Plausibility based comprehension in a neural network model of sentence processing

Most theories of sentence processing assume that distinct processes such as lexical activation and syntactic parsing unfold sequentially with their respective output being detailed and complete. During the past decade, a more nuanced view has been proposed in psycholinguistics following behavioral and electrophysiological experiments. Ferreira et al. (2003) [1] asked human participants to indicate the agent or the patient of the event described by normal active sentences (e.g., "The dog bit the man"), role reversed active sentences (e.g., "The man bit the dog") and their passive versions. Role reversed sentences, often called reversal anomalies (RA) are sentences that are syntactically correct but semantically anomalous because their agent and patient fillers are swapped. In these sentences, the thematic-role assignment (e.g., "man" as *agent* and "dog" as *patient*) violates the expectations imposed by the event semantics. Ferreira's results showed that participants frequently misinterpret passive RA sentences (e.g., "The dog was bitten by the man"). In consequence, Ferreira proposed the *good enough* approach to language comprehension, which assumes that people might sometimes use processing heuristics based on their expectations about events to figure out who is doing what to whom rather than relying on syntactic rules. Related to this, studies conducted by Kuperberg et al. (2003) [4] and Kim & Osterhout (2005) [3] show evidence that RA sentences, despite their semantic abnormality, elicit only a small increase in N400 amplitude compared to normal control sentences, which is surprising because amplitudes of the N400 brain potential are typically increased in semantically anomalous sentences (see [5] for review). These observations were explained as the results of *semantic illusion* according to which the syntax-cued thematic-role assignment is - at least temporarily - overrun by expectations regarding the event semantics [8], hence the small N400 amplitude.

In this study, we investigated whether the **Sentence Gestalt (SG) model**, a connectionist model of language comprehension trained on a large scale corpus, can account for the pattern of behavior elicited by RA sentences (active and passive), based on stimuli such as those used by Kuperberg et al. (2003) and Ferreira et al. (2003). The SG model maps sentences to a representation of the described event approximated by a list of role-filler pairs representing the action and its various participants such as agent, patient, and eventual modifiers (for more details see [7, 6]). The model processes the linguistic input without any inbuilt knowledge of syntactic rules. Instead it learns based on the statistical regularities of its environment to map the linguistic input to the corresponding event representation. In our experiment, we presented the SG model with 360 sentences belonging to 4 matched conditions (2 active and 2 passive, with 90 sentences per condition). Conditions consist of control (C) and reversal anomaly (RA), both active and passive. RA sentences were generated starting from each C sentence. A RA sentence is obtained by reversing the agent and patient fillers of a C sentence. So, for instance, C sentence "After decades in the jungle the research identified the species" is matched by RA "After decades in the jungle the species identified the research".

After feeding the SG model with a whole sentence, the model is tested whether it correctly recognises the semantic roles and the fillers of the sentence's arguments (See Figure 1). **Role accuracy** is assessed by providing probes containing only the filler (as a word embedding) of the agent or patient arguments – the model is expected to assign roles to the provided fillers (Figure 1, left). Conversely, **filler accuracy** is assessed by feeding probes containing only the role of the arguments – the model is expected to provide the fillers to the roles (Figure 1, right). Table 1 shows **role accuracy** across conditions and voices. There is a significant main effect of condition, with significantly higher accuracies for C as compared to RA sentences ($F(1, 32) = 3212.0$, $p < 0.001$) and a main effect of voice, with significant higher accuracies for active as compared to passive sentences ($F(1, 32) = 113.5$, $p < 0.001$). There also was a statistically significant interaction between condition and voice in the average role accuracies of the SG models ($F(1, 32) = 299.7$, $p < 0.001$). In the RA condition, the SG models shows strong tendency to misinterpret *agents* as *patients* 88.27% of times for active and 81.23% of the times for passives. The rate of misinterpretation of *patients* as *agents* is lower, yet still significantly higher than in C sentences. Table 2 shows the **filler accuracy** across condition and voices. It shows a significant main effect of condition, with significantly higher accuracies for C as compared to RA sentences ($F(1, 32) = 815.60$, $p < 0.001$) and a main effect of voice, with significantly higher accuracies for active as compared to passive sentences ($F(1, 32) = 18.75$, $p < 0.001$)). There also was a statistically significant interaction between condition and voice in the average filler accuracies of the SG models ($F(1, 32) = 166.54$, $p < 0.001$).

It has been reported that humans often misinterpret the agent and patient of reversal anomaly sentences because role-filler assignment might sometimes rely on heuristics based on expectations about events, and is not always in line with the syntactic structure of the sentence [2, 1]. The SG model, a simple connectionist model of sentence comprehension trained on mapping sentences to event representations displays similar biases as humans when it comes to comprehend reversal anomaly sentences. The model thus provides a computationally explicit account of plausibility based comprehension.

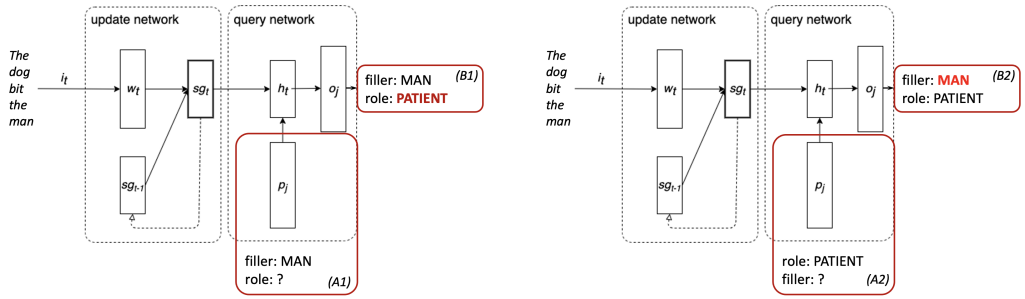

Figure 1: Probing for roles (left): the model is given a sentence and filler, and it is expected to produce the correct role of that word. Probing for fillers (right): the model is given a sentence and semantic role, and it is expected to produce the correct filler.

	C active			
	Ag	Pat	Prd	M*
Ag	**91.98**	6.91	0.00	1.11
Pat	1.60	**91.98**	2.59	3.83
	RA active			
	Ag	Pat	Prd	M*
Ag	**4.81**	88.27	2.59	4.32
Pat	48.27	**50.12**	0.00	1.60
	C passive			
	Ag	Pat	Prd	M*
Ag	**44.57**	51.36	0.00	4.07
Pat	1.60	**90.62**	2.84	4.94
	RA passive			
	Ag	Pat	Prd	M*
Ag	**5.93**	81.23	2.84	10.00
Pat	37.41	**60.62**	0.00	1.98

Table 1: Role probing confusion matrix for our four conditions. Rows indicate correct (target) roles, columns the percentage of correct (in bold) and misclassified fillers.

	C		RA	
	active	passive	active	passive
Ag	96.05	75.93	30.25	50.00
Pat	95.80	91.85	45.19	73.83
avg.	95.93	83.89	37.72	61.91

Table 2: Filler probing percentage accuracy scores.

References

[1] Fernanda Ferreira. The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2):164–203, 2003.

[2] Fernanda Ferreira, Karl Bailey, and Vittoria Ferraro. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15, 11 2002.

[3] Albert E. Kim and Lee Osterhout. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52:205–225, 2005.

[4] Gina R. Kuperberg, Tatiana Sitnikova, David N. Caplan, and Phillip J. Holcomb. Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Brain research. Cognitive brain research*, 17 1:117–29, 2003.

[5] Marta Kutas and Kara D. Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–47, 2011.

[6] Alessandro Lopopolo and Milena Rabovsky. Predicting the n400 erp component using the sentence gestalt model trained on a large scale corpus. *bioRxiv*, 2021.

[7] James L. McClelland, Mark F. St. John, and Roman Taraban. Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4:287–335, 1989.

[8] Mante S. Nieuwland and Jos J. A. van Berkum. Testing the limits of the semantic illusion phenomenon: Erps reveal temporary semantic change deafness in discourse comprehension. *Brain research. Cognitive brain research*, 24 3:691–701, 2005.