# How are early cue effects in Chinese relative clauses reflected in Chinese pre-trained language model?

Anonymous ACL submission

## 1 Introduction

The current project aims at reproduce the early cue effects found in the Chinese relative clause (RC) comprehension study conducted by Wu et al. (2018). In their experiments, the authors take advantage of the pre-nominal characteristic of Chinese RCs to test two competing sentence processing accounts: the working memory-based account and the experience-based account.

Gibson's (1998; 2000) working memory-based Dependency Locality Theory (DLT) predicts sentence processing based on two factors: the storage cost and the integration cost. The former is the memory used to maintain syntactic heads that are needed to establish a phrasal-structural dependency. The latter is the memory consumed to integrate a word into a dependency. The storage cost predicts that the more incomplete dependencies are there, the more difficult the processing is. The integration cost predicts that the longer distance the head and its dependent in a dependency has, the higher the cost is.

The experience-based account has many variants and one of them is the surprisal theory (Hale, 2001, 2003; Levy, 2008). It defines the comprehension difficulties as the surprisal of critical regions (negative log probability of that region). According to this type of theories, a reader, while reading a sequence, keeps updating probabilities of upcoming structures (their expectation) based on the part of sequence that is already seen. If their expectation is confirmed by the word they see (i.e., the structure with a higher probability is confirmed), the surprisal of seeing that word will be low and the sequence will be processed faster. On the other hand, if the actual structure is a low probable structure, the surprisal will be high and processing becomes harder.

Wu et al. (2018) conducted two self-paced reading experiments using Chinese RC. They employed two elements, a classifier and a passive marker, which can be added to the left edge of a RC. The two elements inevitably increases the dependencies in a RC. However, at the same time, they are strong indicators of a RC parsing. The two elements create an arena for the two processing accounts. The experiment results from Wu et al. turned out to be consistent with the experience-based accounts. Adding the two RC parsing indicators facilitate the processing.

The current paper conducts an experiment similar to Wu et al. (2018). The difference is that, instead of being a human experiment, the experiment uses two pre-trained Chinese language models (LM). It turns out that the major effects detected in Wu et al.'s experiment also showed up in the current experiment. However, the exact parts of sentences that they occur are different.

The paper is organized as follows. Section 2 introduces the background and Experiment 2 in Wu et al. (2018). Section 3 connects the experience-based accounts to the surprisal theory and probabilities in LMs. Section 4 describes the experiment conducted in the current project in detail. Section 5 reports the statistical results. Section 6 discusses and compares the results of the current study and Wu et al.'s. Section 7 reports the results of the same experiment ran on a LSTM, followed by a short comparison of the LM performance and the human results. Section 8 concludes.

## 2 Literature review

### 2.1 Wu et al. (2018): early cue effects in Chinese relative clauses

#### 2.1.1 Chinese relative clauses and ambiguity

In the research of sentence processing in pyscholinguistics, RCs are of special interest because they pose challenges for comprehension that simple sentences do not. When reading simple sentences, most dependencies are resolved immediately. For example in (1a), the subject and verb are directly next to each other. In (1b), the matrix verb and its object argument are adjacent.

(1) a. The reporter looked about his surroundings in distress.
b. The stone hit the reporter.

Comprehending a RC, for example (2), can be harder. The reader needs to recognize the RC boundaries (brackets in (2)) and resolve a filler-gap dependency, a relationship between the RC head noun (i.e., *reporter*$_i$) and its corresponding position in the RC (i.e., ___ $_i$).

(2) The reporter$_i$ [$_{RC}$ that the stone hit ___ $_i$] looked about his surrounding in distress.

To process a Chinese RC, a reader also needs to recognize clause boundaries and to solve a filler-gap dependency. However, clause boundary recognition can be difficult because Chinese RCs are prenominal. In (3), the bracketed RC stands before the RC head noun in bold.

(3) [$_{RC}$ shi-kuai za-zhong ___ $_i$ de ]
stone hit Dᴇ reporter
**ji-zhe**$_i$ ao-sang-de huan-gu si-zhou
distressfully look-about surroudings
"The reporter that the stone hit looked about his surroundings in distress."
(Wu et al., 2018, (5a))

If (3) is read word by word, not until the head noun is it clear that (3) is a RC because *shi-kuai za(-zhong)* (stone hit) can continue as a matrix clause (4a), or as a noun-complement structure (4b).[1]

(4) a. shi-kuai za-zhong le wo
stone hit Asp. I
"The stone hit me."
b. shi-kuai za-zhong de ji-lv ...
stone hit Dᴇ odds
"The odds to be hit by a stone ..."

### 2.1.2 Early cues in Chinese relative clauses

There are ways to eliminate the ambiguities in a RC, for example, by adding elements that strongly indicate a RC structure. Wu et al. (2018) chose mis-matching classifiers and the passive marker `bei`. They are added to the left edge of RCs as the early cues. Examples are in (5-7) (adapted from Wu et al., 2018).[2] In (5), a demonstrative-classifier pair is added before the RC. For the purpose here, it is sufficient to know that i) classifiers often co-occur with demonstratives; ii) `wei` is a classifier for human only; iii) the noun that the classifier modifies comes after the classifier.

(5) Classifier

`na` `wei` [$_{RC}$ shi-kuai za-zhong de] ji-zhe
Dᴇᴍ Cʟ stone hit Dᴇ reporter
(Wu et al., 2018, (5b))

The classifier in (5) is a mis-matching one because the noun that occurs after it, *shi-kuai* (stone), is not a human noun phrase. This causes lexical disruption but at the same time raises readers' expectation of seeing a human noun phrase later, which is very likely to be the head noun of a RC.

The second early cue is `bei` as in example (6).

(6) `bei`

[$_{RC}$ `bei` shi-kuai za-zhong de] ji-zhe
Pᴀss stone hit Dᴇ reporter
(Wu et al., 2018, (5c))

The passive marker `bei` minimally requires a verb and a patient. It takes optionally an agent or instrument. In a matrix clause, the patient occurs before `bei`, and the agent/instrument, if present, occurs right after `bei`. When seeing `bei` in (6), the possibility of a relative clause increases because `bei` has to take a patient, which is missing before `bei` and in the sequence *bei shi-kuai za-zhong de* (Pᴀss stone hit Dᴇ).

The two early cues can occur together, as in (7).

(7) Classifier + `bei`

`na` `wei` [$_{RC}$ `bei` shi-kuai za-zhong de]
Dᴇᴍ Cʟ Pᴀss stone hit Dᴇ
ji-zhe
reporter
(Wu et al., 2018, (5d))

In this configuration, the classifier cannot be in a dependency with the noun that comes after `bei` because that is not allowed in Chinese. When encountering the sequence `na wei bei` (Dᴇᴍ Cʟ

[1] There are even more possible ambiguities in a Chinese RC. For a more detailed analysis of ambiguity in Chinese RCs, please see Jäger et al. (2015).

[2] All three sentences express the same proposition as (3). The second half of the sentences, which are exactly the same as in (3), is omitted.

Pass) together, the reader anticipates a noun phrase which matches the classifier and is the obligatory patient of the passivized verb. If no ellipsis is involved, the only way to continue the sequence is to form a relative clause.[3]

### 2.1.3 Predictions by the two accounts

The DLT and the experience-based accounts make different predictions concerning the presence and absence of the early cues.

For DLT, because adding a mis-matching classifier creates an additional dependency in a RC like (5), the storage cost increases. With the classifier being present, the integration cost also increases because the intervening RC delays the accomplishment of the classifier-noun dependency. With `bei` being added after the classifier, the integration cost is even higher because the linear distance between the classifier and the noun is made larger. According to DLT, RCs without any early cues should be processed the fastest, and those with both cues are processed the slowest. In between are the RCs with one early cue.

For the experience-based accounts, the first prediction is that a reader will have difficulty processing the mis-matching classifier-noun pair because the probability of seeing a non-human noun is much lower than seeing a human one based on the experience. This is verified in the corpus study in Wu (2011) and Wu et al. (2018). In the former, it is reported that the mis-matching pattern is 'virtually non-existent'. In the latter, no mis-matching classifier-noun construction is found in the 311 transitive relative clauses extracted from the Chinese Treebank 5.0 corpus (Palmer et al., 2005). The second prediction is that, after seeing a mis-matching classifier and/or `bei`, the reader would have less difficulty processing RC De and head noun, and probably also the regions after that because of the spillover effect.[4] Within the experience-based accounts, RCs with both early cues should be processed the fastest, followed by the ones with only one early cue. The slowest is the RCs with no early cues.

### 2.1.4 Human experiment results

Wu et al. (2018) conducted two self-paced reading experiments with the same target sentences. The difference between the two experiments was that the subject relative clause filler items in the first experiment were substituted with non-relative clauses in the second experiment. The consideration was that high frequency of subject relative clauses might facilitate the processing of the `bei` structure in the target sentences. The results of both experiments were consistent with the predictions made by the experience-based theory but against DLT. The result of the second experiment is reported below.

To analyse the data, Wu et al. (2018) divided their target sentencess into seven regions (8) and aligned them across four conditions (9).

(8) a. RCNP: RC internal noun phrase
b. RCV: RC internal verb
c. De
d. headnoun: RC head noun
e. Adv: matrix clause adverb
f. MainV: matrix clause verb
g. MainO: matrix clause object

(9) a. - `Cl` - `bei`
b. + `Cl` - `bei`
c. - `Cl` + `bei`
d. + `Cl` + `bei`

(The minus sign stands for *without*; the plus sign stands for *with*.)

The reading times were reciprocal-transformed. For each region, a linear mixed-effects model was fitted using the lme4 package in R.[5] The results are summarized in Figure 1 (Wu et al., 2018, Figure 2). Overall, the detected effects demonstrated the spillover effect.

As can be seen in the upper plot, the main effects of the mis-matching classifier were found in

[3] This is confirmed in the norming study in Wu et al. (2018). When being provided the sequence, for example (i), 94.77% of the time the participants finished the sentence as a relative clause. Among the rest of the completed sentences, 1.44% are main clauses with elided noun phrase between the classifier and `bei` (e.g., (ii)), and 2.88% of them are error continuations.

(i) `na` `wei` `bei` shi-kuai za-zhong ______.
Dem Cl Pass stone hit
(Wu et al., 2018, (6d))

(ii) `na` `wei` `bei` shi-kuai za-zhong le tou
Dem Cl Pass stone hit Asp head
"That person's head was hit by a stone."

[4] The easiness and difficulty of processing a critical region might persist longer and be reflected in the regions after the critical region (Ehrlich and Rayner, 1983; Smith and Levy, 2013; Remington et al., 2018, among others).

[5] For information of the exact version of the package and R, and how the data were pre-processed, please refer to Wu et al. (2018, sec. 4.1.4).

Adv, MainV, and MainO regions. The main effects of the passive marker `bei` were pervasive except in the RCNP and MainO region. An interaction of the classifier and the passive marker was only marginally found in MainO.

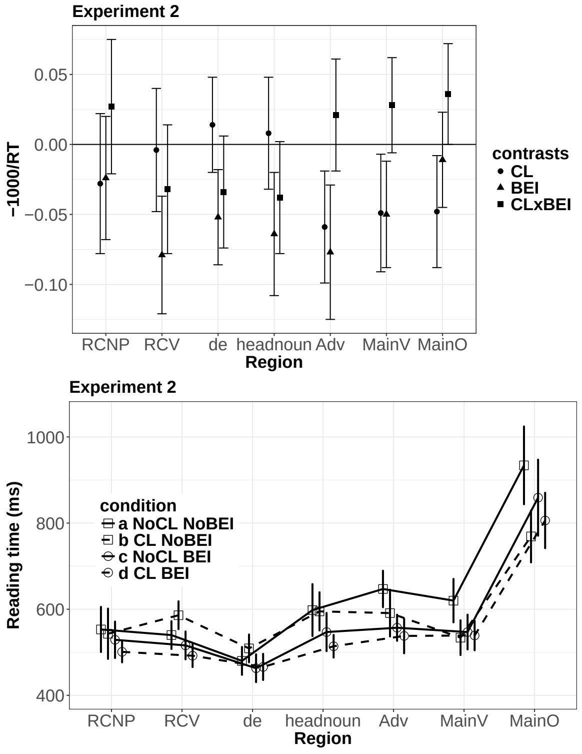


Figure 1: Result summary of the second experiment in Wu et al. (2018)

It is found that the mis-matching classifier and passive marker exhibited facilitatory effects. The facilitatory effects caused by the classifier showed up in the matrix clause regions while `bei` facilitated processing early within the relative clause.

## 3 Experience, probability and LMs

Among the variants of the experience-based accounts of sentence processing, the surprisal theory is of special interest to the current project. Surprisals are defined as the negative log probability of the current word conditioned on the preceding words, which is similar to the ways how large pre-trained LMs are trained. For exmaple, LMs with the GPT-2 structure (Radford et al., 2019) are trained to predict the next word given a prefix based on probabilities. So is a long short-term memory (LSTM) recurrent neural network (RNN). If the surprisal theory is the right approach to account for the early cue effects in Wu et al. (2018), we would expect to see similar effects in pre-trained Chinese LMs.

The idea to use LMs to reproduce human experiment results is not new. It has been done by Van Schijndel and Linzen (2021) to predict garden path effects. The authors trained their own LM and use it to estimate word-by-word probability. The model is a two-layered long short-term memory (LSTM) recurrent neural network (RNN) trained on an 80 million word subset of English Wikipedia and a soap opera dialog corpus with 100 million words (Davies, 2011). The probability is then converted into surprisal. Using the surprisal and the reading time of each word to fit a linear model, the authors find that although surprisal correctly detected the garden path effects but it significantly underestimated the magnitude.

The goal of the current project is similar to Van Schijndel and Linzen's but with two differences. The first difference is that the current project does not focus on the magnitude of the early cue effects. The question is rather whether the presence/absence of the early cues affect the surprisals of a sentence in a statistically significant way. The second difference is that the LM used in the current project is much larger and has a more powerful training structure. Details are provided in the next section.

## 4 Early cue effects in LMs

This section first introduces the LM used in the project. It then presents the experiment process followed by the results.

### 4.1 LM: gpt2-chinese-cluecorpussmall

The pre-trained LM used in the current project is gpt2-chinese-cluecorpussmall (Xu et al., 2020).[6] It is trained on CLUECorpusSmall corpus which has 14G data with 5 billion Chinese characters. The data are crawled online and cover a wide variety of topics and genres. The model is trained using an assemble-on-demand pre-training toolkit Universal Encoder Representations (UER) (Zhao et al., 2019). The concrete model architecture is the same as in GPT2 (Radford et al., 2019). It is a decoder-only transformer with 12 masked self-attention heads in each block (Vaswani et al., 2017). Twelve blocks are stacked on each other.

[6] The model can be accessed from `https://github.com/Morizeyao/GPT2-Chinese` or via HuggingFace `https://huggingface.co/uer/gpt2-chinese-cluecorpussmall`.

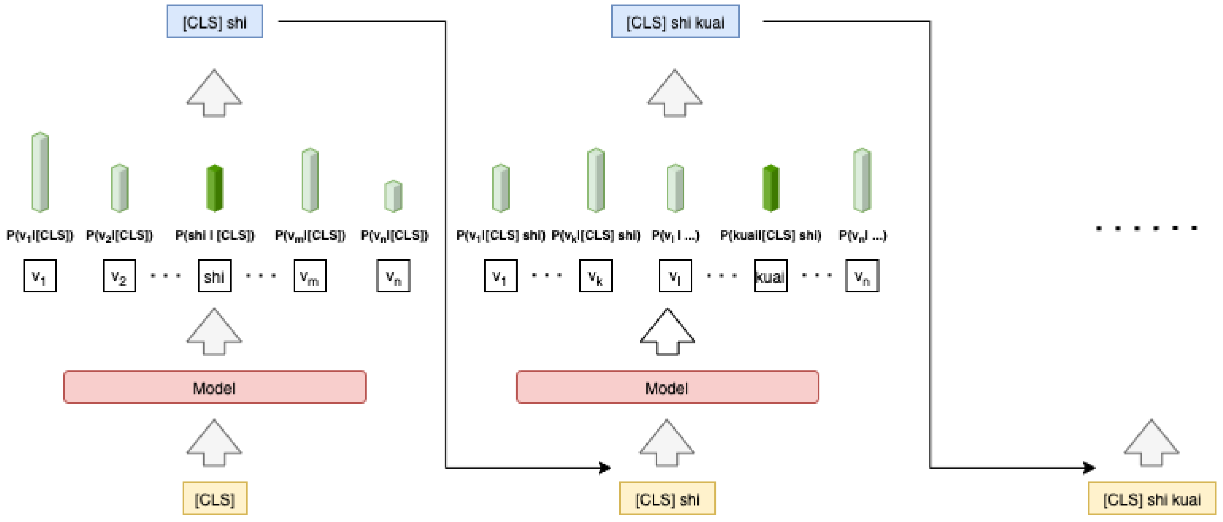


Figure 2: Illustration of target sentence generation process

The model has 102 million parameters. The reason to choose a model trained on this kind of architecture is that the training and predicting processes resemble the self-paced reading setting. That means that only the previous seen words are used to predict the next word. Another popular model architecture is BERT (Devlin et al., 2018) which is trained to predict a masked word given the context before and after the word. This training objective is different from the surprisal definition and the self-paced reading experiment setting. For these reasons, the project uses the gpt2-chinese-cluecorpussmall.

### 4.2 Sentence generation and probability extraction

Instead of using reading times to fit models as in Wu et al. (2018), the current project uses surprisals obtained from the LM introduced in Section 4.1.

The first step was to generate the target sentences using the LM and store the conditional probability of a current word given its prefix.[7] The generation process is illustrated in Figure 2. Suppose we want the probability of the tokens in the sequence *shi-kuai ...* (stone ...). Given the starting point (a special token `[CLS]`), the model returns a probability distribution over the vocabulary. Because the first token of the sequence is *shi*, the probability of that token is stored. In the next step, the output of the previous step (`[CLS] shi`) is the input. The model returns a new probability distribution over the vocabulary. This time, the conditional probability of *kuai* is stored. The generation keeps going until the whole target sentence is generated. There are 96 target sentences. Each sentence has a list of probabilities corresponding to each token in the sentence as in (10).

(10)

| shi | kuai | za | zhong | ... |
|---|---|---|---|---|
| 5.56*e*-04 | 1.13*e*-05 | 2.45*e*-05 | 6.12*e*-04 | |

### 4.3 Calculating region surprisals

The sentences and their token conditional probabilities were generated token by token. However, one region in a RC can contain more than one token. In order for the results to be comparable to Wu et al.'s (2018), region probabilities were calculated based on token probabilities. Following common practice, the region probability is the multiplication of the token probabilities in that region. For example, the probability of the phrase *shi-kuai* (stone) being at the beginning of a sentence is 5.56*e*-04 × 1.13*e*-05 = 6.28*e*-09. Region surprisals were calculated as the negative log of the corresponding probabilities. Transforming probabilities to surprisals makes the results intuitive and comparable to the human experiment.

### 4.4 Data analysis

The procedure of data analysis follows Wu et al. (2018). First, the surprisals were aligned by regions (8) across the four conditions (9). For each region, a linear mixed-effects model was fitted using the lme4 (Bates et al., 2015) and lmerTest packages (Kuznetsova et al., 2017) in R (R Core

[7] The target sentences can be found under the Supporting Information label in this link `https://onlinelibrary.wiley.com/doi/10.1111/cogs.12551`.

Team, 2020, version 4.1.2). All models had the target sentences as the by-item random variable. The model parameter REML was set to FALSE. For the residuals of each fitted model, a significance test was conducted to compare the residual distribution to a normal distribution in order to check if there were serious deviations from normality. All models were initially maximal in the sense of Barr et al. (2013). When a model failed to converge, the random slope of the interaction Cl × bei in the by-item random variable was removed. By doing so, all models converged. The significance level was chosen to be $\alpha = 0.05$.

## 5 Results

Table 1 reports the statistical results for each region. For the regions in bold, the random slope of the interaction Cl × bei in the by-item random variable was removed so that the model can converge. The asterisks after the *p*-values indicate that the results reached statistical significance. Wu et al. (2018) took an absolute *t*-value equal to or above 2 as statistical significance at $\alpha = 0.05$. All statistically significant results in Table 1 meet their criterion. For each region, a Shapiro-Wilk test was conducted on the model residuals. It showed no evidence of non-normality.

At the RCNP region, there was a main effect of Cl. An interaction of Cl and bei was also found.

At the RCV, De, and head noun regions, there were main effects of Cl and bei. No interaction was found.

At the Adv region, there was only a main effect of bei.

At the MainV region, there was only a main effect of Cl reached statistical significance.

At the MainO region, no main effects and no interaction were found.

Figure 3 shows the mean surprisal value of each region in the four conditions. At the bottom is a comparison of the four conditions.

## 6 Discussion

### 6.1 By-region mean surprisals

The mean surprisal values in Figure 3 perfectly match the predictions made by the experience-based approach in section 2.1.3. When there was a mismatching classifier, the surprisal of the RCNP region is higher than there is no mismatching classifier. That is, in the bottom plot in Figure 3, the red and yellow lines are higher than the blue and green lines. In the RCV region, the surprisal is low in the conditions with bei. This is expected since bei obligatorily takes a verb. The influence of the presence of Cl and bei is the most obvious in the De region. The experience-based approach predicts that for human readers, the reading time of the De region and beyond is the fastest when both early cues are present (i.e., condition (d)). It is the slowest when there is no early cue at all (i.e., condition (a)). The reading time with only one cue should lie in between. This cue facilitation prediction matches the pattern seen in Figure 3. In the De region, the yellow line (condition (d)) is at the bottom, the blue line (condition (a)) is on the top, and the other two lines are in the middle. The same pattern is observed in the headnoun region as well but disappear in the farther regions.

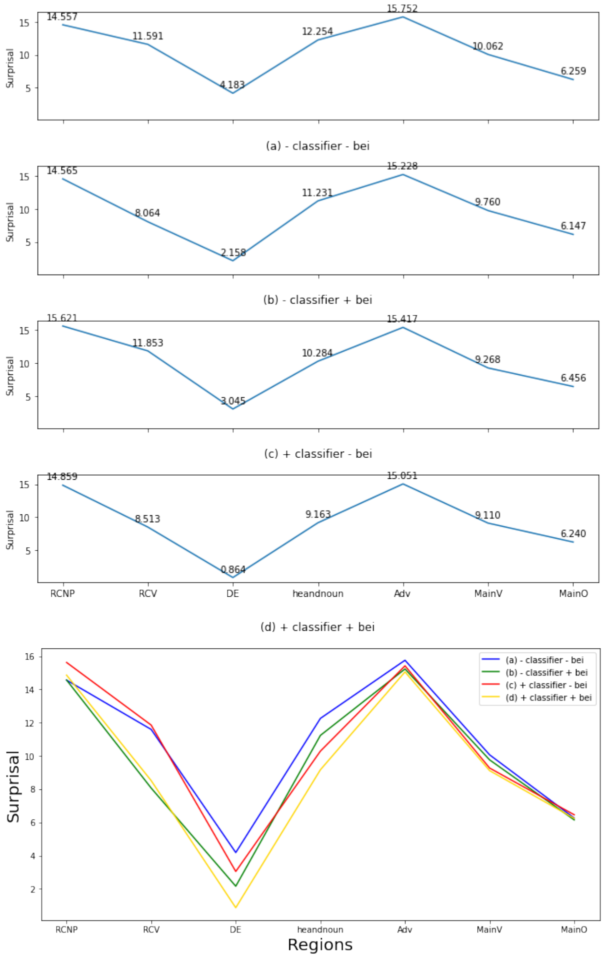


Figure 3: Mean surprisal of each region in the four conditions.

| Region | Contrast | Coef. | SE | *t*-value | *p*-value | Shapiro-Wilk |
|---|---|---|---|---|---|---|
| **RCNP** | Cl | 0.3394 | 0.0989 | 3.432 | 0.002** | 0.700 |
| | bei | -0.188 | 0.141 | -1.337 | 0.194 | |
| | Cl × bei | -0.193 | 0.069 | -2.805 | 0.0098** | |
| **RCV** | Cl | 0.178 | 0.085 | 2.083 | 0.048* | 0.514 |
| | bei | -1.717 | 0.224 | -7.681 | 0.000*** | |
| | Cl × bei | 0.047 | 0.052 | 0.897 | 0.379 | |
| **De** | Cl | -0.608 | 0.048 | -12.606 | 0.000*** | 0.149 |
| | bei | -1.051 | 0.086 | -12.196 | 0.000*** | |
| | Cl × bei | -0.039 | 0.033 | -1.193 | 0.244 | |
| headnoun | Cl | -1.009 | 0.110 | -9.166 | 0.000*** | 0.313 |
| | bei | -0.536 | 0.130 | -4.126 | 0.000*** | |
| | Cl × bei | -0.025 | 0.063 | -0.391 | 0.699 | |
| **Adv** | Cl | -0.128 | 0.202 | -0.631 | 0.534 | 0.784 |
| | bei | -0.223 | 0.065 | -3.408 | 0.002** | |
| | Cl × bei | 0.040 | 0.034 | 1.169 | 0.254 | |
| **MainV** | Cl | -0.361 | 0.094 | -3.851 | 0.000*** | 0.691 |
| | bei | -0.115 | 0.059 | -1.932 | 0.065 | |
| | Cl × bei | 0.036 | 0.024 | 1.511 | 0.144 | |
| MainO | Cl | 0.072 | 0.062 | 1.170 | 0.254 | 0.112 |
| | bei | -0.082 | 0.042 | -1.972 | 0.060 | |
| | Cl × bei | -0.026 | 0.014 | -1.857 | 0.076 | |

Table 1: Main effects of Cl and bei, and their interaction by region. The models fitted for the regions in bold are reduced models. The *p*-values marked with * mean the results reach statistical significance.

### 6.2 Statistical results

### 6.3 Absence of spillover effects

Facilitatory effects caused by Cl and/or bei are detected both in Wu et al.'s 2018 human experiment and the current LM experiment. However, the effects showed up in different patterns.

In the human experiment, main effects and/or interactions were found in all seven regions except for the RCNP region. Most statistically significant results were found in the matrix clause regions (Adv, MainV, and MainO). In other words, there are spillover effects. In the current study, the facilitatory effects show up as lower surprisals (i.e., negative correlations). The majority of the main effects and/or interactions were found within the RC regions, namely RCNP, RCV, De, and headnoun. The spillover effects are not as evident.

The absence of spillover effects is not surprising. Spillover effects in human experiments mainly caused by cognitive demands (Ehrlich and Rayner, 1983; Remington et al., 2018). But for computational models, there is no cognitive process. All they rely on is probability (experience in the case of human experiments). As such, any effects the previous context has on the current token will display themselves immediately and show how far their impact lasts.

### 6.4 Lexical-disruption effects

Wu et al. claim that there are lexical-disruption effects caused by Cl in the combined region De + headnoun but not directly in RCNP. The disruption effects were canceled out by the presence of bei. In the current experiment, the statistically significant disruption effects of Cl immediately show up in the RCNP region. The presence of Cl has a positive correlation with the surprisal of that area (coef. = 0.3394, $p = 0.002 < 0.05$). The effects are still marginally significant in the RCV region (coef. = 0.178, $p = 0.048 < 0.05$) but do not extend to the other regions.

However, the disruption effects caused by Cl are canceled by the presence of bei, as the surprisal is negatively correlated to Cl × bei in the RCNP region. The lexical-disruption effects are not statistically significant in the RCV region under the Cl × bei condition. Beyond the RCNP and RCV regions, there is no significant disruption effects.

### 6.5 Facilitatory effects

Wu et al. found that CL and bei elicited facilitatory effects. The effects from CL spilt over to the matrix clause regions while bei facilitated processing early within the relative clause. Translating the facilitatory effects into the context of surprisals, it means lower surprisal values (negative coefficients) are expected. In the current experiment, the existence of CL had a statistically significant negative correlation with the region surprisals of DE, headnoun, and MainV. The facilitatory effects caused by bei exit throughout the whole sentence but are only significant in the RCV, DE, headnoun, and Adv regions. The co-existence of CL and bei only elicit significant facilitatory effects in the RCNP region. The majority of the facilitatory effects exist within the RC.

### 6.6 Advantages of LMs

To evaluate the experience-based account of sentence processing, especially for the surprisal theory, having access to conditional probability is essential and crucial. In their paper, Wu et al. point out that while the surprisal theory gives a concrete way to define processing difficulties, it is not easy to get the probability estimates. Previous studies use language corpora (e.g., Wu, 2011; Jäger et al., 2015; Wu et al., 2018) or built their own LMs (Van Schijndel and Linzen, 2021) to obtain probabilities. However, they have shortcomings.

Corpora have mainly two disadvantages. First, although it is easy to get frequencies of sequences in a corpus, it can be hard to get the conditional probabilities of sequences. Second, not all possible token combinations are available in corpora. For example, in her corpus study, Wu (2011) finds that the local classifier-noun mis-matching pattern does not exist in the corpora she searched, even though such pattern does exist in Chinese. In Wu et al. (2018), the authors attribute this issue to the limited size of corpora. To overcome this limitation, Wu et al. had to conduct a sentence-completion norming study with ninety-three participants to obtain the probability of a RC across four conditions (9).

Not only the corpora, building a LM specifically for one project might also be affected by the size of data. As listed in (11), the corpus used in Van Schijndel and Linzen (2021) is larger than the Chinese Treebank 5.0 used in Wu et al. (2018) by several orders of magnitude. However, it is still smaller than the corpus (11c) on which the LM in the current experiment is trained. For LMs, the larger the dataset size is, the more comprehensive a model understands a language, and the better a model performs on downstream tasks (Devlin et al., 2018; Radford et al., 2019; Xu et al., 2020).

(11) Corpus sizes

a. Wu et al. (2018)
Chinese Treebank 5.0
**824,983** Chinese characters

b. Van Schijndel and Linzen (2021)
Subset of English Wikipedia and soap opera dialog
**180 million** English words

c. Current study
CLUECorpusSmall
**5 billion** Chinese characters

The performance of a LM is also dependent on its model architecture. Van Schijndel and Linzen's (2021) LM has two layers, which is much shallower than LMs like BERT (Devlin et al., 2018), GPT2 (Radford et al., 2019) and GPT3 (Brown et al., 2020), each of which has at least twelve blocks with multiple layers in each block. It has been shown, for example, in Brown et al. (2020), that models with more layers (hence, more parameters) perform better on downstream tasks even without specific fine-tuning.

Compared to language corpora and small LMs, the advantages of large scale pre-trained LMs are obvious. They are trained on unlimited data covering various topics and genres with sophisticated powerful architecture. Hence, it is not surprising that the experiment results of the human experiment can be reproduced in the LM, although the results showing a different pattern concerning spillover effects. For future linguistic study, it would be interesting to see whether surprisals obtained from large scale LMs can detect the garden path effects as in Van Schijndel and Linzen as well as reach the magnitude. In general, if the surprisals extracted from the models mimic human experiment results in a systematic way, the methodology used here may substitute norming experiments to elicit conditional probabilities for future study. This has the potential to facilitate experiment design because it is inexpensive, time-saving, and efficient.

# 7 A LSTM trained on cluecorpussmall

A gpt2 structure LM is powerful but it does not resemble human cognition. A LSTM neural net, although being computationally less powerful and slow to train, seems to have more similarities to human because of its long- and short-term memory setting. To see whether a LSTM can perform better than gpt2-chinese-cluecorpussmall, the same experiment was run on a LSTM, which is trained on the same training data set as gpt2-chinese-cluecorpussmall. Results are presented in this section. However, it is worthnoting that the results of gpt2-chinese-cluecorpussmall and the LSTM are not strictly comparable because of two differences in the models. First, the LSTM has only 25.8 million parameters while gpt2-chinese-cluecorpussmall has 102 million. Second, the two LMs have very difference structure, with the LSTM being much shallower than gpt2-chinese-cluecorpussmall.

## 7.1 Results of LSTM

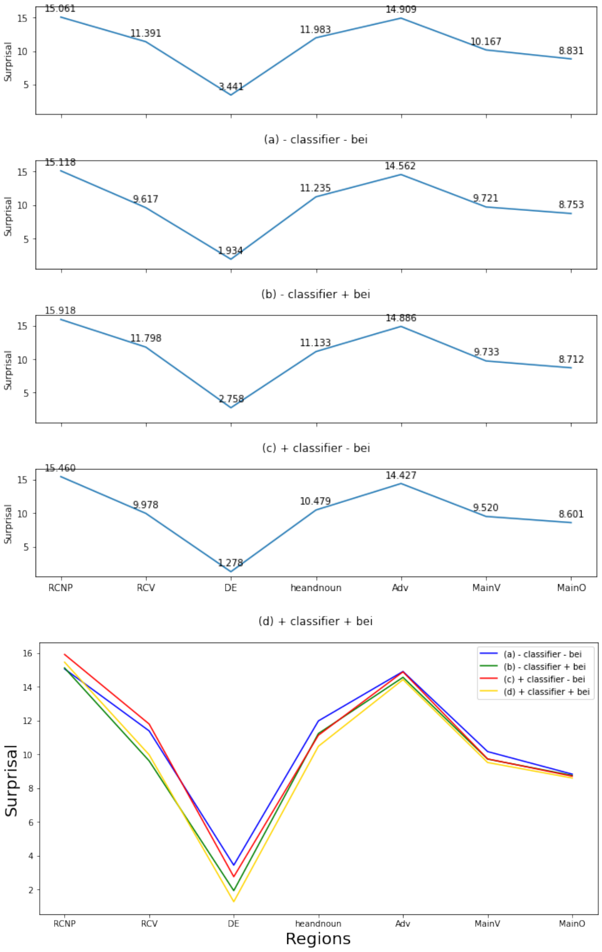
Table 2 reports the by region statistical results of the LSTM. At the RCNP region, there was a main effect of CL. Unlike in Table 1, there is no interaction of CL and `bei`. At the RCV, DE, and head noun regions, there were main effects of CL and `bei`. No interaction was found. At the Adv region, there was a main effect of `bei`. At the MainV region, there were main effects of CL and `bei` as well as an interaction of both. At the MainO region, no main effects and no interaction were found.

Figure 4 shows the mean surprisal value of each region in the four conditions. At the bottom is a comparison of the four conditions.

## 7.2 Discussion of LSTM

As the gpt2 model, there were lexical-disruption effects detected in the RCNP region when a mismatching CL was present. Its influence extended into the RCV region. Otherwise, the existence of CL lowered the surprisal of the most regions.

What was different from the gpt2 model is that although the co-existence of `bei` with CL in this region removed the disruption effects, it did not statistically significantly decrease the surprisal (coef. = -0.129, $p$ = 0.118). Also, the surprisal of the MainV region was significantly affected by the early cues, while in the gpt2 model, only CL significantly lower the surprisal value.



(a) - classifier - bei

(b) - classifier + bei

(c) + classifier - bei

(d) + classifier + bei

Figure 4: Mean surprisal of each region in the four conditions.

Overall, the performance of the LSTM is like a fusion of the gpt2 model results and the human results. On the one hand, the facilitatory effects of the early cues occurs early in the RC area, showing no cognitive demand delay. On the other hand, the effects caused by the early cues spilt over into the main clause region while in the gpt2 model only CL did that. Concerning the interaction of CL and `bei`, the LSTM model does not accord with the predictions made by the experience-based approach. Nor did the human results.

# 8 Conclusion

The currect project conducted the same experiment as reported in Wu et al. (2018) Experiment 2 but with the large scale pre-trained Chinese LM gpt2-chinese-cluecorpussmall and the LSTM trained on the same data. In general, lexical-disruption effects and facilitatory effects were found. In the gpt2 model, the results matched all predictions made by the experience-based ap-

| Region | Contrast | Coef. | SE | *t*-value | *p*-value | Shapiro-Wilk |
|---|---|---|---|---|---|---|
| RCNP | Cl | 0.300 | 0.122 | 2.451 | 0.022* | 0.079 |
| | bei | -0.100 | 0.133 | -0.755 | 0.457 | |
| | Cl × bei | -0.129 | 0.080 | -1.622 | 0.118 | |
| **RCV** | Cl | 0.192 | 0.063 | 3.062 | 0.005** | 0.394 |
| | bei | -0.898 | 0.174 | -5.162 | 0.000*** | |
| | Cl × bei | -0.012 | 0.049 | -0.234 | 0.817 | |
| De | Cl | -0.335 | 0.034 | -9.918 | 0.000*** | 0.627 |
| | bei | -0.747 | 0.081 | -9.166 | 0.000*** | |
| | Cl × bei | 0.007 | 0.017 | 0.403 | 0.69 | |
| **headnoun** | Cl | -0.401 | 0.082 | -4.919 | 0.000*** | 0.501 |
| | bei | -0.350 | 0.096 | -3.661 | 0.001** | |
| | Cl × bei | 0.023 | 0.037 | 0.639 | 0.529 | |
| Adv | Cl | -0.039 | 0.105 | -0.373 | 0.713 | 0.536 |
| | bei | -0.201 | 0.067 | -3.01 | 0.006** | |
| | Cl × bei | -0.028 | 0.031 | -0.923 | 0.365 | |
| **MainV** | Cl | -0.159 | 0.042 | -3.770 | 0.001*** | 0.346 |
| | bei | -0.165 | 0.044 | -3.709 | 0.001** | |
| | Cl × bei | 0.058 | 0.021 | 2.831 | 0.009** | |
| **MainO** | Cl | -0.068 | 0.039 | -1.754 | 0.092 | 0.048* |
| | bei | -0.047 | 0.032 | -1.493 | 0.149 | |
| | Cl × bei | -0.008 | 0.013 | -0.629 | 0.535 | |

Table 2: Main effects of Cl and bei and their interaction by region. The models fitted for the regions in bold are reduced models. The *p*-values marked with * mean the results reach statistical significance.

proach but showed few or no spillover effects. In the LSTM results, although the results matched the predictions but not all of them reached a statistically significant level. There were slight spillover effects.

It has been shown that using pretrained large-scale LMs can be promising for human experiment design as it offers insights for potential outcomes and is convenient for getting conditional probabilities, which are important for many human studies. By comparing the results of the gpt2 model to the LSTM, although neither of them replicated the human results, the gpt2 model showed a pattern as the human results without any delay caused by cognitive demand.

## References

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1):1–48. https://doi.org/10.18637/jss.v067.i01.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* .

Mark Davies. 2011. Corpus of American soap operas: 100 million words.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Kate Ehrlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of verbal learning and verbal behavior* 22(1):75–87.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1):1–76.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain* 2000:95–126.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics*.

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research* 32(2):101–123.

Lena Jäger, Zhong Chen, Qiang Li, Chien-Jer Charles Lin, and Shravan Vasishth. 2015. The subject-relative advantage in chinese: Evidence for expectation-based processing. *Journal of Memory and Language* 79:97–120.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13):1–26. https://doi.org/10.18637/jss.v082.i13.

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

Martha Palmer, Fu-Dong Chiou, Nianwen Xue, and Tsan-Kuang Lee. 2005. Chinese Treebank 5.0 ldc2005t01. *Web Download* .

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Roger W Remington, Jennifer S Burt, and Stefanie I Becker. 2018. The curious case of spillover: Does it tell us much about saccade timing in reading? *Attention, Perception, & Psychophysics* 80(7):1683–1690.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–319.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science* 45(6):e12988.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. pages 5998–6008.

Fuyun Wu. 2011. Frequency issues of classifier configurations for processing mandarin object-extracted relative clauses: A corpus study .

Fuyun Wu, Elsi Kaiser, and Shravan Vasishth. 2018. Effects of early cues on the processing of chinese relative clauses: Evidence for experience-based theories. *Cognitive science* 42:1101–1133.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *ArXiv* abs/2003.01355.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019* page 241.