Who did what to Whom? Language models and humans respond diversely to features affecting argument hierarchy construction

Xiaonan Xu
University of Cologne, Germany

Haoshuo Chen
Nokia Bell Labs, USA

Abstract
Pre-trained transformer-based language models have achieved state-of-the-art performance in many areas of NLP. It is still an open question whether the models are capable of integrating syntax and semantics in language processing like humans. This paper investigates if models and humans construct argument hierarchy similarly with the effects from telicity, agency, and individuation, using the Chinese structure “NP1+BA/BEI+NP2+VP”. We present both humans and six transformer-based models with prepared sentences and analyze their preference between BA (view NP1 as an agent) and BEI (NP2 as an agent). It is found that the models and humans respond to (non-)agentive features in telic context and atelic feature very similarly. However, the models show insufficient sensitivity to both pragmatic function in expressing undesirable events and different individuation degrees represented by human common nouns vs. proper names. By contrast, humans rely heavily on these cues to establish the thematic relation between two arguments NP1 and NP2. Furthermore, the models tend to interpret the subject as an agent, which is not the case for humans who align agents independently of subject position in Mandarin Chinese.\(^1\)

1 Introduction
Pre-trained transformer-based language models (LMs) keep achieving state-of-the-art performance in NLP tasks. Many studies have indicated that pre-trained LMs can learn syntactic knowledge (e.g., Linzen et al. 2016; Gulordava et al. 2018 for subject-verb agreement, Wilcox et al. 2018 for filler-gap dependencies, Futrell et al. 2019 for garden-path effects) and semantic knowledge (e.g., Zhao et al. 2021 for telicity, Kementchedjhieva et al. 2021 for causality bias, Misra et al. 2020 for semantic priming, Misra et al. 2021 for typicality, Ettinger 2020 for role reversal and same-category distinctions). However, to what extent LMs can acquire knowledge in the syntax-semantics interface is still an open question. To answer this question, we explore arguments hierarchy construction which identifies the thematic roles of arguments in the semantic domain and aligns arguments and subject/object in the syntactic domain. In this hierarchy, the active, controlling agent (prototypical actor) outranks the affected patient (prototypical undergoer), i.e., who did what to whom? (Van Valin Jr, 1990; Van Valin and LaPolla, 1997; Bornkessel et al., 2005). The mapping between thematic roles (agent/patient) and syntactic structure (subject/object) varies depending on various features.

In this paper, we investigate whether pre-trained transformer-based LMs and humans behave similarly in the argument hierarchy construction using the Chinese structure “NP1+BA/BEI+NP2+VP”. This structure provides a unique opportunity to examine the alignment through the occurrence of BA/BEI (Deng et al., 2018), without interference from morphology or word order. For example, human name Zhang-san (NP1) in the subject position of sentence (1a) with BA is interpreted as an agent, and human name Li-si (NP2) in the object position is viewed as a patient. By contrast, if BEI occurs as in (1b), subject Zhang-san is viewed as a patient, and object Li-si is considered an agent. This inverse interpretation depending on BA/BEI allows us to use word prediction to study LMs without task-specific fine-tuning. It also avoids tokenization issues since both BA and BEI are single characters.

\[(1a)\] 张三 [把] 李四 [杀] 死 [了] .
Zhangsan [ba] lisi [shasha] -le
‘Zhangsan killed Lisi.’

\[(1b)\] 张三 [被] 李四 [杀] 死 [了] .
Zhangsan [bei] lisi [shasha] -le
‘Zhangsan was killed by Lisi.’

\(^1\)Dataset for both humans and language models, and analysis code are available at https://github.com/NLPbelllabs/WhoWhom.git
The construction of argument hierarchy can be affected by different cues related to telicity, agency, and individuation via notion transitivity (Hopper and Thompson, 1980; De Mattia-Viviès, 2009; Virtanen, 2015). For example, a cue emphasizing the agentive property of NP1 (e.g., by adding the adverbial *volitionally*) increases the probability of NP1 being viewed as an agent (Cruse, 1973), making BA more natural than BEI. By contrast, a cue denoting the non-agentive property of NP1 (e.g. by adding the clause *what happened to NP1 was that...*) decreases the probability of NP2 being viewed as an agent, making BEI more natural than BA. To examine the effects of these cues, we carry out a human acceptability judgment experiment using sentences with BA/BEI and compare the result with the probability of masked token BA/BEI predicted by the six pre-trained transformer-based LMs: BERT-base, ELECTRA-large, RoBERTa-base, ERNIE 1.0, and MacBERT-base/large. The results show that the models and humans construct similar argument hierarchy with atelic feature, and both agentive and non-agentive feature in telic context. However,

(A) LMs show insufficient sensitivity to the pragmatic function of BEI in forming adversative passives with disposal verbs, but humans depend on it in establishing thematic relation between the arguments.

(B) LMs and humans present different responses to various degrees of individuation encoded in human common nouns vs. proper names. Humans often perceive proper nouns as agents. However, LMs are inclined to interpret common nouns as agents.

(C) Unlike Mandarin Chinese native speakers who do not align the agent role depending on subject position, LMs tend to interpret the subject as an agent in telic context.

2 Materials

We prepare a dataset including the sentences highlighting telicity-, agency-, and individuation-related features. To avoid gender effect, we choose frequently used male surnames and first names to form NP1 and NP2 in the structure "NP1+BA/BEI+NP2+VP". For each condition, we make a hypothesis about human judgment in BA/BEI-preference based on previous studies about features in the structure.

2.1 Telicity

2.1.1 Atelic-condition

We use dynamic atelic verbs and imperfective aspect -*zhe* to build atelic sentences. The dynamic verbs such as *la 'pull' in (2a) and xun-chi ‘reprimand’ in (2b) with imperfective -*zhe* represent durative events without inherent endpoints (Vendler, 1957; Smith, 2012; Xiao and McEnery, 2004a). BEI with dynamic verbs can collocate with imperfective aspect -*zhe* (Cook, 2019; Xiao et al., 2006). But the co-occurrence of BA with dynamic verbs and -*zhe* is rarely found (Tsung and Gong, 2021). We expect a preference for BEI over BA in the atelic-condition.

(2a) 郭杰 把/被 张伟 拉 着。guo-jie ba/bei zhang-wei la -zhe
Guojie BA/BEI Zhangwei pull -IMPF
‘Guojie is pulling Zhangwei.’

(2b) 赵涛 把/被 吴波 训斥 着。zhao-tao ba/bei wu-bo xun-chi -zhe.
Zhaotao BA/BEI Wubo reprimande -IMPF
‘Zhaotao is being reprimanded by Wubo.’

2.1.2 Telic-condition

(3a) 郭杰 把/被 张伟 拉 到 了 门口。guo-jie ba/bei zhang-wei dao-le men-kou
Guojie BA/BEI Zhangwei pull arrive -PERF door
‘Guojie pulled Zhangwei to the door.’

(3b) 赵涛 把/被 吴波 训斥 了 一顿。zhao-tao ba/bei wu-bo xun-chi le yi-dun.
Zhaotao BA/BEI Wubo reprimande -PERF one-CL
‘Zhaotao reprimanded Wubo.’

A modifier specifying an endpoint can change an atelic verb at the lexical level into a telic situation at clause level (Vendler, 1957; Xiao and McEnery, 2004a). We set up two types of telic modifiers. The first one uses prepositional phrases (PPs) like *dao...men-kou ‘arrive at the door’ denoting a spatial endpoint (3a). The second one uses yi-dun ‘one+CL’ indicating an temporal endpoint, where the specific verbal classifier dun is used to measure the count of a durative event (3b) (McEnery

---

2 Markers signaling viewpoint aspect, such as perfective marker -*le* in the examples (1, 3-10) or imperfective marker -*zhe* in (2), are necessary for the grammatical correctness of Chinese sentences (Li and Thompson, 1989). In atelic-condition, we choose the imperfective marker -*zhe* to emphasize ongoing, uncompleted events.
and Xiao, 2007; Li and Thompson, 1989). We combine one-half of atelic verbs like la ‘pull’ with PPs to build spatially telic VPs (3a) and the other half verbs with yi-dun to form temporally telic VPs (3b). Both telic VPs co-occur with the perfective marker -le and are used in the following agency- and individuation-related conditions.

One crucial distinction between the spatially and temporally telic sentences is that the former with dao ‘arrive’ denotes an instantaneous, non-durative event, and the latter describes a durative event approaching an endpoint incrementally. Linguistic studies suggest that both BA and BEI are compatible with a telic situation (Liu, 1997; Yang, 1995; Xiao and McEnery, 2004b). We examine whether BA and BEI are acceptable in both temporally and spatially context in the telic-condition.

2.2 Agency
Adopting cues highlighting agentive or non-agentive feature can modify the thematic roles mapped to NPs. We form three condition groups: (1) a manner adverbial ‘volitionally’ vs. ‘unfortunately’, (2) a subordinate clause with ‘do’ vs. ‘happen’, and (3) a purpose phrase with ‘in order to’ (Gruber, 1967; Cruse, 1973) to construct sentences.

2.2.1 Volition and non-volition-condition
The Chinese adverbial gu-yi ‘volitionally’ after NP1 in (4) presents the intention of NP1 to carry out an action (Cruse, 1973) and drives NP1 to be interpreted as an agent. It harmonizes with BA, which indicates NP1 as an agent, but conflicts with BEI, which signals NP1 as a patient. By contrast, the adverbial bu-xing ‘unfortunately’ in (5) demonstrates a non-volitional, passive property of NP1. It agrees with BEI but contradicts BA.

2.2.2 Do- and happen-condition
The do/happen-clause is another way to test agentive and non-agentive property. For example, John in John punched Bill is viewed as an agent, as What happened to John was punch Bill is odd (Cruse, 1973). On the contrary, John in John was punched by Bill is viewed as non-agent, as What happened to John was that he was punched by Bill is normal and What John did was that he was punched by Bill is abnormal. We place the do/happen-clause as in (6) and (7) to modify agentive/non-agentive feature of NP1.

2.2.3 Aim-condition
A third widely discussed test for the agency is the modifiability by a phrase in order to. For example, John in John looked into the room in order to learn who was there is viewed as a willful agent (Gruber, 1967). Similarly, the purpose phrase wei-le da-dao mu-di ‘in order to achieve goal’ after the NP1 in (8) emphasizes NP1’s purpose, which matches NP1’s agent role with BA and contradict NP1’s patient role with BEI.

In sum, we predict that the tested telic context show consistent BA/BEI-preference under the effect of agency, that is, the volition-, do- and aim-
2.3 Individuation

Human common nouns like ‘worker’ are regarded to be less identifiable and individuated than human proper names like Guo-jie, which are more likely to be perceived as agents in human comprehension (Faurud, 1996; Yamamoto, 1999; Dixon, 1979; Timberlake, 1977). In NP2-com-condition (9), frequently used occupation names like “worker” are used as common nouns for NP2 and male human names are used as proper names for NP1. NP1-com-condition (10) is in reverse. We predict that humans prefer BA for NP2-com-condition and BEI for NP1-com-condition as the proper names are more likely to be viewed as agents.

Human BA/BEI-preference can be attributed to human sensitivity to different ways of referring such as common nouns vs. proper names. It is uncertain whether LMs own this sensitivity. Therefore, we predict that LMs may behave differently. For grammatical correctness, each common noun occurs with a numeral yi ‘one’ and the general classifier ge (Zhang, 2013).

NP2-com-condition:

(9) 郭杰 把 拉到了门口。
  guo-jie ba/he la dao -le men-kou
  Guojie in order to achieve goal
  ‘Guojie aiming to achieve his goal.’

NP1-com-condition:

(10) 一个工人 把 杨华 拉到了门口。
  yi-ge gong-ren ba/he yang-hua la dao -le men-kou
  ‘(that) he pulled Zhangwei to the door.’
  ‘(that) he was pulled by Zhangwei to the door.’

3 Experiment

3.1 Human Judgment Task

We prepare 18 verbs to form 36 sentences either with BA or with BEI for each of the 9 conditions, resulting in 324 sentences in total5. To avoid repeating verbs and NPs, we split these sentences evenly over 18 lists following a Latin-Square design, with 18 sentences in each list. Every list contains each condition twice and each of the 18 verbs once. Additional 10 sentences which are either semantically or syntactically incorrect were added to each list as fillers. Each of the lists was pseudo-randomized so that two test items from a single condition did not appear sequentially.

We conducted an acceptability judgment experiment using a four-point-scale questionnaire to obtain human ratings. Participants are required to mark the sentences following this instruction: entirely acceptable sentences should be marked with 1; sentences containing some expression which is acceptable to some degree, but not fully acceptable, should be marked with 2; sentences containing some expression which is unacceptable to some degree, but not fully unacceptable, should be marked with 3; and sentences containing some expression which is fully unacceptable should be marked with 4. A larger score indicates a sentence is less acceptable.

This human judgement experiment was administered on the Chinese website of wenjuanxing6, 121 university students from mainland China participated in this experiment voluntarily. Their ages range from 18 to 25 years old, with a mean age of 20.6 years. Fifty-six of them are female. They all reported a monolingual Mandarin Chinese background except one female. Her and the other 11 participants’ data were filtered out because of their low judgment scores (meaning high acceptable) on unacceptable filler items sentences (mean < 3.5).

3.2 LM Prediction

We replace BA/BEI in our sentences with a masked token and measure the output at the corresponding position for BA and BEI in different conditions for six pre-trained transformer-based LMs: BERT-base (Devlin et al., 2018), RoBERTa-base (Liu et al., 2019), ELECTRA-large (Clark et al., 2020), ERNIE 1.0 (Sun et al., 2019), MacBERT-base and MacBERT-large (Cui et al., 2020), implemented in

5We publish all the sentences at Github.
6https://www.wjx.cn
the Huggingface Transformers library (Wolf et al., 2019). Even though these LMs have different pre-training tasks and use different databases in different sizes (see Table 5 in Appendix), we expect that they show (or tend to show) a consistent rather than inconsistent performance in the prediction of BA/BEI for each condition.

3.3 Measure

We define $B_{hum}$ as BA/BEI-preference bias $B$ for humans based on $Accep$ which is the judgment score for each sentence $S$. $B_{hum}$ quantifies the preference of a sentence to occur with BA or BEI. It is negative with BA preferred and positive with BEI preferred.

$$B_{hum} = Accep(BA|S) - Accep(BEI|S)$$  \hspace{1cm} (1)

For LMs, surprisal is defined as the inverse log probability of a word ($w_i$) conditioned on the surrounding words in a context $C$:

$$Surp(w_i|C) = log \frac{1}{p(w_i|C)}$$  \hspace{1cm} (2)

Due to the fact that BA and BEI are not exclusive to each other, we follow Misra et al. (2020) and define BA/BEI-preference bias $B$ for LM $B_{LM}$ as the surprisal difference between BA and BEI.

$$B_{LM} = Surp(BA|C) - Surp(BEI|C)$$  \hspace{1cm} (3)

$B_{LM}$ is negative if BA is preferred and positive if BEI is preferred. $B_{LM}$ has been applied as a linking function between human expectations and LM’s output (Hale, 2001). In this paper, we employ $B_{LM}$ and $B_{hum}$ to test the BA/BEI-preference of humans and LMs under the effects of various features.

4 Results

Average $B_{LM}$ and $B_{hum}$ are visualized in Figure 1. $B_{LM}$ is averaged for every condition within each LM. $B_{hum}$ is averaged over all the participants for every condition. We further examine average $Accep$ and average $Surp$ for BA and BEI from RoBERTa-base for each condition in Figure 2 (other LMs present similar results, see Figure 5 in Appendix). The human $Accep$ for all items in each condition show a lower averaged coefficient of variation over all the conditions than $Surp$ of all LMs (0.42 vs. 0.64, detailed results see Figure 4 in Appendix). Statistically, the temporally telic and spatially telic context in all the conditions except for telic- and NP2_con-condition show quite consistent pattern regarding the BA/BEI-preference in both human $Accep$ and $Surp$ of LMs, suggesting that the difference between temporally telic and spatially telic context play a limited role in the BA/BEI-preference for these conditions. Thus we compare the results between temporally telic and spatially telic context only for telic- and NP2_con-condition. The human $Accep$ and $Surp$ of each LM for each condition are fitted with a linear mixed-effects model using the lme4 package in R (Bates et al., 2015). The model treated variable BA/BEI as a fixed effect with a random intercept for each verb (detailed results see Table 3 in Appendix).

4.1 Telicity

In atelic-condition, positive $B_{hum}$ ($p \leq 0.001$) and $B_{LM}$ ($p \leq 0.05$ for all the LMs), see Figure 1 and Table 3 in Appendix, confirm our prediction of BEI-preference for humans and LMs. In telic-condition, Figure 2 shows that the human acceptability of BA and BEI are relatively high (low judgement scores), which supports our prediction that BA and BEI are
Table 1: Preference comparison between BA and BEI for humans and LMs in the temporally and spatially telic context for telic- and NP\textsubscript{2}\textsubscript{com}-condition. (ba: statistically significant BA-preference, bei: statistically significant BEI-preference. **: $p \leq 0.01$, ***: $p \leq 0.001$)

| condition | context       | Humans | BERT-base | RoBERTa-base | ELECTRA-large | ERNIE 1.0 | MacBERT-base | MacBERT-large |
|-----------|---------------|--------|-----------|--------------|--------------|-----------|--------------|--------------|
| telic     | temporal telic| bei*** | ba***     | –            | bei**        | –         | –            | ba***        |
|           | spatially telic| –      | ba***     | ba***        | ba***       | ba***    | –            | ba***        |
| NP\textsubscript{2}\textsubscript{com} | temporally telic | –      | bei*      | bei**        | bei***      | bei**    | bei***       | –            |
|           | spatially telic | ba***  | –         | ba**         | ba**        | –         | –            | ba***        |

Figure 2: Average Accep from human acceptability judgment experiment (A) and average Surp for RoBERTa-base (B) for each condition. The 9 conditions belong to three groups: telic/atelic-condition is related to telicity (Sec. 2.1), do/happen/aim/non-volition/volition-condition is related to agency (Sec. 2.2) and NP\textsubscript{2}\textsubscript{com}/NP\textsubscript{1}\textsubscript{com}-condition is related to individuation (Sec. 2.3). The values from telic-condition are set as reference lines.

both acceptable in the telic context. However, positive $B_{\text{hum}}$ ($p \leq 0.01$) and negative $B_{LM}$ ($p \leq 0.05$ except ELECTRA-large) in Figure 1 reveal distinction between humans and LMs.

Results of humans and each LM in both temporally and spatially telic context of telic-condition are further compared at Table 1. While participants preferred BEI ($p \leq 0.001$) for the temporally telic sentences, LMs show inconsistent results. As LMs prefer BA ($p \leq 0.01$) consistently for the spatially telic sentences, no significant preference is found in human judgment.

4.2 Agency

Figure 1 shows consistent negative $B_{LM}$ and $B_{\text{hum}}$ for do/aim/volition-condition (all with $p \leq 0.001$) and consistent positive $B_{LM}$ and $B_{\text{hum}}$ (all with $p \leq 0.001$) for non-volition-condition. A small discrepancy is found in happen-condition, where participants preferred BEI ($p \leq 0.001$) but three LMs out of six do not present clear BEI-preference (see Table 3 in Appendix). Mostly-aligned preferences between humans and LMs for agency-related conditions suggest that both rely heavily on the agentive/non-agentive features in the tested telic context to construct argument hierarchy as predicted.

We observe an interesting discrepancy between humans and LMs in the responses to the agency-related and telic-condition sentences. Participants scored almost all the agency-related sentences above the reference lines (telic-condition), see Figure 2(A), but the results of the LMs do not present this apparent offset, see Figure 2(B). This discrepancy between human and model results is likely contributed by the differences in the mechanism of human judgment and LM prediction. Masked language models behave as a classifier which assigns probability to BA and BEI in sentence context depending on their relative compatibility to the other tokens in the vocabulary. Therefore, the probability of BA/BEI does not directly reflect the adequacy of the whole sentence. In contrast, participants score the acceptability of each sentence as a whole. Acceptability of other factors inside the sentence such as attached adverbials/subordinate clauses may also play a role in participants’ judgment.

4.3 Individuation

In NP\textsubscript{1}\textsubscript{com}-condition, LMs prefer BA ($p \leq 0.001$) but no significant preference is observed in human judgment for telic context. In NP\textsubscript{2}\textsubscript{com}-condition, humans show BA-preference ($p \leq 0.001$) but three LMs out of six show clear BEI-preference, see Table 3 in Appendix. We compare further between different telic contexts in NP\textsubscript{2}\textsubscript{com}-condition, see Table 1. In temporally telic NP\textsubscript{2}\textsubscript{com}-condition, LMs show a mostly consistent BEI-preference ($p \leq 0.05$ except MacBERT-large) but no significant preference is found in human judgment. In spatially telic NP\textsubscript{2}\textsubscript{com}-condition, humans prefer BA ($p \leq 0.001$)
but inconsistent preference is observed for LMs. These results clearly show that LMs differ from humans in their interpretation of human common NPs like yi-ge gong-ren ‘one-CL worker’ and proper names like Zhang-wei.

A follow-up study is carried out to confirm the negligible influence from yi-ge ‘one-CL’ and examine the thematic relation between common nouns (C, like gong-ren ‘worker’) and proper names (P, like Zhang-wei) in LMs. We focus on the spatially telic context since LMs show a more consistent performance in this context than that in the temporally telic context in telic-condition, as indicated in Table 1. The telic-, NP1-com- and NP2-com-condition in spatially telic context is renamed as P/P-, Ccl/P- and P/Ccl-condition, in the format of "[NP1]/NP2]-condition”. Ccl represents a common noun phrase composed of a numeral, a classifier and a common noun, e.g., yi-ge gong-ren ‘one-CL worker’. For a comprehensive comparison, we add two more conditions Ccl/Ccl and C/P. Table 2 exemplifies all the five conditions.

The BA/BEI-preference of six LMs is obtained for each condition (detailed results see Table 4) and their average $B_{LM}$ is shown in Figure 3. Figure 3 shows consistent negative $B_{LM}$ for P/P-condition ($p \leq 0.01$) and Ccl/Ccl-condition ($p \leq 0.06$ except BERT-base) where subject and object are equal in the degree of individuation (both are P or both are Ccl). This result implies that the spatially telic context is inclined to prefer BA under the condition that both NPs are equal in the individuation degree.

Compared to P/P- and Ccl/Ccl-condition, BA-preference increases (larger negative $B_{LM}$) in Ccl/P-condition and decreases (smaller negative even positive $B_{LM}$) in P/Ccl-condition. The results suggest that the unequal individuation degree between Ccl and P also imposes an effect on the preference. The agentic interpretation of Ccl over P strengthens the BA-preference in Ccl/P-condition and weakens the BA-preference in P/Ccl-condition.

Furthermore, ‘one-CL’ in common NPs shows no significant effect on preference, as C/P-condition agrees with Ccl/P-condition in the BA-preference ($p \leq 0.05$ for all LMs in both conditions). In sum, these results suggest that LMs deliver a more agentic interpretation of the common nouns than that of the proper names in the spatially telic context.

Table 2: Examples of NPs for different conditions with a spatially telic context. (P: proper name, C: common noun, Ccl: common noun phrase with a numeral and a classifier)

| condition | NP1 | NP2 |
|-----------|-----|-----|
| P/P       | ‘guo-ji’ (P) | ‘Zhangwei’ (P) |
| Ccl/Ccl   | ‘yi-ge gong-ren’ (Ccl) | ‘yi-ge si-ji’ (Ccl) |
| Ccl/P     | ‘yi-ge gong-ren’ (Ccl) | ‘Zhangwei’ (P) |
| P/Ccl     | ‘yi-ge gong-ren’ (Ccl) | ‘Zhangwei’ (Ccl) |
| C/P       | ‘gong-ren’ (C) | ‘Zhangwei’ (P) |

Figure 3: Average $B_{LM}$ of six LMs for items with a spatially telic context. The value of zero is set as a reference line.

5 Discussion

This study compares LMs and human behavior in argument hierarchy construction. The results show that LMs and humans perform more similarly with atelic feature than with telic feature. In telic context, LMs and humans show similar behaviour with (non-)agentive features, but differently with individuation-related features. We discuss these (dis)similarities from the following four perspectives.

LMs rely on non-durative property to construct argument hierarchy in a telic context. In telic-condition, spatially telic sentences with adverb dao ‘arrive’ (like 3a) signal non-durative events and show a consistent preference for all LMs, while temporally telic sentences (like 3b) describe durative events and display an inconsistent preference among the LMs. A previous study has suggested that non-duration plays a crucial role for LMs to make telic interpretation (Zhao et al., 2021). Our results further develop the importance of non-durative property: LMs rely more strongly on the non-durative property (compared to durative property) to construct a consistent argument hierarchy.
in a telic context.

**LMs lack sufficient sensitivity in pragmatic function to make the human-like prediction.** BEI has a specific pragmatic function in forming adversative passives which express undesirable, unfortunate events (Li and Thompson, 1989; Chao and Zhao, 1968; Philipp et al., 2008) and often comes with disposal verbs denoting unfavorable meaning like piping ‘criticize’ and da ‘hit’ (Cook, 2019; Wenfang and Susumu, 2013; Loar, 2012). The majority of the temporally telic sentences (7 out of 9) contain disposal verbs whose close connection with BEI may directly contribute to the human BEI-preference in the temporally telic-condition. The pragmatic function of BEI may also increase human BEI-preference for happen-condition. The verb fa-sheng ‘happen’ has a negative prosody (i.e., is likely to occur in a negative context) (Zhang and Ping, 2006; Xiao and McEnery, 2006; Sinclair and Sinclair, 1991), making BEI natural to occur in happen-condition in our results.

However, LMs fail to show sensitivity in this pragmatic function of BEI, as no human-like preference is found for both temporally telic- and happen-condition. Our results are in line with previous study that pre-trained transformer-based LMs have shortage in acquiring pragmatic knowledge (Ettinger, 2020).

**LMs are inclined to interpret the subject as an agent in a spatially telic context.** As both NP1 and NP2 are proper nouns, humans show high acceptability of both BA and BEI in a spatially telic context. It indicates that participants do not interpret argument hierarchy based on the linear position of arguments, at least in Mandarin Chinese (Philipp et al., 2008; Bornkessel and Schlesewsky, 2006), that is, the sequence subject-verb-object does not determine the argument assignment. However, LMs show a clear preference for BA in a spatially telic context where both NPs are common nouns (C1/C1-condition) or proper names (telic-condition), indicating that LMs intend to interpret the subject in the telic context as an agent. This BA-preference in LMs may be explained by 1) unbalanced occurrences between active and passive voice, as more active sentences increase the probability of subjects interpreted as agents, and 2) a higher occurrence frequency of BA over BEI during training. The occurrence frequencies of active/passive and BA/BEI in the LMs’ training corpus worth further investigation.

**Individuation degree plays a different role between LMs and humans in spatially telic context.** Proper names have a higher degree of individuation than common nouns. A proper name is more likely to function as an agent than a common NP (Yamamoto, 1999; Dixon, 1979), which agrees with the results in spatially telic context for humans: 1) BA-preference in NP2-com-condition and 2) high acceptability of BEI in NP1-com-condition.

However, LMs show an opposite tendency in viewing a common NP in spatially telic context as an agent through BA-preference for NP1-com-condition for all LMs. The follow-up study in spatially telic context further confirms the agentive interpretation of common nouns in LMs.

LMs fall short to interpret proper names as agents, which may be attributed to their low occurrence frequency during training. Moreover, almost each character in proper names has separate semantic meanings. We use Zhang-wei as an example. Zhang is usually used as a classifier for flat objects like table and paper and wei forms a number of adjectives meaning great and grand. Therefore, LMs may have difficulty in interpreting the combination of these characters as human names (Lake and Murphy, 2021; Yu and Ettinger, 2020).

6 **Future work**

Note that telic predicates in the agency- and individuation-related conditions are necessary to build items in the Chinese structure "NP1+BA/BEI+NP2+VP" (Xiao et al., 2006), which is also verified by the high acceptability of BA and BEI in telic-condition (low judgment scores in Figure 2(A)) in our experiment. Future work could continue to explore LMs’ sensitivity to agency- and individuation-related features isolated from telic context in syntax-semantics-interface. Moreover, as we treat LMs as a whole and pay attention to their final predictions of BA/BEI to compare with human judgment in our study, more probing measures, such as attention probing, could be taken to deepen our understanding about internal performance of LMs.

---

8 In NP1-com-condition, humans show high acceptability for both BA and BEI as indicated in Figure 2(A). The high acceptability of BA for NP1-com-condition may be contributed by the tendency of BA-construction with a definite NP2 (Ye et al., 2007).
7 Conclusion

This study uses BA/BEI-preference in the Chinese structure “NP1+BA/BEI+NP2+VP” to examine if pre-trained transformer-based language models construct similar argument hierarchy like humans, i.e., the interpretation of Who did what to Whom, with the effect of telicity-, agency- and individuation-related features. The results show that LMs and humans behave similarly for atelic and non-agentive/agentive features, but differently to telic and individuation-related features in the tested context. Specifically, their discrepancy in the temporally telic context suggests that unlike humans, LMs lack sufficient sensitivity to pragmatic function of BEI describing undesirable events with disposal verbs. The different BA/BEI-preference in the sentences with human common vs. proper nouns between LMs and humans indicates that unlike humans who perceive proper nouns as agents, LMs tend to interpret common nouns as agents.

Acknowledgements

We sincerely appreciate anonymous reviewers for their valuable comments and suggestions, which helped us to improve the quality of the manuscript.

References

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.

Ina Bornkessel and Matthias Schlesewsky. 2006. The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. Psychological review, 113(4):787.

Ina Bornkessel, Stefan Zysset, Angela D Friederici, D Yves Von Cramon, and Matthias Schlesewsky. 2005. Who did what to whom? the neural basis of argument hierarchies during language comprehension. Neuroimage, 26(1):221–233.

Yuen Ren Chao and Yuanren Zhao. 1968. A grammar of spoken Chinese. University of California Press.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Angela Cook. 2019. The use of the passive marker bei in spoken mandarin. Australian Journal of Linguistics, 39(1):79–106.

D Alan Cruse. 1973. Some thoughts on agentivity. Journal of linguistics, 9(1):11–23.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. arXiv preprint arXiv:2004.13922.

Monique De Mattia-Viviës. 2009. The passive and the notion of transitivity. Review of European Studies, 1(2):94–109.

Xiangjun Deng, Ziyin Mai, and Virginia Yip. 2018. An aspectual account of ba and bei constructions in child mandarin. First Language, 38(3):243–262.

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.

Robert MW Dixon. 1979. Ergativity. Language, 55(1):59–138.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8:34–48.

Kari Fraurud. 1996. Cognitive ontology and NP form. In Thorstain Fertheim and Jeanette K. Gundel, editors, Reference and Referent Accessibility, pages 65–88. John Benjamins Publishing Company.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. arXiv preprint arXiv:1903.03260.

Jeffrey S Gruber. 1967. Look and see. Language, 43(4):937–947.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. arXiv preprint arXiv:1803.11138.

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.

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. Language, 56(2):251–299.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised mary because he? implicit causality bias and its interaction with explicit cues in lms. arXiv preprint arXiv:2106.01060.

Brenden M Lake and Gregory L Murphy. 2021. Word meaning in minds and machines. Psychological Review.

Charles N Li and Sandra A Thompson. 1989. Mandarin Chinese: A functional reference grammar, volume 3. Univ of California Press.
Robert D Van Valin Jr. 1990. Semantic parameters of split intransitivity. Language, 66(2):221–260.

Zeno Vendler. 1957. Verbs and times. The Philosophical Review, 66(2):143–160.

Susanna Virtanen. 2015. Transitivity in Eastern Mansi: An information structural approach. Ph.D. thesis, University of Helsinki.

Fan Wenfang and Kuno Susumu. 2013. Semantic and discourse constraints on Chinese bei-passives. Linguistics and the Human Sciences, 8(2):205–240.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? arXiv preprint arXiv:1809.00042.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Mita Zhang. 2013. Classifier Structures in Mandarin Chinese. De Gruyter Mouton.
A Appendix

Figure 4: Coefficient of variation of human Accep (A) and Surp averaged across six LMs (B) for each condition with BA and BEI. We find that in human Accep, the preferred one between BA and BEI shows a higher coefficient than the other one (e.g., the do-condition prefers BA and BA has a higher coefficient than BEI) for all the conditions except for telic-condition. In telic-condition where both BA and BEI are high acceptable in human judgment, their coefficients are also at a relatively high level. LMs show a similar trend.
Figure 5: Average $Surp$ for BERT-base, ELECTRA-large, ERNIE1.0, MacBERT-large and MacBERT-base. The values from the telic-condition are set as reference lines.

| Factor | Condition | Humans | BERT-base | RoBERTa-base | ELECTRA-large | ERNIE 1.0 | MacBERT-base | MacBERT-large |
|--------|-----------|--------|-----------|--------------|--------------|-----------|--------------|---------------|
| Telicity (Sec. 2.1) | atelic | bei*** | bei*** | bei*** | bei*** | bei*** | bei*** | bei*** |
|        | telic | bei*** | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |
| Agency (Sec. 2.2) | aim | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |
|       | do | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |
|       | happen | bei*** | ba*** | bei*** | – | – | bei*** | bei*** |
|       | non-volition | bei*** | bei*** | bei*** | bei*** | bei*** | bei*** | bei*** |
|       | volition | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |
| Individuation (Sec. 2.3) | NP2 COM | ba*** | bei*** | – | bei*** | bei*** | – | – |
|       | NP1 COM | – | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |

Table 3: Preference comparison between BA and BEI for humans and LMs for telicity-, agency- and individuation-related conditions (ba: statistically significant BA-preference, bei: statistically significant BEI-preference. Formula: $Surp/\text{Accept} \sim \text{BA}/\text{BEI} + (1/\text{verb})$. ∗ : $p \leq 0.05$, ∗∗ : $p \leq 0.01$, ∗∗∗ : $p \leq 0.001$)

| Speically telic context | BERT-base | RoBERTa-base | ELECTRA-large | ERNIE 1.0 | MacBERT-base | MacBERT-large |
|-------------------------|-----------|--------------|--------------|-----------|--------------|---------------|
| P/P-condition | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |
| Ccl/P-condition | – | ba** | ba** | ba** | ba** | ba** |
| P/C-cl-condition | ba*** | ba*** | ba*** | ba*** | ba*** | ba*** |
| C/P-cl-condition | – | ba*** | beii*** | ba** | ba** | ba** |

Table 4: Preference comparison between BA and BEI for LMs for individuation-related conditions in Section 4.3 (ba: statistically significant BA-preference, bei: statistically significant BEI-preference. Formula: $Surp \sim \text{BA}/\text{BEI} + (1/\text{verb})$. ∗ : $p \leq 0.05$, ∗∗ : $p \leq 0.01$, ∗∗∗ : $p \leq 0.001$, 0.05 ≤ m ≤ 0.06)

| LMs | Tasks | Chinese Database |
|-----|-------|------------------|
| BERT-base | MLM, next sentence prediction | 25M sentences (Devlin et al., 2018) |
| ERNIE 1.0 | MLM, dialogue, language model task | 173M sentences (Sim et al., 2019) |
| RoBERTa-base | MLM | 5.4B words (Cui et al., 2020) |
| ELECTRA-large | replaced token, detection task | 5.4B words (Cui et al., 2020) |
| MacBERT-base/large | MLM as correction, sentence-order prediction | 5.4B words (Cui et al., 2020) |

Table 5: Comparison between models with respect of tasks in their pre-training process and size of Chinese database (MLM: masked LM task).