frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 19 December 2024
DOI 10.3389/fcomm.2024.1442093

OPEN ACCESS

EDITED BY
Luca Campanelli,
University of Alabama, United States

REVIEWED BY
William Schuler,
The Ohio State University, United States
Byung-Doh Oh,
New York University, United States

*CORRESPONDENCE
Eunjeong Oh
✉ eoh@smu.ac.kr

RECEIVED 08 June 2024
ACCEPTED 05 December 2024
PUBLISHED 19 December 2024

CITATION
Noh K, Oh E and Song S (2024) Testing language models' syntactic sensitivity to grammatical constraints: a case study of *wanna* contraction.
Front. Commun. 9:1442093.
doi: 10.3389/fcomm.2024.1442093

Testing language models' syntactic sensitivity to grammatical constraints: a case study of *wanna* contraction

Kangsan Noh[1], Eunjeong Oh[2]* and Sanghoun Song[1]

[1]Department of Linguistics, Korea University, Seoul, Republic of Korea, [2]Department of English Education, Sangmyung University, Seoul, Republic of Korea

Wanna contraction refers to the reduction of *want to* to *wanna*. Interestingly, native English speakers contract *want to* in object extraction questions but not in subject extraction questions. The present study investigated whether language models such as bidirectional encoder representations from transformers (BERT) adhere to this grammatical subtlety. *Wanna* contraction involves two factors: subject–object asymmetry and contraction. Disentangling these two ensures that when language models accurately identify illicit instances of *wanna* contraction, the detection stems from their understanding of the contraction, rather than the intervention by subject–object asymmetry. For this objective, we conducted three independent experiments. We tested whether language models detect illicit cases of contraction by maintaining constant contraction (Experiment 1) and question types (Experiment 2). We predicted that higher surprisal values would be assigned to ungrammatical instances. The overall results of the two experiments were in line with our prediction (87.5 and 75%, respectively). In addition, the analysis of by-word surprisal also indicates that the models generate higher surprisal values for subject extraction questions in illicit *wanna* instances (Experiment 3). Thus, the models' processing patterns of *wanna* contraction turn out to be close to those of native English speakers, suggesting their role as a research tool in linguistic experiments.

KEYWORDS

wanna contraction, *wh*-trace, language model, surprisal, subject–object asymmetry

1 Introduction

Among the important features of natural languages is that they are linear and hierarchical, rather than simply a linear arrangement of lexical items. Some linguistic phenomena representing this non-linear aspect of natural languages involve linguistic dependencies (Chomsky, 1957; Ross, 1967). The recent advancement in deep neural language models has led both theoretical and computational linguists to examine how neural language models process such dependencies. For instance, whether neural language models can capture English subject–verb dependencies as first-language (L1) English speakers do is an interesting line of research. This is because when sentences without a present-tense verb (e.g., The keys to the cabinet ___) are presented, language models must decide whether the number of the vacant verb is singular or plural (in this case, the answer is the plural *are*). Interestingly, neural language models have been shown to be sensitive to subject–verb dependencies (Linzen et al., 2016; Goldberg, 2019; Jawahar et al., 2019). The results of such experimental studies are informative and valuable to the extent that they have shown that neural language models successfully detect hierarchical structures such as subject–verb dependencies.

Considering the testing of neural language models' syntactic knowledge, researchers have suggested that this testing has implications for understanding human language acquisition (Linzen and Baroni, 2021). Neural language models differ cognitively from human learners as they are trained on large datasets. Nonetheless, their success in processing grammatical constraints may contribute to the debate on innate mechanisms in language acquisition. Although it would be inaccurate to claim that neural language models are entirely tabula rasa, their innate mechanisms appear to differ from those of humans. Consequently, if neural language models process a syntactic phenomenon in a way consistent with human syntactic competence, this may suggest that an innate principle for that phenomenon is not strictly necessary. In this regard, we believe that *wanna* contraction can be a phenomenon of interest. *Wanna* contraction is a grammatical phenomenon in which the verb *want* and the infinitival marker *to* are reduced to the form *wanna*, as in (1b). What makes *wanna* contraction interesting is that L1 English speakers do not contract *want to* in some cases, as in (2b).

(1) a. Who do you want to see tomorrow?
b. Who do you wanna see tomorrow?
(2) a. Who do you want to come tomorrow?
b. *Who do you wanna come tomorrow?

For cases such as (2b), many have tried to explain why contraction is not permitted (Lakoff, 1970; Lightfoot, 1976; Chomsky et al., 1977; Chomsky and Lasnik, 1977; Rotenberg, 1978; Postal and Pullum, 1982). Some have argued that this grammatical constraint underlying *wanna* contraction must be innate as even children seem to be aware of when contraction is impermissible (Chomsky, 1980; Crain and Pietroski, 2001). Previous research on *wanna* contraction has also shown that second-language (L2) speakers are generally less conservative than L1 English speakers, considering illicit sentences such as (2b) acceptable (O'Grady et al., 2008; Rezaeian et al., 2017).

Given the abovementioned characteristics of *wanna* contraction, we presume that testing language models' processing of *wanna* contraction would be more interesting in two ways. First, the processing of *wanna* contraction involves the understanding of an invisible *wh*-trace that blocks contraction in certain conditions. The processing of *wanna* contraction is related to an understanding of the asymmetry between subjects and objects as the contraction is allowable only in object extraction questions (Schachter and Yip, 1990; Stromswold, 1995; Juffs and Harrington, 1995; Ito, 2018). Second, the processing of *wanna* contraction is interesting because L1 and L2 speakers exhibit a contrast: L1 speakers mostly adhere to the constraint, while L2 speakers do not. Previous research on *wanna* contraction examined how it was processed by L1 and L2 learners (Thornton, 1990; Kweon, 2000; Kweon and Bley-Vroman, 2011), and we aim to broaden the research scope by considering artificial learners. We do not exclusively argue for or against specific linguistic theories on *wanna* contraction. Rather, we recognize the situation that *wanna* contraction is disallowed in some cases and this constraint seems to be acknowledged by L1 English speakers. By testing how language models process the grammatical constraint underlying *wanna* contraction, we investigate whether they show syntactic sensitivity to *wanna* contraction as L1 English speakers do.

In sum, the goal of the present study is to assess whether artificial learners, represented by neural language models, exhibit syntactic sensitivity to the grammatical constraint underlying *wanna* contraction. In other words, we aim to observe the behavior of neural language models through the adoption of an experimental design from a psycholinguistic perspective. Specifically, our approach is grounded in the premise that *wanna* contraction is disallowed in subject extraction questions, a constraint to which L1 speakers strongly adhere.

Building upon this premise, we present the following research questions: First, can neural language models capture the grammatical constraint underlying *wanna* contraction? Second, are there additional factors that might impact the processing of *wanna* contraction by neural language models? We aim to answer these two questions by conducting a series of experiments designed to test language models' syntactic sensitivity to *wanna* contraction.

2 Background

For neural language models, we employed two types of language models specialized for natural language understanding. First, we utilized bidirectional encoder representations from transformers (BERT), a bidirectional variant of transformer networks that considers both the left and the right contexts of the masked word (Devlin et al., 2019). Second, we employed the robustly optimized BERT approach (RoBERTa), an enhanced replication of BERT that includes longer training, more data, and different masking patterns (Liu et al., 2019).

Masked language models, such as BERT and RoBERTa, are trained to predict intentionally masked words within a sentence. During training, certain words are replaced with a special [MASK] token, and the model learns to predict the original word based on its context. Rather than processing the sentence from left to right, as autoregressive models such as generative pretrained transformers (GPTs) do, masked language models predict the masked word by considering both the preceding and following words. This bidirectional approach allows them to leverage context from both sides, providing a richer understanding of sentence structure.

3 Experiment 1: controlling for contraction

3.1 Methods

3.1.1 Materials

Based on the experimental design by Zukowski and Larsen (2011), we use a two-by-two table design for the distribution of *wanna* contraction, as shown in Table 1.

TABLE 1 Distribution of *wanna* contraction.

	No contraction	Contraction
Object	Who do you want to take to the party?	Who do you wanna take to the party?
Subject	Who do you want to come to the party?	*Who do you wanna come to the party?

Asterisks (*) refer to the ungrammatical sentences.

As shown in Table 1, *wanna* contraction is disallowed in subject extraction questions. In other words, subject extraction questions present illicit conditions for *wanna* contraction, while object extraction questions present licit conditions. Regarding the types of *wh*-words (*who*, *what*, *where*, *when*, and *why*), all *wh*-words except for *who* and *what* were excluded. This exclusion is because *where*, *when*, and *why* only occur as object extraction questions and, therefore, cannot form a contrasting pair (Zukowski and Larsen, 2011).

Based on Table 1, each dataset comprises four conditions, as in Table 2.

Each dataset comprises four sentences, and the fourth one (4) represents an illicit condition for *wanna* contraction, as it involves a subject extraction question with a contraction [+ Contraction, Subject Extraction Question].

We constructed two distinct datasets: Datasets A and B. Each dataset comprises the four conditions outlined in Table 2. In Dataset A, verbs differ between the two question types. Intransitive and transitive verbs were utilized in subject extraction questions and object extraction questions, respectively. This choice is due to the differing extraction sites of *wh*-words in the two question types. In subject extraction questions, *wh*-words are originally positioned between the verb *want* and the infinitival marker *to*. By contrast, they are extracted from a canonical position behind embedded verbs in object extraction questions. We used word frequency data from the Corpus of Contemporary American English (Davies, 2008) to mitigate unintended frequency effects, using the list of the top 60,000 lemmas. From this list, we selected pairs of intransitive and transitive verbs with similar frequencies. In Dataset B, we controlled for embedded verbs, distinguishing between the two question types based on the last token in each sentence. In subject extraction questions, the last token is an adverb, whereas it is a preposition in object extraction questions. We constructed Dataset B because only embedded verbs are masked in Dataset A. Therefore, to diversify our experimental design, we masked adverbs and prepositions in Dataset B. Table 3 provides samples of each dataset.

Regarding the size of our dataset, each dataset comprises 100 sets of *who* question (400 sentences) and 100 sets of *what* question sets (400 sentences). Thus, each dataset includes 200 sets (800 sentences) and, combined, the two datasets total 400 sets (1,600 sentences).

3.1.2 Modeling procedure

As previously mentioned, we utilize both the BERT and RoBERTa models. Specifically, we used the following models: BERT-base-uncased, BERT-large-uncased, RoBERTa-base, and RoBERTa-large. Details of each model are provided in Table 4.

Both BERT-base-uncased and BERT-large-uncased models are case-insensitive. For instance, they do not distinguish between *Word* and *word*. Meanwhile, RoBERTa-base and RoBERTa-large are both case-sensitive,

TABLE 2 Basic format of an individual dataset.

Condition	Contraction	Question type	Grammaticality
(1)	–	Object	Licit
(2)	–	Subject	Licit
(3)	+	Object	Licit
(4)	+	Subject	Illicit

TABLE 3 Exemplary cases of each dataset.

Dataset	Conditions		Sentence
	Question	Contraction	
A (Masking of embedded verbs)	Object	–	Who do you want to **meet** at the dorm?
		+	Who do you wanna **meet** at the dorm?
	Subject	–	Who do you want to **wait** at the dorm?
		+	*Who do you wanna **wait** at the dorm?
	Object	–	What do you want to **show** at the party?
		+	What do you wanna **show** at the party?
	Subject	–	What do you want to **happen** at the party?
		+	*What do you wanna **happen** at the party?
B (Masking of prepositions and adverbs)	Object	–	Who do you want to meet the students **with**?
		+	Who do you wanna meet the students **with**?
	Subject	–	Who do you want to meet the students **today**?
		+	*Who do you wanna meet the students **today**?
	Object	–	What animal do you want to play **with**?
		+	What animal do you wanna play **with**?
	Subject	–	What animal do you want to play **outside**?
		+	*What animal do you wanna play **outside**?

The bolded part refers to the target word that are going to be masked. Asterisks (*) refer to the ungrammatical sentences.

TABLE 4 Details of language models.

Models	Layers	Hidden size	Attention heads	Parameters
BERT-base-uncased	12	768	12	110 M
BERT-large-uncased	24	1,024	16	340 M
RoBERTa-base	12	768	12	125 M
RoBERTa-large	24	1,024	16	355 M

TABLE 5 Data formats for measuring surprisal in Experiment 1.

Dataset	Sentence	Item 1	Item 2
A	Who do you wanna [MASK] at the dorm?	meet	wait
B	Who do you wanna meet the students [MASK]?	with	today

which means that they differentiate between *Word* and *word*. Because the *wanna* contraction is not affected by case, we utilized both case-sensitive and case-insensitive models.

We utilized surprisal values to assess the ability of language models to detect violations of *wanna* contraction. While the term 'surprisal' originally refers to the logarithm of the reciprocal of a probability (Tribus, 1961), it is also used to characterize the informational value of a given event. In this context, surprisal functions as a complexity metric, quantifying the difficulty of processing a given linguistic expression (Hale, 2001; Levy, 2008). Given the inverse relationship between probabilities and their reciprocals, as probabilities approach zero, their reciprocals increase. Essentially, this implies that surprisal values are higher for events with low probabilities.

We employed the cloze test method, in which an appropriate token is required for the masked part of a given sentence. Specifically, we calculated surprisal values for each masked part to identify any processing difficulties encountered by language models (Wilcox et al., 2018; Chaves and Richter, 2021). We masked one specific region of a sentence and provided two different items, as in Table 5.

Then, the language models were required to provide surprisal values for each item. In Dataset A, the probabilities of pairs of transitive verbs (e.g., *meet*) and intransitive verbs (e.g., *wait*) are calculated at [MASK]. The use of intransitive verbs leads to the formation of subject extraction questions. Consequently, when the contraction is applied, using intransitive verbs results in ungrammaticality (e.g., **Who do you wanna wait at the dorm?*). In contrast, the use of transitive verbs leads to the formation of object extraction questions, which are grammatical regardless of the contraction (e.g., *Who do you wanna meet at the dorm?*). In Dataset B, the probabilities of pairs of prepositions (e.g., *with*) and adverbs (e.g., *today*) are calculated at [MASK]. Specifically, the use of adverbs results in the formation of subject extraction questions. Therefore, employing adverbs also leads to ungrammaticality when the contraction is applied (e.g., **Who do you wanna meet the students today?*). In contrast, employing prepositions is grammatical as it leads to the formation of object extraction questions (e.g., *Who do you wanna meet the students with?*).

Based on the formats presented in Table 5, we propose the following hypotheses: The surprisal value of Condition [+ Contraction, Subject Extraction Question] will be higher than that of Condition [+ Contraction, Object Extraction Question]. This expectation arises from the fact that Condition [+ Contraction, Subject Extraction Question] is the only case in which *wanna* contraction is disallowed, making it more likely to be perceived as surprising, as in (3).

(3) a. Who do you wanna meet at the dorm? [+ Contraction, Object Extraction]
b. *Who do you wanna wait at the dorm? [+ Contraction, Subject Extraction]

That is, we expect the mean difference in surprisal between the two question types to be statistically significant when the contraction is applied.

3.2 Results

In summary, we considered a total of 16 cases (8 cases per dataset) in Experiment 1, and 14 of them (87.5%) were consistent with our predictions. That is, the mean difference in surprisal between the two question types was not statistically significant only in two cases. The results of a paired *t*-test based on the surprisal values in Dataset A and Dataset B are provided in Tables 6, 7, respectively. Those consistent with our predictions are highlighted in boldface.

4 Experiment 2: controlling for question types

In Experiment 2, we conduct a comparison between Conditions [− Contraction, Subject Extraction Question] and [+ Contraction, Subject Extraction Question]. The comparison ensures control over question types (subject extraction vs. object extraction). Thus, what makes Experiment 2 different from Experiment 1 is that the key focus lies in the contraction (*want to* vs. *wanna*). By controlling for question types, Experiment 2 exclusively tests the effect of the presence or absence of contraction.

4.1 Methods

4.1.1 Materials

The materials used in Experiment 1 are also used in Experiment 2. We use Datasets A and B again. However, in Experiment 2, we used the whole sentences as an input for the cloze test instead of masking one specific part of the sentence, as illustrated in the following section.

TABLE 6 Surprisal significance for Dataset A (Experiment 1).

Question type	Model		Paired *t*-test	
Who	BERT	base	**t = −6.4906**	**p = 3.421e-09** (***)
		large	**t = −5.1981**	**p = 1.084e-06** (***)
	RoBERTa	base	**t = −5.1981**	**p = 1.084e-06** (***)
		large	**t = −4.2465**	**p = 4.907e-05** (***)
What	BERT	base	*t* = 1.1701	*p* = 0.2448 (NS)
		large	**t = −2.2203**	**p = 0.02868** (*)
	RoBERTa	base	**t = −2.6846**	**p = 0.008514** (***)
		large	*t* = −1.8297	*p* = 0.0703 (NS)

The bolded part refers to the target word that are going to be masked. The symbol (***) indicates that the p-value is smaller than 0.001. Asterisk (*) indicate that the p-value is smaller than 0.05.

TABLE 7 Surprisal significance for Dataset B (Experiment 1).

Question type	Model		Paired *t*-test	
Who	BERT	base	**t = −28.424**	**p < 0.001** (***)
		large	**t = −26.14**	**p < 0.001** (***)
	RoBERTa	base	**t = −19.734**	**p < 0.001** (***)
		large	**t = −25.948**	**p < 0.001** (***)
What	BERT	base	**t = −19.171**	**p < 0.001** (***)
		large	**t = −20.394**	**p < 0.001** (***)
	RoBERTa	base	**t = −18.791**	**p < 0.001** (***)
		large	**t = −20.984**	**p < 0.001** (***)

The bolded part refers to the target word that are going to be masked. The symbol (***) indicates that the *p*-value is smaller than 0.001.

TABLE 8 Data formats for measuring surprisal in Experiment 2.

Dataset	Item 1	Item 2
A	Who do you want to wait at the dorm?	*Who do you wanna wait at the dorm?
B	Who do you want to meet the students today?	*Who do you wanna meet the students today?

The bolded part refers to the target word that are going to be masked. Asterisk (*) refer to the ungrammatical sentences. Asterisks indicate that the *p*-value is smaller than 0.05.

4.1.2 Modeling procedure

We employ the same language models used in Experiment 1: BERT-base-uncased, BERT-large-uncased, RoBERTa-base, and RoBERTa-large. These models are used again for the implementation of the cloze test method. However, instead of masking one specific region of a sentence (embedded verbs and prepositions/adverbs), we measure the surprisal by using the whole sentences as an input. This is because the exact number of tokens in each sentence varies depending on the use of the contraction. When *wanna* contraction is applied, the number of tokens decreases by one compared to sentences without the contraction. Therefore, we measured the by-word surprisal values for each token and calculated the mean for each sentence to compare sentences with and without *wanna* contraction, as shown in Table 8.

In Dataset A, we use the sentences with intransitive verbs (e.g., *wait*) as an input. Using intransitive verbs leads to the formation of subject extraction questions. Consequently, when the contraction is applied, the use of intransitive verbs results in ungrammaticality (e.g., **Who do you wanna wait at the dorm?*). In Dataset B, we use the sentences ending with adverbs (e.g., *today*) as an input, and this leads to the formation of subject extraction questions. Therefore, applying the contraction to sentences with adverbs also results in ungrammaticality (e.g., **Who do you wanna meet the students today?*).

Based on this format, we propose the following hypotheses: The surprisal value for Condition [+ Contraction, Subject Extraction Question] will be higher than that for Condition [− Contraction, Subject Extraction Question]. This expectation arises from the fact that *wanna* contraction is not allowed in subject extraction questions and is therefore more likely to be perceived as surprising, as illustrated in (4).

(4) a. Who do you want to wait at the dorm? [− Contraction, Subject Extraction]
b. *Who do you wanna wait at the dorm? [+ Contraction, Subject Extraction]

That is, we predict a statistically significant mean difference in surprisal between the two question types when contraction is applied in subject extraction questions.

4.2 Results

In summary, we considered a total of 16 cases (8 cases per dataset) in Experiment 2, and 12 of them (75%) were consistent with our predictions. That is, the mean difference in surprisal

TABLE 9 Surprisal significance for Dataset A (Experiment 2).

Question type	Model		Paired *t*-test	
Who	BERT	base	**t = −19.322**	**p < 2.2e-16**
		large	**t = −15.548**	**p < 2.2e-16**
	RoBERTa	base	*t* = 7.708	*p* = 1.008e-11
		large	**t = −29.726**	**p < 2.2e-16**
What	BERT	base	**t = −25.586**	**p < 2.2e-16**
		large	**t = −22.367**	**p < 2.2e-16**
	RoBERTa	base	*t* = 5.0623	*p* = 1.916e-06
		large	**t = −27.567**	**p < 2.2e-16**

The bolded part refers to the target word that are going to be masked.

TABLE 10 Surprisal significance for Dataset B (Experiment 2).

Question type	Model		Paired *t*-test	
Who	BERT	base	**t = −38.502**	**p < 2.2e-16**
		large	**t = −23.719**	**p < 2.2e-16**
	RoBERTa	base	*t* = 5.4118	*p* = 4.355e-07
		large	**t = −29.979**	**p < 2.2e-16**
What	BERT	base	**t = −24.889**	**p < 2.2e-16**
		large	**t = −21.191**	**p < 2.2e-16**
	RoBERTa	base	*t* = −1.4395	*p* = 0.1532
		large	**t = −29.294**	**p < 2.2e-16**

The bolded part refers to the target word that are going to be masked.

between the two question types was not statistically significant in four cases. Specifically, only the RoBERTa-base model deviated from our predictions. The results of a paired *t*-test based on the surprisal values in Dataset A and Dataset B are provided in Tables 9, 10, respectively. Those consistent with our predictions are highlighted in boldface.

5 Experiment 3: measuring by-word surprisal

In Experiment 3, we measure the by-word surprisal values for every sentence region, in addition to measuring the surprisal values for a specific sentence region. This expansion is required because the previous experiment designs are confined to either embedded verbs (Dataset A) or prepositions/adverbs (Dataset B), preventing the drawing of fully conclusive interpretations. Therefore, the purpose of Experiment 3 is to examine whether sentence regions other than embedded verbs and prepositions/adverbs exhibited noticeable differences.

5.1 Methods

5.1.1 Materials

From Dataset A, we selected 50 *who* questions with prepositional phrases (e.g., *Who do you want to meet at the party?*). The other half of Dataset A was excluded as it comprised *who* questions with either adverbs (e.g., *Who do you want to help immediately?*) or adverbial phrases (e.g., *Who do you want to contact the most?*). This exclusion was due to the non-ideal nature of comparing single-word adverbs (e.g., *immediately*) with multiple-word adverbial phrases (*the most*) for measuring by-word surprisal values. In addition, controlling for the first region of adverbs or adverbial phrases was not possible. The first region of adverbial phrases encompasses various parts of speech (e.g., *right away* vs. *the most*), unlike prepositional phrases, in which the first region is always a preposition.

5.1.2 Modeling procedure

We employ the same language models used in Experiment 1 and Experiment 2: BERT-base-uncased, BERT-large-uncased, RoBERTa-base, and RoBERTa-large. These models are again used for the implementation of the cloze test method for each sentence region.

5.2 Results

Using the selected 50 sentences comprising prepositional phrases from Dataset A, each sentence region was masked to calculate the mean of the surprisal values for each region. Figure 1 visualizes the mean by-word surprisal values in each sentence region.

Figure 1 illustrates the contrast between subject extraction questions and object extraction questions, particularly concerning the regions *Who*, *wanna*, and VERB. The regions *Who* and VERB consistently show that by-word surprisal values are higher in subject extraction questions regardless of contraction. However, the region *wanna* indicates that both the BERT and RoBERTa models managed to produce higher surprisal values for subject extraction questions. Focusing on the regions *want to* and *wanna*, the results of a paired *t*-test are presented, as in Table 11.

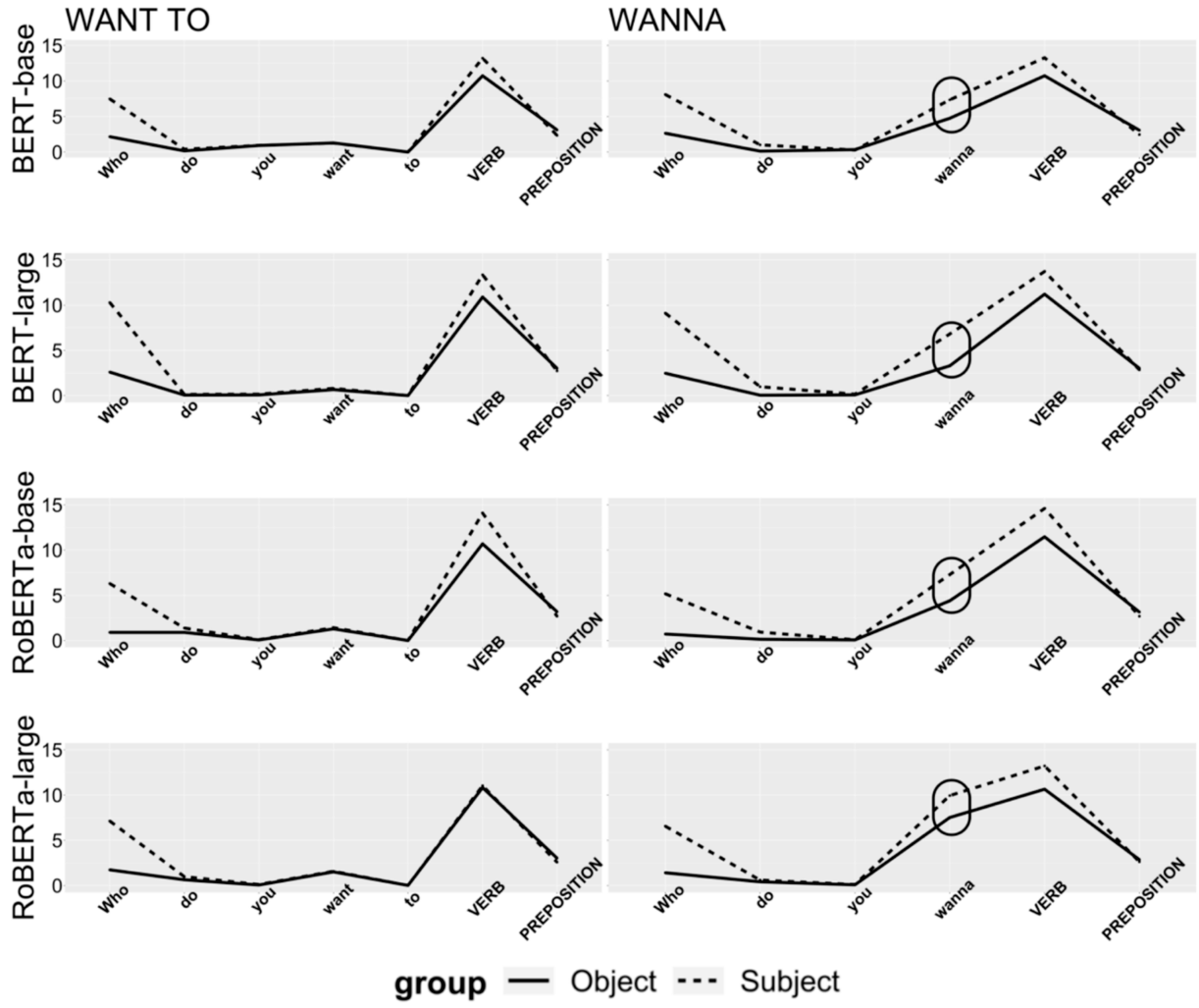

FIGURE 1
Mean by-word surprisal values in sentence regions.

TABLE 11 Surprisal significance for the regions *wanna* and *want to*.

Question type	Model		Paired *t*-test	
wanna	BERT	base	**t = 4.5658**	**p = 3.367e-05** (***)
		large	**t = 5.9236**	**p = 3.069e-07** (***)
	RoBERTa	base	**t = 4.7928**	**p = 1.568e-05** (***)
		large	**t = 5.073**	**p = 6.02e-06** (***)
want	BERT	base	$t = -0.15718$	$p = 0.8757$
		large	$t = 0.94407$	$p = 0.3498$
	RoBERTa	base	$t = 0.62979$	$p = 0.5318$
		large	$t = 0.2931$	$p = 0.7707$
to	BERT	base	$t = 1.22$	$p = 0.2283$
		large	**t = 2.1314**	**p = 0.03809** (*)
	RoBERTa	base	$t = -0.78321$	$p = 0.4373$
		large	$t = -1.7129$	$p = 0.09305$

The bolded part refers to the target word that are going to be masked. The symbol (***) indicates that the *p*-value is smaller than 0.001. Asterisk (*) indicate that the *p*-value is smaller than 0.05.

Table 11 demonstrates that the mean difference in surprisal between the two question types at the region *wanna* is statistically significant in every model. However, in contrast to the region *wanna*, the means of the surprisal values for the regions *want* and *to* show no statistically significant difference except for one case.

6 Discussion

We now address our research questions. First, do language models capture the grammatical constraint underlying *wanna* contraction? Second, are there additional factors influencing language models' processing of *wanna* contraction? Here are our answers to these two questions.

Our answer to the first question is as follows: The BERT and RoBERTa models do identify illicit cases of *wanna* contraction. Given that *wanna* contraction is not allowed in subject extraction questions, we hypothesized that the surprisal values for subject extraction questions would be higher than those for object extraction questions when the contraction was employed. Thus, the mean difference in surprisal between the two question types was expected to be statistically significant when contraction was present. The results reveal that a substantial portion of Experiment 1 (87.5%) was in line with our predictions. Then, in Experiment 2, we solely considered the existence of contraction (*want to* vs. *wanna*) to prevent the effect of question types (subject vs. object). By using only subject extraction questions, we expected the surprisal values for the questions with contractions to be higher than those without contractions. The results show that a substantial portion of Experiment 2 (75%) was in line with our expectations.

Our answer to the second question is that factors such as subject–object asymmetry and statistical sparsity may play a role. Subject–object asymmetry is inherently related to *wanna* contraction as question types (subject vs. object) decide whether the contraction is permitted or not. Statistical sparsity also needs to be considered as the form *wanna* is less common in corpus data than the form *want to*. Nonetheless, in the following two sections, we argue that the present study reveals the language models' syntactic sensitivity to *wanna* contraction despite these intervening factors.

6.1 Subject–object asymmetry

The findings from Experiment 3 unmistakably indicate that, when processing sentences with *wanna* contraction, neural language models are markedly influenced by the extraction site of *wh*-words, specifically subject–object asymmetry in *wh*-movement. The surprisal values for subject extraction questions were generally higher than those for object extraction questions. Given that the distinction between subject extraction questions and object extraction questions involves subject–object asymmetry, it is possible that this asymmetry influenced our results. It is worth noting that a consistent preference for object extraction over subject extraction has been observed in both L1 and L2 speakers (Schachter and Yip, 1990). Furthermore, research has shown that English L1 children tend to favor object extraction over subject extraction (Stromswold, 1995). For instance, in the experiment, all 11 participating children produced object extraction questions such as (5a), whereas only one out of the 11 produced subject extraction questions such as (5b).

(5) a. Who do you think Mary invited ____ to the party? [Object Extraction]
b. Who do you think ____ invited Bill to the party? [Subject Extraction]
(Stromswold, 1995:40)

Processing subject extraction questions is also challenging for L2 learners. For instance, both L1 English speakers and Chinese EFL learners experienced greater difficulty in processing (6a) as compared to (6b), requiring additional reaction time for the former during judgment tasks (Juffs and Harrington, 1995).

(6) a. Who does Tom expect ____ to fire the manager? [Subject Extraction]
b. Who does Tom expect to fire ____? [Object Extraction]
(Juffs and Harrington, 1995:496)

This contrast between the two types of gaps is relevant to *wanna* contraction because the contraction is not permitted when subjects are extracted. Subject–object asymmetry has been suggested to influence L2 learners' processing of *wanna* contraction (Ito, 2018).

Our datasets also distinguish between subject and object extractions, as illustrated in (7).

(7) a. *Who do you wanna wait at the dorm? [Subject Extraction]
b. Who do you wanna meet at the dorm? [Object Extraction]

In (7a), *who* is extracted from the subject position, whereas in (7b), *who* is extracted from the object position within the infinitival phrase. The findings consistently demonstrate that both the BERT and the RoBERTa models generated higher surprisal values for subject extraction questions such as (7a), regardless of the contraction. Higher surprisal values indicate increased difficulty in processing unexpected sentences, suggesting that subject extraction questions were indeed unexpected and posed greater challenges for both models.

However, despite the effect of subject–object asymmetry, the findings from Experiment 3 reveal that the language models do distinguish licit cases of *wanna* contraction from illicit ones. The mean difference in surprisal between the two question types was always statistically significant in the region *wanna* but not in the regions *want to*. In other words, both language models exhibited a meaningful degree of syntactic sensitivity to *wanna* contraction when we measured mean by-word surprisal values, effectively mitigating the impact of subject–object asymmetry in embedded verb positions.

6.2 Statistical sparsity

Despite English-speaking children's processing of *wanna* contraction, it seems to be less common in speech data compared

to the *want to* counterpart. For instance, utterances with *wanna* contraction were less common in adult speech from the CHILDES database compared to those without *wanna* contraction (Zukowski and Larsen, 2011). It is also worth noting that *wanna* contraction is not frequently found in the corpus data. For instance, there is a stark contrast between the two forms *want to* and *wanna* in the Corpus of Contemporary American English (COCA, Davies, 2008). While the form *want to* is found 628,967 times throughout the corpus, the form *wanna* is found only 78,858 times in total. A similar contrast is also found in the Wikipedia Corpus (Davies, 2015), on which both BERT and RoBERTa models were pretrained. Even though the form *want to* appears 63,434 times throughout the corpus, the form *wanna* appears only 4 times. Given this contrast, we can conclude that the language models show syntactic sensitivity to the illicit cases of *wanna* contraction despite the statistical sparsity of the phenomenon.

7 Conclusion

The primary contribution of this study lies in our assessment of whether the BERT and RoBERTa models accurately capture the grammatical constraint underlying *wanna* contraction. Considering that *wanna* contraction involves complex linguistic properties such as *wh*-movement, subject/object extractions, and contraction, examining the neural language models' capability to manage this phenomenon can provide insights into how effectively they process natural languages, similar to human speakers. In Experiments 1 and 2, both the BERT and the RoBERTa models were largely in line with our initial predictions (87.5% in Experiment 1 and 75% in Experiment 2). The measurement of by-word surprisal values in Experiment 3 reveals that both language models yielded higher surprisal values in subject extraction questions for the sentence region *wanna* but not for *want to*. This contrast leads us to argue that the BERT and RoBERTa models demonstrate syntactic sensitivity to *wanna* contraction despite the impact of subject–object asymmetry and statistical sparsity of this phenomenon.

One potential limitation of our study pertains to the influence of our datasets on the results. Our dataset design primarily focuses on comparing surprisal values between subject extraction questions and object extraction questions. Consequently, *wh*-questions containing *wanna* contraction within these datasets might be infrequently used by L1 English speakers, making them challenging to locate in corpus data. Therefore, the scarcity of question forms featuring *wanna* contraction may have had an impact. In addition, there is a need to test artificial learners other than BERT or any other state-of-the-art language models. Specifically, it would be interesting to explore whether the constraint underlying *wanna* contraction can be captured by language models that are equipped with a moderate amount of data. The BERT model was trained with a massive amount of input data (approximately 3.3. billion words), making it difficult to directly compare it with human learners. In other words, we need to employ language models that are comparable to human counterparts in terms of the amount of input data.

Despite the abovementioned limitations, we believe that the present study has demonstrated that neural language models, such as BERT, can serve as useful tools for both theoretical and computational linguists to simulate how natural language is processed. In addition, we hope to contribute a valuable empirical investigation into *wanna* contraction, a research focus that has targeted L1 and L2 speakers, offering novel experimental data produced by artificial learners.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KN: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. EO: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. SS: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Research Foundation (NRF), Korea, under the project BK21 FOUR (4299990414427).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2024.1442093/full#supplementary-material

References

Chaves, R. P., and Richter, S. N. (2021). Look at that! BERT can be easily distracted from paying attention to morphosyntax. *Proc. Soc. Comput. Lingu.* 4, 28–38. doi: 10.7275/b92s-qd21

Chomsky, N. (1957). Syntactic structures. Berlin: de Gruyter.

Chomsky, N. (1980). Rules and representations. *Behav. Brain Sci.* 3, 1–15.

Chomsky, N., Culicover, P. W., Wasow, T., and Akmajian, A. (1977). "On wh-movement" in Formal syntax. eds. P. Cullicover, T. Wasow and A. Akmajian (New York: Academic Press), 71–132.

Chomsky, N., and Lasnik, H. (1977). Filters and control. *Linguist. Inq.* 8, 425–504.

Crain, S., and Pietroski, P. (2001). Nature, nurture, and universal grammar. *Linguist. philos.* 24, 139–186. doi: 10.1023/A:1005694100138

Davies, M.. (2008). The Corpus of contemporary American English (COCA). Available at: https://www.english-corpora.org/coca/ (Accessed March 30, 2023).

Davies, M.. (2015). The Wikipedia Corpus. Available at: https://www.english-corpora.org/wiki/ (Accessed March 30, 2023).

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K.. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies, Minneapolis, Minnesota. Association for Computational Linguistics.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. Available at: https://arxiv.org/abs/1901.05287 (Accessed May 5, 2023).

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model: In second meeting of the north american chapter of the association for computational linguistics. Available at: https://aclanthology.org/N01-1021 (Accessed May 10, 2023).

Ito, Y. (2018). Acquisition of Contraction Constraint by Japanese learners of English. *J. Pan-Pac. Assoc. Appl. Linguist.* 22, 19–41. doi: 10.25256/PAAL.22.1.2

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th annual meeting of the Association for Computational Linguistics. 3651–3657. Florence, Italy. Association for Computational Linguistics.

Juffs, A., and Harrington, M. (1995). Parsing effects in second language sentence processing. *Stud. Second. Lang. Acquis.* 17, 483–516. doi: 10.1017/S027226310001442X

Kweon, S. O. (2000). The acquisition of English contraction constraints by advanced Korean learners of English: Experimental studies on wanna contraction and auxiliary contraction. Honolulu, HI: University of Hawaii at Manoa.

Kweon, S. O., and Bley-Vroman, R. (2011). Acquisition of the constraints on wanna contraction by advanced second language learners: universal grammar and imperfect knowledge. *Second. Lang. Res.* 27, 207–228. doi: 10.1177/0267658310375756

Lakoff, G. (1970). Global rules. *Language* 46, 627–639. doi: 10.2307/412310

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lightfoot, D. (1976). Trace theory and twice-moved NPs. *Linguist. Inq.* 7, 559–582.

Linzen, T., and Baroni, M. (2021). Syntactic structure from deep learning. *Annu. Rev. Linguist.* 7, 195–212. doi: 10.1146/annurev-linguistics-032020-051035

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist.* 4, 521–535. doi: 10.1162/tacl_a_00115

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., and Chen, D., (2019). Roberta: a robustly optimized bert pretraining approach. Available at: https://arxiv.org/abs/1907.11692.

O'Grady, W., Nakamura, M., and Ito, Y. (2008). Want-to contraction in second language acquisition: an emergentist approach. *Lingua* 118, 478–498. doi: 10.1016/j.lingua.2007.01.006

Postal, P. M., and Pullum, G. K. (1982). The contraction debate. *Linguist. Inq.* 13, 122–138.

Rezaeian, M., Sadighi, F., Yamini, M., and Bagheri, M. S. (2017). Investigating "wanna" contraction through an emergentist approach among Iranian EFL learners applying usage-based model of language acquisition. *Lingua* 196, 55–73. doi: 10.1016/j.lingua.2017.06.008

Ross, J. R. (1967). Constraints on variables in syntax. Cambridge: Massachusetts Institute of Technology.

Rotenberg, J. (1978). The syntax of phonology. Cambridge: Massachusetts Institute of Technology.

Schachter, J., and Yip, V. (1990). Grammaticality judgments: why does anyone object to subject extraction? *Stud. Second. Lang. Acquis.* 12, 379–392. doi: 10.1017/S0272263100009487

Stromswold, K. (1995). The acquisition of subject and object wh-questions. *Lang. Acquis.* 4, 5–48. doi: 10.1080/10489223.1995.9671658

Thornton, R. (1990). Adventures in long distance moving: the acquisition of complex wh-questions. Mansfield, CT: University of Connecticut.

Tribus, M. (1961). Information theory as the basis for Thermostatics and thermodynamics. *J. Appl. Mech.* 28, 1–8. doi: 10.1115/1.3640461

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP, Stroudsburg: Association for Computational Linguistics.

Zukowski, A., and Larsen, J. (2011). Wanna contraction in children: retesting and revising the developmental facts. *Lang. Acquis.* 18, 211–241. doi: 10.1080/10489223.2011.605043