Korean Journal of English Language and Linguistics, Vol 22, October 2022, pp. 1101-1115
DOI: 10.15738/kjell.22..202210.1101

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

An Experimental Investigation of Discourse Expectations in Neural Language Models*

Eunkyung Yi (Ewha Womans University) **Hyowon Cho** · **Sanghoun Song** (Korea University)

Received: August 25, 2022
Revised: October 20, 2022
Accepted: October 30, 2022

Eunkyung Yi (1st author)
Assistant Professor, Dept. of English Education, Ewha Womans University
Tel: 02) 3277-4699
Email: eyi@ewha.ac.kr

Hyowon Cho (co-author)
Undergraduate Student, Dept. of Linguistics, Korea University
Tel: 02) 3290-2170
Email: snhan9658@naver.com

Sanghoun Song (corresponding author)
Associate Professor, Dept. of Linguistics, Korea University
Tel: 02) 3290-2177
Email: sanghoun@korea.ac.kr

ABSTRACT

Yi, Eunkyung, Hyowon Cho and Sanghoun Song. 2022. An experimental investigation of discourse expectations in neural language models. *Korean Journal of English Language and Linguistics* 22, 1101-1115.

The present study reports on three language processing experiments with most up-to-date neural language models from a psycholinguistic perspective. We investigated whether and how discourse expectations demonstrated in the psycholinguistics literature are manifested in neural language models, using the language models whose architectures and assumptions are considered most appropriate for the given language processing tasks. We first attempted to perform a general assessment of a neural model's discourse expectations about story continuity or coherence (Experiment 1), based on the next sentence prediction module of the bidirectional transformer-based model BERT (Devlin et al. 2019). We also studied language models' expectations about reference continuity in discursive contexts in both comprehension (Experiment 2) and production (Experiment 3) settings, based on so-called Implicit Causality biases. We used the unidirectional (or left-to-right) RNN-based model LSTM (Hochreiter and Schmidhuber 1997) and the transformer-based generation model GPT-2 (Radford et al. 2019), respectively. The results of the three experiments showed, first, that neural language models are highly successful in distinguishing between reasonably expected and unexpected story continuations in human communication and also that they exhibit human-like bias patterns in reference expectations in both comprehension and production contexts. The results of the present study suggest language models can closely simulate the discourse processing features observed in psycholinguistic experiments with human speakers. The results also suggest language models can, beyond simply functioning as a technology for practical purposes, serve as a useful research tool and/or object for the study of human discourse processing.

KEYWORDS

discourse expectation, implicit causality bias, neural language model, BERT, GPT-2, LSTM, next sentence prediction, coreference resolution, surprisal

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A03042760). We thank the anonymous reviewers for their valuable comments. We also would like to thank Unsub Shin for his feedback on an earlier draft. Any remaining errors are solely our responsibility.

1. Introduction

The underlying mechanisms of human language have been extensively explored for decades by language scientists from many academic disciplines including linguistics, cognitive psychology and computer science. The research enhanced to a great extent our understanding of how language works and ultimately, how speakers or, more in general, human beings function communicatively and intellectually as language is one of the most critical features that characterize humans. In the era of digitalization and automation, however, language scientists are facing new and unforeseen challenges. While linguists are benefiting from unprecedentedly large amounts of genuine (and frequently free) language data they can use in building or confirming a theory, they are also witnessing automated language processors such as neural language models that function well enough without help of the theoretical constructs that linguists have believed to be essential in natural language processing. For example, some latest deep neural networks or deep-learning language models are shown to be highly successful in performing various linguistic tasks with unsupervised training on real texts, namely with no explicit information on how to analyze them linguistically. This seems to make some linguists rethink the underlying mechanisms of language and, at the same time, urges them to examine to what extent such language models are successful in what linguistic tasks, compared to human linguistic capabilities. A close evaluation of how similar or different human and neural language processors are in certain linguistic tasks may also contribute to our understanding of human linguistic mechanisms in turn. In this context, the present study attempts to investigate the linguistic capabilities of up-to-date deep neural language models in discourse processing and evaluate to what extent they are close to or different from what we know about the human discourse processor.

In the realm of computational linguistics and natural language processing, much effort has been made to construct computational algorithms or systems that can process discourses or texts properly, i.e., those that understand and detect relationships between sentences. Earlier attempts include Rhetorical Structure Theory (Mann and Thompson 1987) and Penn Discourse TreeBank (Miltsakaki et al. 2004, Prasad et al. 2008, 2014). They perform discourse parsing largely based on discourse connectives and pre-determined labels of possible intersentential meaning relationships. It was shown that they are, to some degree, successful for some natural language processing tasks such as text summary, inference, sentiment analysis and machine translation. As noted above, recent neural network language models trained on large quantities of corpora are shown to significantly outperform those traditional models (e.g., Shi and Demberg 2019). Recent literature in computational linguistics suggests language models can encode some abstract linguistic representations not only at the syntactic and semantic levels (e.g., Linzen et al. 2016, Wilcox et al. 2018, 2019) but also at the discourse-pragmatic level (e.g., Jeretic et al. 2020), although the study on the robustness of such representations is still in progress.

A body of psycholinguistics research demonstrated language processing is incremental and predictive (see Pickering and Gambi (2018) for a review). Speakers were shown to expect what comes next proactively at every step of incoming linguistic units in language comprehension. Similarly, speakers were shown to exhibit the tendency to produce what is highly expected in specific contexts (e.g., Garvey and Caramazza 1974). Building on the recent development of neural network language models and the psycholinguistic findings, we investigate in this study whether neural language models' predictions in the course of discourse processing are close to human discourse expectations. We first examined discourse expectations about coherence or story continuity in Experiment 1, i.e., what comes next is coherent with what is said previously, and, second, examined expectations about reference continuity in comprehension and production settings in Experiments 2 and 3, respectively, i.e., with which referent the next story is continued given what is said. Using the same stimuli, we attempt to compare

the discourse expectation patterns manifested in neural language models with those observed in human speakers. What follows introduces the linguistic and psycholinguistic bases for the three experiments.

As noted in Hobbs (1979), when speakers process successive utterances that constitute a discourse, "some desire for coherence is operating." Discourse coherence, for example, guarantees that "he" in (1a) can refer only to John, which makes the two successive sentences be about the same entity. Despite the continued aboutness or coreference between *John* and *he*, this short discourse in (1a) does not stand up easily as a coherent story. It is an unusual or unlikely story to one's knowledge or understanding of the world. However, the "desire for coherence" is strong enough to drive comprehenders to any possible explanations that may make the story sound marginally coherent (or at least making-sense) such as "Instanbul was famous for spinach then." This suggests that the degree to which a discourse is coherent is related to how *probable* or *likely* a story is to occur based on world knowledge. In other words, the more probable a story is to one's knowledge of the world, the more easily or conveniently it is understood as a coherent one. Drawing on this relationship between coherence and probability, we examine in Experiment 1 whether language models can reasonably assess the degree to which a discourse is coherent or reasonably expected. We prepare two sets of 1,000 discourses consisting of two successive sentences, i.e., human-constructed intersentential discourses vs. randomly-paired sentences, and compare the probabilities of the second sentences in the two groups to verify whether neural models can discriminate between expected and unexpected discourses in human communication.

(1) a. John took a train from Paris to Istanbul. He likes spinach. (Hobbs 1979)
b. John took a train from Paris to Istanbul. _____

Garvey and Caramazza (1974) showed some verbs or the events they denote are associated with different patterns of causal attribution. The phenomenon is referred to as Implicit Causality (IC) biases and has been used in the literature to investigate people's cognitive, linguistic and/or social tendency in perceiving and producing causal relations (e.g., Rohde, Levy and Kehler 2011). It is also useful to test reference resolution in discourse processing research. When speakers read a sentence stimulus illustrated in (2), for example, they encounter a pronoun in the subordinate clause that refers to either NP_1 or NP_2 in the main clause. Studies showed that speakers process the pronoun faster when it refers to the NP that corresponds to the IC bias of the verb (e.g., Caramazza, Grober, Garvey and Yates 1977). For example, speakers were faster in processing *he* in (2a) that refers to NP_1 'the man' than *she* in (2b) that refers to NP_2 'the actress' as the verb to *confess* in the main event is known to be NP_1-biased. Similarly, when the verb is NP_2-biased such as to *criticize*, speakers were faster in processing *he* in (2d) that refers to NP_2 'the priest' than *she* in (2c) that refers to NP_1 'the woman.' However, such biases were not confirmed with neutral or non-IC verbs such as to *argue* as in (2e-f). The results confirmed that language processing is expectation-based such that speakers' expectations generated by IC-bias verbs modulate the processing of the upcoming linguistic input. We examine in Experiment 2 whether such preference patterns in refence choice observed in human discourse processing can be replicated in neural language models.

(2) (NP_1-biased IC verb)
a. The man$_{NP1}$ *confessed* to the actress$_{NP2}$ because <u>he</u> (= NP_1) ...
b. The man$_{NP1}$ *confessed* to the actress$_{NP2}$ because <u>she</u> (= NP_2) ...
(NP_2-biased IC verb)
c. The woman$_{NP1}$ *criticized* the priest$_{NP2}$ because <u>she</u> (= NP_1) ...
d. The woman$_{NP1}$ *criticized* the priest$_{NP2}$ because <u>he</u> (= NP_2) ...
(Neutral/non-IC verb)

e. The man$_{NP1}$ *argued* with the lady$_{NP2}$ because <u>he</u> (= NP$_1$) ...
f. The man$_{NP1}$ *argued* with the lady$_{NP2}$ because <u>she</u> (= NP$_2$) ...

IC biases were also confirmed in the context of production (Garvey and Caramazza 1974). They were demonstrated by speakers' continuations to sentence fragments, as illustrated in (3). The incomplete subordinate clause headed by *because* is intended to elicit a causal explanation to the event depicted in the main clause. Notably, the subject of the *because* clause is given in a pronoun that can refer either to NP$_1$ or to NP$_2$ of the main clause. For example, *he* in (3a) may refer either to 'the man' or to 'the priest'; *she* in (3b) can refer to 'the woman' or 'the actress'; *they* in (3c) can refer to 'the crew' or 'the critics.' However, previous studies showed speakers tend to exhibit biases in their choice of reference for the pronouns depending on the verbs. Namely, after the clause with to *confess*, an NP$_1$-biased verb, in (3a), speakers tend to produce a sentence where *he* refers to 'the man' (NP$_1$) rather than to 'the priest' (NP$_2$), choosing 'the man' as the causer of the confessing event as in *because he wanted absolution*. Whereas, for the verb to *criticize*, an NP$_2$-biased verb, in (3b), they tend to choose 'the actress' (NP$_2$) as the causer of the event rather than 'the woman' (NP$_1$) as in *because she was rude to the staff*. For verbs like to *argue* as in (3c), speakers tend not to show any clear bias in their causal attribution, i.e., a neutral or non-IC verb. We examine in Experiment 3 whether IC-bias patterns in reference choice observed in human speakers are also manifested in the sentence generation or production module of neural language models.

(3) a. The man$_{NP1}$ *confessed* to the priest$_{NP2}$ because he _____ (NP$_1$-biased IC verb)
b. The woman$_{NP1}$ *criticized* the actress$_{NP2}$ because she _____ (NP$_2$-biased IC verb)
c. The crew$_{NP1}$ *argued* with the critics$_{NP2}$ because they _____ (Neutral/non-IC verb)

The organization of this paper is as follows. Section 2 reports on Experiment 1 that investigates BERT's (Devlin et al. 2019) capability of assessing discourse expectations or coherence by comparing random sentence pairs with the actual production data obtained in a discourse completion experiment. Section 3 reports on Experiment 2 that examines whether neural language models exhibit human-like discourse expectations for upcoming reference in the comprehension of causal discourses, based on LSTM (Hochreiter and Schmidhuber 1997) and surprisal (Levy 2008). Section 4 reports on Experiment 3 that investigates whether human-like reference choices can also be replicated in GPT-2's (Radford et al. 2019) generation of causal discourses. Section 5 summarizes the results and concludes the paper.

2. Experiment 1: BERT's Evaluation of Discourse Coherence

We first conduct an overall evaluation of an up-to-date language model's capacity in evaluating discourse coherence. As introduced above, two successive sentences are considered coherent in human discourses when the second sentence story is highly probable. Drawing on the relationship between coherence and probability, we examine in this experiment whether a language model, BERT (Devlin et al. 2019), can reasonably estimate the degree to which the next (or second) sentence is coherent with its preceding one. More specifically, we measure BERT's estimations of discourse coherence based on next or second sentence probabilities in both human-produced intersentential discourses (naturally intended to be coherent) and randomly paired successive sentences and compare the two sets of results to each other. We predict, first, that the language model returns significantly different second-sentence probabilities for the two sets of discourses and, second, that the second sentence

probabilities in human-constructed discourses tend to be significantly higher than those in random sentence pairs. We use BERT in this experiment because its training paradigm is most close to the present task, i.e., measuring the probability of next sentences in the masked position. BERT is trained with masked language modeling and next sentence prediction objectives (see Devlin et al. 2019 for more details). What follows details the experiment.

2.1 Method

2.2.1 Material

We prepared two sets of 1,000 short discourses, each consisting of two sentences: One set was curated from human-constructed discourses and the other was constructed from randomly paired successive sentences. For the collection of human data, we retrieved 1,000 two-sentence long discourses obtained in Yi and Koenig's (2021) story continuation experiment. Each discourse consists of a sentence stimulus that serves as a prompt given to participants and their actual continuation to the prompt, as illustrated in (4). For example, "Carl fixed the computer for Margaret" is presented to participants as a stimulus and "Margaret paid him $50" is the next sentence that a participant provided to continue the story. In that experiment, participants were allowed complete freedom in constructing the second sentence as long as they think the story makes sense.

(4) Examples of human-produced discourses
a. Carl fixed the computer for Margaret. Margaret paid him $50.
b. Fred made pasta for Alice. He was born in Italy.

(5) Examples of artificially-made discourses (random sentence pairings)
a. We don't serve anything. But I think there's been a mistake.
b. The mine shut down. A staircase in Mexico inspired this work.

To prepare artificially-made discourses, we first collected 2,000 utterances from the Corpus of Contemporary American English (Davis 2008) that are three- to seven-word-long and end with a period. Then, we randomly paired them up to construct 1,000 discourses consisting of two successive sentences as illustrated in (5). As one can expect, the majority of random pairs do not seem to make any sense, e.g., (5b), but some, by chance, sound fully or marginally acceptable when interpreting with the "desire for coherence," e.g., (5a).

2.1.2 Data generation using BERT-base and data analysis

As alluded to above, we used the representative transformer-based language model BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019) in this experiment. BERT employs the attention mechanism as a building block to embed a context-based or sentence-level representation into a neural language model and is trained with two learning objectives, namely, Mask Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM trains the model in a way it pays attention to the context around a mask and predicts what linguistic expression should be filled into the masked position, which is similar to performing a fill-in-the-blank task. In NSP, the model is fed with pairs of sentences and is asked to decide whether each pair consists of actual adjacent sentences appearing in the training corpus or not, i.e., distinguishing true pairs from random pairs. These mechanisms of BERT allow us to calculate the probability of the second sentence given the first one as a context

in pairs of adjacent sentences. We used BERT-base (a default model) with 110 million learning parameters, 768 hidden layers, 12 transformer blocks, and maximum 512-word context windows.

We computed second-sentence probabilities in 2,000 pairs of adjacent sentences, i.e., 1,000 random and 1,000 human-produced pairs, using BERT-base. We then analyzed the data using two sample independent *t*-test to determine whether the probability distributions of the two groups statistically differ from each other. We further made other notable observations in the two distributions.

2.2 Result and Discussion

As predicted, the model produced relatively higher probabilities for human-constructed intersentential discourses than for the random sentence pairs. As illustrated in Figure 1, the model produced over 90% second-sentence probabilities for 92.3% of the human-constructed discourses and only for 23.6% of the random pairs. In contrast, the model produced less than 10% probabilities only for 6.8% of the human pairs and for 72.1% of the random pairs. The result of the two-sample independent *t*-test showed the distributions of second-sentence probabilities are statistically different between human-produced discourses and random sentence pairs ($t = 42.993$, $p = .000$). It should also be noted in the data that BERT produced extreme probability scores, i.e., either very high (over 90%) or very low (below 10%), and rarely yielded medium scores. Only 0.9% and 4.3% of the human and random pairs, respectively, yielded probabilities between 10% and 90%.

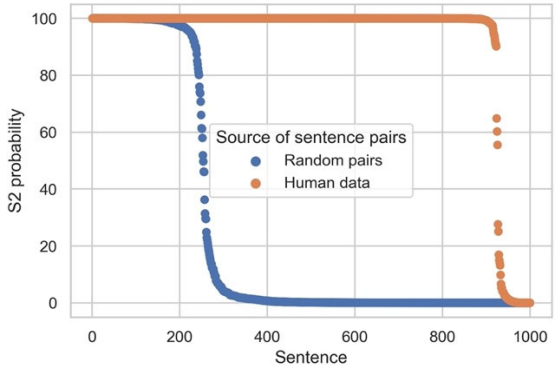

Figure 1. The Distribution of 2nd Sentence Probabilities of Random Pairs and Human Production Data Computed by BERT-base

The result demonstrates, first, that the neural language model BERT can evaluate discourse coherence to a significant degree. It also shows that, although probability is gradient in itself, the model seems to make binary-like decisions, i.e., whether a discourse is coherent or not, by producing extremely high or low probabilities for the second sentences in sentence pairs.

3. Experiment 2: LSTM's Reference Expectations in Comprehension

Section 2 above reported on an experiment that evaluates a neural language model's overall capacity in capturing discourse coherence between two successive sentences. The results showed the neural model can detect human-

like discourse coherence and discriminate them from random and thus highly improbable and incoherent discourses. In this experiment, we move on to test a more specific discourse-related phenomenon often called coreference resolution, i.e., choice of referent when alternatives are available. As introduced in Section 1, language processing is known to be largely expectation-based. Speakers usually have expectations about upcoming linguistic units at almost every level of granularity. For example, listeners are likely to expect a certain phoneme or syllable to occur next more than others given the phonetic and phonological information processed earlier in speech perception. They tend to predict a certain part of speech or phrase to follow based on words and phrases that have been preceded over the course of online sentence comprehension. Studies showed discourse expectations are also at work and speakers expect some content, i.e., a specific referent or discourse coherence relation, to occur next given the story that has preceded (e.g., Rohde 2008). In the present experiment with a neural language model, we examine whether neural language models can simulate discourse expectations or more specifically reference expectations demonstrated in human language processing. We used Implicit Causality biases introduced above to construct the material and used surprisal (Hale 2001) and LSTM (Hochreiter and Schmidhuber 1997) to estimate language models' expectations for a referent in a discourse context.[1]

3.1 Method

3.1.1 Material

We prepared 240 sentence stimuli in the frame of *NP_1 V NP_2 because Pronoun …* that is often used in classic psycholinguistic experiments to test IC-biases (e.g., Garvey and Caramazza 1974). As illustrated in (6), we used third person pronouns, *he/him* and *she/her*, for the NP1 and NP2 positions. In order to avoid ambiguity in reference resolution, gender was counterbalanced using two unambiguous combinations such as *he*$_{(NP1)}$-*her*$_{(NP2)}$ and *she*$_{(NP1)}$-*him*$_{(NP2)}$. Each combination is appended with two different incomplete *because* clauses. In one, the subject of the *because* clause refers to NP1; in the other, it refers to NP2. For the verb position, we used three types of IC-bias verbs in past tense, namely, NP1-biased (*confessed to*), NP2-biased (*criticized*) and neutral or no-bias (*greeted*) verbs. Twenty verbs for each type are listed in (7).

(6) a. **He** *confessed to/criticized/greeted* **her** because ***he***$_{=NP1}$/***she***$_{=NP2}$…
b. **She** *confessed to/criticized/greeted* **him** because ***she***$_{=NP1}$/***he***$_{=NP2}$…

(7) a. (NP_1-biased) *aggravate*, *amaze*, *amuse*, *annoy*, *apologize to*, *bore*, *charm*, *confess to*, *deceive*, *disappoint*, *exasperate*, *fascinate*, *frighten*, *humiliate*, *infuriate*, *inspire*, *intimidate*, *offend*, *scare*, *surprise*
b. (NP_2-biased) *assist*, *blame*, *comfort*, *congratulate*, *correct*, *detest*, *envy*, *fear*, *hate*, *help*, *mock*, *notice*, *pacify*, *praise*, *reproach*, *scold*, *stare at*, *thank*, *trust*, *value*
c. (Neutral) *chat with*, *cook with*, *dine with*, *encounter*, *greet*, *hang with*, *meet*, *meet with*, *run into*, *run with*, *sing with*, *sit with*, *see*, *stand with*, *study with*, *talk with*, *wait for*, *walk with*, *watch*, *work with*

[1] The reviewers pointed out that the same experiment can be carried out using the masked language models, such as BERT. We agree it is possible, but we tried to use a model that fits best for what is assumed and tested in each experiment. Thus, we chose to use LSTM based on the memory cells since the current experiment lays focus on the sequential properties of linguistic items.

3.1.2 Data generation based on LSTM and surprisal and data analysis

For present purposes, we used the long short-term memory (LSTM) model (Hochreiter and Schmidhuber 1997). LSTM is based on recurrent neural networks (Elman 1990), which lays focus on linguistic units sequentially like other traditional language models such as N-gram. But it is known to achieve higher performance than ordinary RNN models by updating information selectively; the less important a piece of information is, the more the information is ignored in the training. The way this language model works is conceptually analogous to the way humans process linguistic input incrementally and predictively. We used the Google LSTM model (Chelba et al. 2013, Jozefowicz et al. 2016) trained on the One Billion Word Benchmark in this experiment.

Drawing on the mechanisms of this language model, we calculate *surprisal* on a particular linguistic unit (Hale 2001, Levy 2008, Smith and Levy 2013). Surprisal, or the negative log probability of a word given a context, is known to correlate with the degree of cognitive effort or difficulty that humans experience in processing a linguistic unit in online sentence comprehension. Many psycholinguistic studies confirmed that surprisal is correlated with, for example, reaction time in self-paced reading experiments. Namely, it takes a relatively long time for speakers to process a word with a high surprisal score when reading through a sentence. In other words, surprisal scores represent the degree to which a sequence of linguistic units, i.e., the preceding sequence plus the current word, sounds (un)natural or (un)expected. Highly expected words tend to have low surprisal scores. Conversely, unexpected words tend to have high surprisal scores. We measure surprisal at the last position of an input sequence, i.e., *he* or *she* after *because*, that refers either to NP1 or NP2.

We perform a two-way ANOVA to analyze the effect of verb type (NP1-biased, NP2-bised or neutral) and reference choice (NP1 or NP2) for the continued subjects in the *because* clause on their surprisal scores. We predict that the interaction between verb type and reference choice is statistically significant, showing the language model performs discourse processes in a way human processors do. We further analyze each verb type to examine whether surprisal scores significantly differ between NP1- and NP2-referents of the subject in the *because* clause.

3.2 Result and Discussion

As illustrated in Figure 2, the results revealed the patterns of surprisal scores differ in three verb types as was reported in previous psycholinguistic research. Namely, NP1 referents tend to yield lower surprisal scores (M = 1.32, SD = 0.348) than NP2 referents (M = 2.20, SD = 0.777) for the NP1-biased IC verbs. Conversely, NP1 referents tend to yield higher surprisal scores (M = 1.90, SD = 0.854) than NP2 referents (M = 1.52, SD = 0.674) for the NP2-biased IC verbs. The difference in mean surprisal scores between NP1 and NP2 referents was the smallest for the verbs known to have no IC biases, i.e., M = 1.68, SD = 0.565 and M = 1.52, SD = 0.421, respectively.

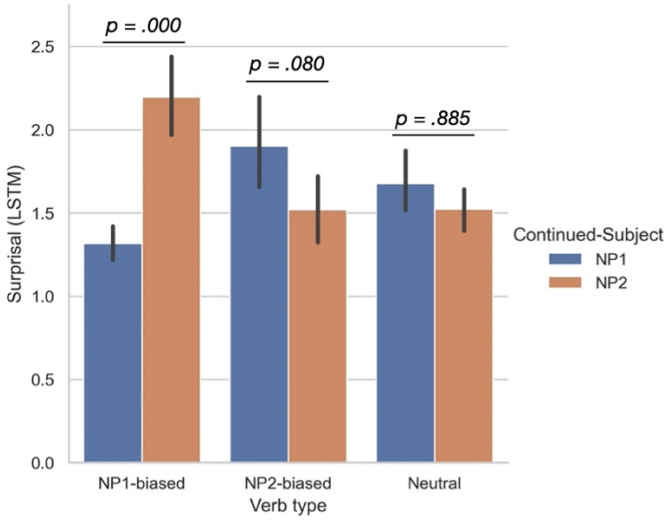

Figure 2. Surprisal of Continued-Subject (referring to NP1 or NP2) by Verb Type

The results of a two-way ANOVA revealed that, as predicted, there is a statistically significant interaction between verb type and reference choice on the effect of surprisal at the continued-subject position ($F(2, 234) = 22.525$, $p = .000$). We found no significant simple main effects: Verb type did not have a statistically significant effect on surprisal of the continued subject ($F(2, 234) = 1.302$, $p = .274$). Reference choice also did not have a significant effect ($F(1, 234) = 1.965$, $p = .162$). We performed post hoc comparisons using Tukey's test to analyze the differences in surprisal between NP1 and NP2 referents within each verb type. The results showed that the difference is significant ($p = .000$) for the NP1-biased verbs and marginal ($p = 0.08$) for the NP2-biased verbs while it is not significant ($p = .885$) for the neutral verbs.

Our results showed that the surprisal score at the subject of the *because* clause can be modulated jointly by what IC bias a verb has in the prompt sentence and which preceding argument (NP1 or NP2) the subject of the continued story refers to. As shown in previous experiments on human subjects, we found in this experiment with the LSTM model that surprisal on the continued subject tends to be lower when it refers to the argument the verb is biased towards with respect to causal attribution than when it does not. In a psycholinguistic perspective, as noted above, surprisal represents the processing cost or the amount of cognitive effort one makes at a certain word. Our result suggests that the LSTM model also exhibits or "experiences" human-like patterns of difficulty in reference resolution in processing inter-clausal discourses. It also suggests that the neural model, LSTM, is a viable experimental method in investigating human discourse processes particularly in the incremental comprehension setting.

4. Experiment 3: GPT-2's Choice of Reference in Next Sentence Production

We have shown in Section 3 that a neural language model LSTM replicates the results of psycholinguistic experiments that investigated reference resolution in language comprehension. In the present experiment, we attempt to examine whether language models can also generate or "produce" sentences with a choice of referent that accords with speakers' preference in the same discourse contexts. We also utilize Implicit Causality biases in

constructing our material and conduct an experiment with a neural language model called GPT-2. Details are in what follows.

4.1 Method

4.1.1 Material

We constructed 360 incomplete sentences as sentence stimuli, adapting the material used in the previous experiment. As before, sixty verbs in three IC bias types listed in (7) above, i.e., twenty NP1-biased, twenty NP2-biased and twenty neutral verbs, occur with NP1 (subject) and NP2 (object). NP1 and NP2 are gender-crossed combinations of three different types. In one, they occur in 3rd person singular pronouns such as *he-her* and *she-him* as in (8a-b) or in a full noun phrase consisting of either a definite or indefinite determiner and a noun such as *a/the man-a/the woman* or *a/the woman-a/the man* as in (8c-d). Genders are crossed in each stimulus to avoid ambiguity in reference resolution. Then they are followed by the connective *because*, as illustrated in (8).

(8) a. **He** *confessed to/criticized/greeted* **her** because…
b. **She** *confessed to/criticized/greeted* **him** because…
c. **A/the man** *confessed to/criticized/greeted* **a/the woman** because…
d. **A/the woman** *confessed to/criticized/greeted* **a/the man** because…

4.1.2 Data generation using GPT-2

We put the incomplete sentence fragments into a neural language model called GPT-2 (Generative Pre-trained Transformer; Radford et al. 2018, 2019). GPT-2 is a neural language model tested reliable and suitable for sentence generation. It has a transformer-based architecture like BERT used in Experiment 1 above but works in a unidirectional way such that each state in generation is calculated sequentially and incrementally from the beginning. Among different versions of GPT-2 depending on the size of training data, we used the 'large' version for the present experiment. We let the model generate the rest of the incomplete sentences at six different levels of *temperature*, i.e., 0.1, 0.3, 0.5, 0.7, 0.9 and 1.0. Temperature is one of the significant hyperparameters in using deep-learning generation models. It is used to adjust the level of randomness in predictions and avoid too much stereotyped patterns in the output. Higher temperatures tend to allow for more random or less fixed expressions.[2] We repeated sentence generation ten times at each temperature level so that we can observe the model's generation preferences with more output samples. We end up with 3,600 sentences as output at each temperature. In total, we had 21,600 sentences generated from six different temperature levels.

4.1.3 Data coding and analysis

We coded each sentence output as to which NP the subject of the *because* clause refers to, namely, the model's choice of causal attribution. If a fragmental stimulus is continued with a referent referring to NP1 after *because*, it is coded as *NP1 continuation*. If it is continued with NP2, it is coded as *NP2 continuation*. If the subject of the

[2] For more information about this parameter, see Goodfellow et al. (2016; §17.5.1., p. 605) or a blog post of Hugging Face (https://huggingface.co/blog/how-to-generate).

because clause refers to neither NP1 nor NP2, the sentence is coded as *other continuation*. Lastly, if *because* is not followed by a clausal complement, e.g., *because of the time*, it is coded with *NA*.

We analyzed the data in two respects. First, we made close observations on the overall distributional differences in continuation types depending on IC bias types as well as on temperatures. Second, we conducted logistic regression analyses to examine which referent, namely either NP1 or NP2, the model prefers to choose for the subject of the *because* clause, i.e., the causer of the event denoted by the main clause. We report the results in the next section.

4.2 Result and Discussion

The model's continuation types vary as to the temperature parameter. At lower temperatures, the model tends to continue strictly with either NP1 or NP2 as the referent of the subject in the *because* clause. However, the higher temperature the model is at, the more 'other' continuations it generates. Table 1 below summarizes the counts of each type of continuations at six different temperatures. For example, at the relatively low temperatures such as 0.1 and 0.3, more than 99% of the completed *because* clauses begin with NP1 or NP2 as the referent of their subjects. NP1 and NP2 continuations are down to 73.6% at the temperature of 1.0 and 'other' continuations constitute 26.4% of the generated output.

Table 1. The Counts of Continuation Types Depending on Six Levels of the Temperature Parameter

Temperature	Continuations with NP1 or NP2			Continuations with neither NP1 nor NP2		
	NP1	NP2	Total (%)	Other	NA	Total (%)
0.1	937	2658	3595 (99.86)	0	5	5 (0.14)
0.3	1039	2532	3571 (99.19)	0	29	29 (0.81)
0.5	1159	2308	3467 (96.31)	32	101	133 (3.69)
0.7	1164	2037	3201 (88.92)	110	289	399 (11.08)
0.9	1085	1720	2805 (77.92)	395	400	795 (22.08)
1.0	1072	1577	2649 (73.58)	585	366	951 (26.42)

The results also revealed the model's general preference for NP2 continuations over NP1 continuations, as illustrated in Figure 3. The model continued the story more frequently with NP2 than with NP1 regardless of verbs' IC biases. This tendency is pervasive in all six different temperature levels we tested. But the degree to which the model is biased towards NP2 for the subject of the *because* clause seems to vary depending on verbs' IC-bias types. In other words, the model is more biased towards NP2 continuations after NP2-biased verbs in the main clause and less so after NP1-biased verbs. As predicted, the model is least biased after neutral verbs in all temperatures. The results of statistical analysis on these trends are reported below.

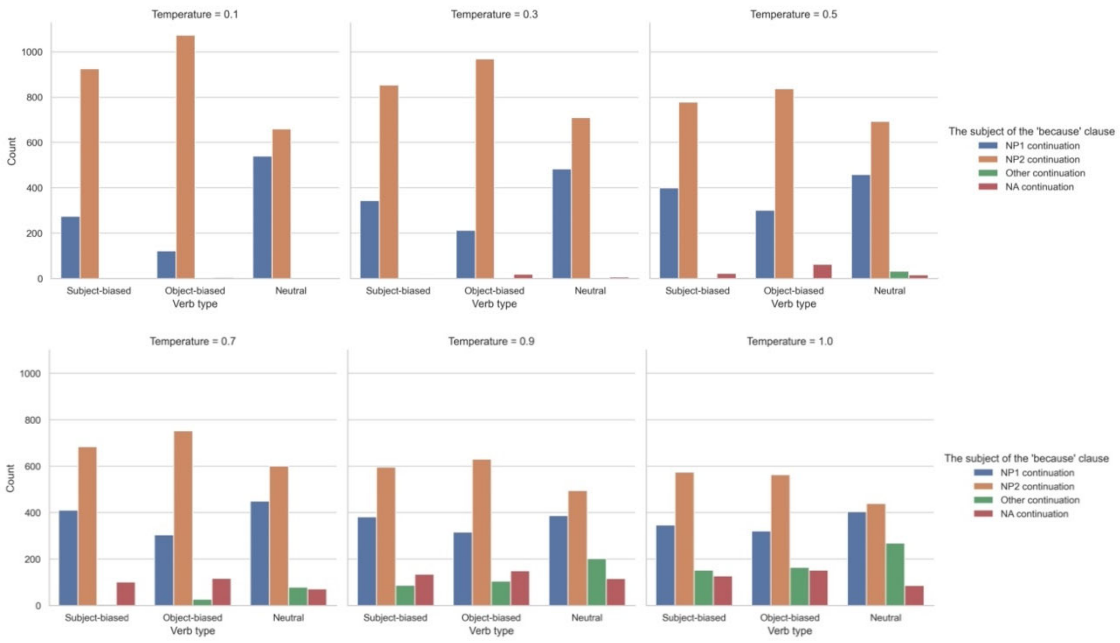

Figure 3. Reference Choice for the Subject of the *because* Clause in GPT-2's Next Sentence Production (at Six Different Temperatures)

We performed logistic regression analyses with verb type as the predictor variable and the choice of continued referent as the outcome variable. For these analyses, we subset the data to include only NP1 and NP2 continuations, i.e., focusing on the choice between NP1 and NP2 while excluding *other* and *NA* continuations. The models test whether the choice of reference in the model's continuation is modulated by verb type. The results showed that the model's reference choice after NP1-biased and NP2-biased verbs, respectively, significantly differs from that after neutral verbs at all six temperature levels, as illustrated in the second column in Table 2. The biases in the choice between NP1 and NP2 continuations tend to be the smallest after neutral verbs, medium-sized after NP1-biased and the largest after NP2-biased verbs. We repeated the same logistic regression analyses, but with NP1-biased verbs as the reference level (i.e., baseline) to verify whether the results of NP1-biased and NP2-biased verbs also statistically differ from each other. As illustrated in the third column in Table 2, we found the reference choice after NP2-biased verbs significantly differs from that after NP1-biased verbs in five temperature levels except for the highest 1.0 temperature (b = 0.058, p = 0.548). In short, the neural model's continuations after IC verbs tend to be different from those after neutral verbs irrespective of the temperature settings while the differences in continuations after NP1- and NP2-biased verbs are reliably significant in the relatively lower temperature settings.

Table 2. The Results of Logistic Regression Analyses at Six Temperatures (Varying the Reference Level)

Temp	NP1- and NP2-biased verbs against neutral verbs (Reference level = neutral verbs)					NP2-biased and neutral verbs against NP1-biased verbs (Reference level = NP1-biased verbs)				
0.1		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)
	(Intercept)	0.20067	0.05803	3.458	0.000544 ***	(Intercept)	1.21302	0.06868	17.661	< 2e-16 ***
	NP1-biased	1.01235	0.08991	11.259	< 2e-16 ***	NP2-biased	0.96117	0.11766	8.169	3.12e-16 ***
	NP2-biased	1.97352	0.11178	17.656	< 2e-16 ***	Neutral	-1.01235	0.08991	-11.259	< 2e-16 ***
0.3		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)
	(Intercept)	0.38525	0.05898	6.532	6.50e-11 ***	(Intercept)	0.90812	0.06387	14.218	< 2e-16 ***
	NP1-biased	0.52287	0.08694	6.014	1.81e-09 ***	NP2-biased	0.61156	0.09914	6.169	6.88e-10 ***
	NP2-biased	1.13443	0.09606	11.809	< 2e-16 ***	Neutral	-0.52287	0.08694	-6.014	1.81e-09 ***
0.5		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)
	(Intercept)	0.41198	0.06018	6.846	7.61e-12 ***	(Intercept)	0.66777	0.06158	10.845	< 2e-16 ***
	NP1-biased	0.25579	0.08610	2.971	0.00297 **	NP2-biased	0.35495	0.09115	3.894	9.86e-05 ***
	NP2-biased	0.61073	0.09021	6.770	1.29e-11 ***	Neutral	-0.25579	0.08610	-2.971	0.00297 **
0.7		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)
	(Intercept)	0.28935	0.06234	4.642	3.46e-06 ***	(Intercept)	0.51180	0.06246	8.194	2.52e-16 ***
	NP1-biased	0.22245	0.08824	2.521	0.0117 *	NP2-biased	0.39391	0.09231	4.267	1.98e-05 ***
	NP2-biased	0.61636	0.09222	6.683	2.34e-11 ***	Neutral	-0.22245	0.08824	-2.521	0.0117 *
0.9		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)
	(Intercept)	0.24613	0.06785	3.627	0.000286 ***	(Intercept)	0.44314	0.06556	6.759	1.39e-11 ***
	NP1-biased	0.19701	0.09435	2.088	0.036801 *	NP2-biased	0.24684	0.09513	2.595	0.00947 **
	NP2-biased	0.44384	0.09673	4.589	4.46e-06 ***	Neutral	-0.19701	0.09435	-2.088	0.03680 *
1.0		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)
	(Intercept)	0.08536	0.06891	1.239	0.215	(Intercept)	0.50330	0.06800	7.402	1.35e-13 ***
	NP1-biased	0.41794	0.09681	4.317	1.58e-05 ***	NP2-biased	0.05853	0.09755	0.600	0.548
	NP2-biased	0.47648	0.09818	4.853	1.22e-06 ***	Neutral	-0.41794	0.09681	-4.317	1.58e-05 ***

To summarize, we found human-like reference choices in the sentence continuations generated by GPT-2. Overall, the results are similar to those observed in the production experiments on human discourse expectations after IC-bias verbs. However, it should be noted that the similarities were statistically reliable only in the temperatures below 1.0. In other words, the neural language model behaves more like humans when the level of randomness is less granted.

5. Conclusion

In this study, we conducted three language processing experiments with most up-to-date neural language models in a psycholinguistic perspective and examined whether we can observe human-like language processing features or, more specifically, discourse expectation patterns in neural language models' processing of discourses.

In Experiment 1, we assessed the overall discourse processing capacity of neural language models using the bidirectional transformer-based neural language model, BERT. We tested whether the model can capture discourse (in)coherence between two sentences that characterizes human discourse. We computed the probabilities of the second sentences in both human-constructed and randomly-matched sentence pairs and examined whether the model can discern human-constructed sentence pairs from random pairs. In Experiments 2 and 3, we investigated

neural models' discourse processing capacity in reference resolution based on Implicit Causality biases. We examined whether neural models exhibit human-like discourse expectations for reference choice in the course of sentence comprehension using the unidirectional (or left-to-right) RNN-based LSTM in Experiment 2 and in the course of sentence production using transformer-based GPT-2 in Experiment 3.

The results of the three experiments revealed that neural language models exhibit discourse processing patterns similar to those demonstrated in previous psycholinguistic experiments. First, in Experiment 1, BERT showed significantly higher next-sentence probabilities for human-constructed discourses than random sentence pairs, suggesting the model discriminates coherent discourses from incoherent ones to a significant extent. Second, the results of Experiments 2 and 3 showed LSTM and GPT-2 exhibit human-like preferences in reference resolution in the comprehension and production of two-clausal sentences.

The present study suggests language models can simulate the discourse processing features such as expectations of an upcoming story and reference choice based on discourse coherence that were previously observed in psycholinguistic experiments with human speakers. Although neural language models mainly function as a language technology used for natural language processing tasks such as text summarization and translation, our results further suggest that they can also serve as a useful research tool and/or object for the study of human discourse processing.

References

Caramazza, A., E. Grober, C. Garvey and J. Yates. 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior* 16, 601-609.

Davis, F. and M. van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 396-407. November 19-20. Association for Computational Linguistics.

Garvey, C. and A. Caramazza. 1974. Implicit causality in verbs. *Linguistic Inquiry* 5(3), 459-464.

Goodfellow, I., Y. Bengio and A. Courville. (2016). *Deep Learning*. MIT press.

Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1-8. Association for Computational Linguistics.

Hobbs, J. R. 1979. Coherence and coreference. *Cognitive Science* 3, 67-90.

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

Jin, X., X. Wang, X. Luo, S. Huang and S. Gu. 2020. Inter-sentence and Implicit Causality extraction from Chinese Corpus. In H. Lauw, R. W. Wong, A. Ntoulas, E. P. Lim, S. K. Ng and S. Pan, eds., *Advances in Knowledge Discovery and Data Mining*, 739-751. Springer, Cham.

Jeretic, P., A. Warstadt, S. Bhooshan and A. Williams. 2020. Are natural language inference models impressive? Leanring implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8690-8705. Association for Computational Linguistics.

Jurafsky, D. and J. H. Martin. To appear. *Speech and Language Processing*. 3rd ed. Prentice Hall.

Khetan, V., R. Ramnani, M. Anand, S. Sengupta and A. E. Fano. 2022. Causal-BERT: Language models for causality detection between events expressed in text. In K. Arai ed., *Intelligent Computing*, 965-980. Springer, Cham.

Kishimoto, Y., Y. Murawaki and S. Kurohashi. 2020. Adapting BERT to implicit discourse relation classification

with a focus on discourse connectives. In *Proceedings of the 12th Conference on Language Resources and Evaluation* (LREC 2020), 1152-1158.

Liang, S., Z. Wanli, Z. Shi, S. Wang, J. Wang and X. Zuo. 2022. A multi-level neural network for implicit causality detection in web texts. *Neurocomputing* 481, 121-132.

Linzen, T., E. Dupoux and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics* 4, 521-535.

Mann, W. C. and S. A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report RS-87-190, Information Sciences Institute.

Miltsakaki, E., R. Prasad, A. K. Joshi and B. L. Webber. 2004. The Penn Discourse Treebank. *LREC*.

Pickering, M. J. and C. Gambi. 2018. Predicting while comprehending language: a theory and review. *Psychological Bulletin* 144(10), 1002-1044.

Prasad, R., N. Dinesh, A. Lee, E. Milt- sakaki, L. Robaldo, A. K. Joshi and B. L. Webber. 2008. The Penn Discourse TreeBank 2.0. *LREC*.

Prasad, R., B. L. Webber and A. Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics* 40(4), 921-950.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI tech report.

Radford, A., K. Narasimhan, T. Salimans and I. Sutskever. 2018. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/language understanding paper. pdf

Rohde, H. 2008. *Coherence-driven Effects in Sentence and Discourse Processing*. Unpublished doctoral dissertation, University of California, San Diego.

Rohde, H., R. Levy and A. Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition* 118, 339-358.

Shi, W. and V. Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019, Poster)*, November 3-7, 2019, Hong Kong, China.

Stewart, A. J., M. J. Pickering and A. J. Sanford (1998). Implicit consequentiality. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 1031-1036. Mahwah, N: Lawrence Erlbaum Associates.

Yi, E. and J.-P. Koenig. 2021. Grammar modulates discourse expectations: evidence from causal relations in English and Korean. *Language and Cognition* 13(1), 99-127.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

