Zurich.
[2]Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich.
[3]Département de Linguistique, University of Geneva.

Converging evidence from event apprehension (Wilson, Papafraou, Bunger, & Trueswell, 2011), language processing (Ferreira, 2003; Bornkessel-Schlesewsky & Schlesewsky, 2009; Kemmerer, 2012) and default grammar rules (Dryer, 2013; Napoli & Sutton-Spence, 2014) shows that humans have a bias towards the Agent participant in events (*Agent bias*). We contrast two hypotheses on its evolution: (i) The *biologically-driven hypothesis* ascribes the bias to an innate property that is already present in prelinguistic infants (Durrant et al., 2021) and that is potentially shared with other animals (Mascalzoni, Regolin, & Vallortigara, 2010); in this case, the effects on language stem directly from an innate principle. (ii) The *experience-driven hypothesis* ascribes the bias to our experience with the distributional properties in language use; in this case, the bias in processing is acquired, and the parallels between cognition and language reflect parallel evolution. To probe these hypotheses we focus on the specific effect that Agent Bias has on the processing of role-ambiguous noun phrases. Here the effect reveals itself in a transient prediction of an agent role which triggers an electrophysiological deflection, typically an N400 effect, when the prediction fails later in the sentence (Bornkessel-Schlesewsky & Schlesewsky, 2014). The experience-driven hypothesis suggests that this effect can be simulated with computational models whose knowledge of language is derived from distributional patterns in the linguistic input and general architectural assumptions but without any explicit agent bias built into them. Previous work on English (Levy, 2008; Frank, Otten, Galli, & Vigliocco, 2015; Szewczyk & Federmeier, 2022) has shown that N400 effects correlate with *surprisal*, the negative log probability of a word given the previous input extracted from such models. We use previously conducted EEG experiments which contain an initial ambiguous NP to compare the extent to which suprisal values in these models can capture N400 effects. Under the experience-driven hypothesis, we expect higher surprisal values in the conditions that also trigger an

N400 effect. Under the biologically-driven hypothesis, we expect higher surprisal values not to correlate with the conditions that raise an N400 effect.

We selected experimental data from three languages that are maximally distinct in how they code Agents: Hindi (Bickel, Witzlack-Makarevich, Choudhary, Schlesewsky, & Bornkessel-Schlesewsky, 2015), which flags Agents by an "ergative" case maker only under specific conditions (perfective transitive clauses); Basque (Isasi-Isasmendi Andaluze et al., 2022), which flags all Agents as ergative; and German (Haupt, Schlesewsky, Roehm, Friederici, & Bornkessel-Schlesewsky, 2008), which does not flag Agents by case but treats them as general subjects. All experiments showed that a sentence-initial noun phrase disambiguated to a Patient reading elicits an N400 response, although experimental conditions varied (Hindi: role-ambigous vs unambigous Patients $\times$ aspect; Basque: ambiguous Agents vs Patients $\times$ unambigous Agents vs Patients; German: ambigous Agents vs Patients $\times$ dative vs accusative-assigning verbs).

We extracted the surprisal values of two types of language model architectures: a Long Short-Term (LSTM) model (Hochreiter & Schmidhuber, 1997), a variant of a recurrent neural network, and a Transformer model (Vaswani et al., 2017). LSTMs process units incrementally by recurrence, and hence contain a memory bottleneck. Transformers, on the contrary, have 'attention layers' which allow for direct access to parts of the previous input (Vaswani et al., 2017). Hierarchical Bayesian Models are fitted to estimate surprisal values given the same conditions as in the experiments. We then qualitatively evaluate whether the surprisal values show the same trends as the N400, in line with the experience-driven hypothesis.

The results (see figure in supplementary material) for the Hindi LSTM show a slight trend in line with the N400 findings, but estimates overlap between conditions (89% CIs imperfective role-ambiguous: [13.29,14.55] vs role-marked: [12.9,14.12], perfective amb.: [11.57,12.74] vs marked: [11.24,12.31]); the Transformer model replicates the N400 findings slightly better in the imperfective aspect (role-marked [16.53,17.88] vs role-unmarked [15.74,17.02]), unlike in the experiments, where it held across aspects. Both Basque models replicates the N400 findings (LSTM A: [12.55,13.61 vs P: [13.63,14.47]; Transformer A: [16.18,16.26] vs P: [16.24,16.32]) but, unlike in the experiments, they generalize the pattern to unambigous cases (LSTM A: [12.71,13.78] vs P: [13.44,14.29]; Transformer: [16.18,16.26] vs P: [16.23,16.31]), where no N400 was observed. The German models replicate the N400 findings but the effect is strong and reliable only in the Transformer model (for acc-verbs: A: [3.31, 4.1] vs P: [5.73,6.48]; LSTM: A: [8.96,9.76] vs P: [9.67,10.29]).

We conclude that there is at best marginal evidence for the experience-driven hypothesis. The surprisal values replicate N400 effects only in German Transformer models. In all other cases, surprisal values are either more general (Basque) or less general (Hindi) than N400 findings; in addition most effects have overlapping credibility intervals.

Acknowledgments

This work was funded by the NCCR Evolving Language supported by Swiss National Science Foundation Agreement (#51NF40_180888).

References

Bickel, B., Witzlack-Makarevich, A., Choudhary, K. K., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2015). The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLOS ONE*, *10*, e0132819.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). Minimality as vacuous distinctness: Evidence from cross-linguistic sentence comprehension. *Lingua*, *119*(10), 1541–1559.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2014). Competition in argument interpretation: Evidence from the neurobiology of language. *Competing motivations in grammar and usage*, *107*, 126.

Dryer, M. S. (2013). Order of subject, object and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Durrant, S., Jessop, A., Chang, F., Bidgood, A., Peter, M. S., Pine, J. M., & Rowland, C. F. (2021). Does the understanding of complex dynamic events at 10 months predict vocabulary development? *Language and Cognition*, *13*(1), 66–98.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive psychology*, *47*(2), 164–203.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Haupt, F. S., Schlesewsky, M., Roehm, D., Friederici, A. D., & Bornkessel-Schlesewsky, I. (2008). The status of subject–object reanalyses in the language comprehension architecture. *Journal of Memory and Language*, *59*(1), 54–96.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Isasi-Isasmendi Andaluze, A., Sauppe, S., Andrews, C., Laka, I., Meyer, M., & Bickel, B. (2022). Semantic role parsing in intransitives: EEG insights from basque. *34th Annual CUNY Conference on Human Sentence Processing*.

Kemmerer, D. (2012). The cross-linguistic prevalence of sov and svo word orders reflects the sequential and hierarchical representation of action in broca's area. *Language and Linguistics Compass*, *6*, 50-66.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Mascalzoni, E., Regolin, L., & Vallortigara, G. (2010). Innate sensitivity for self-propelled causal agency in newly hatched chicks. *Proceedings of the National Academy of Sciences*, *107*(9), 4483–4485.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: implications for all language. *Frontiers in psychology*, *5*.

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wilson, F., Papafraou, A., Bunger, A., & Trueswell, J. (2011). Rapid extraction of event participants in caused motion events. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).