Syntactic locality in Chinese in-situ and ex-situ wh-questions in transformer-based deep neural network language models

Giuseppe Samo (Beijing Language and Culture University)
Xu Chen (Qufu Normal University)

1 Introduction: Wh-in-situ & Wh-ex-situ in Chinese

Chinese is predominantly considered to be a wh-in-situ language (Huang 1982 *inter alia*), in contrast to wh-ex-situ languages in which the interrogative element is fronted to the beginning of the clause, such as English or Italian (Chomsky, 1977; Rizzi, 1997). Tendentially, wh-ex-situ languages ban lack of movement (cf. echo questions, see Badan et al. 2017), and, as well, wh-in situ languages ban or minimize syntactic wh-movement (Cheng 1991), although intralinguistic variation exists, such as the familiar case of French and some northern Italian dialects (cf. Cheng and Rooryck 2000, Bonan 2021: ch. 1 and reference therein for a detailed review). Let us take, as an initial reference, the asymmetry among the four Chinese examples in (1).

(1) a. ni zuotian mai le shenme?
you yesterday buy PAST what

b. ni zuotian mai le na-ge shuiguo?
you yesterday buy PAST which-CL fruit

c. ??/* shenme, ni zuotian mai le?
what you yesterday buy PAST

d. ?? na-ge shuiguo, ni zuotian mai le?
which-CL fruit you yesterday buy PAST

In natural settings, Chinese speakers do allow the wh-element to occur in the sentence initial position, although their frequency is quite rare (Wu, 1999). While in (1a) and (1b) wh-in-situ is an unmarked option, the ex-situ in (1c) and (1d) can be considered as partially marginal (??) or ungrammatical (*) (Pan, 2014). Indeed, a binary delimitation on grammaticality in Chinese—with wh-in-situ grammatical and wh-ex-situ ungrammatical—seems to be too coarse because on the basis of informal judgement, the gradation of acceptability varies according to the environment (e.g. the pronominal or XP nature of the fronted element or of the subject), where the ex-situ wh-item occurs.

Besides, bare wh-item and complex wh-item also exhibit differently, as the contrast between (1c) and (1d) shows. No crossing and therefore no standard intervention effects emerge for both types of wh-in-situ. Being an element fronted from a generation position (Belletti 2018 for an overview, and reference therein), elements of syntactic locality (Rizzi, 2004) are triggered. If the fronted object and the intervening subject differ, sentences are parsed better by speakers (Friedmann et al. 2009 and related works). For example, while in (1c) the pronominal element (henceforth, PRO) *shenme* crosses a pronominal subject *ni* 'you', in (1d), the pronoun *ni* 'you' is crossed by a dissimilar in nature item, the maximal projection (henceforth, XP) *na-ge shuiguo* 'which fruit' (see Samo and Merlo 2021 for a computational approach of locality exploring grammatical sentences in large-scale datasets).

In this paper, we test how wh-ex-situ, and their dimensions with respect to the complexity/simplicity of the fronted elements and their dissimilarity with the intervening subjects, are parsed by the mechanistic model provided by artificial neural networks (Linzen and Baroni, 2021; Wilcox et al., 2018), following the line of works in which word embeddings models have been tested for syntactic locality and intervention effects (Merlo 2019; Merlo and Ackermann 2018 *inter alia*). Beyond the pure interest in language modelling, our results can be harvested by theoretical syntacticians, in terms of how many lexical elements may encode syntactic properties. We briefly introduce locality in language models in section 2, while the study and a discussion will be presented in section 3. Section 4 concludes.

2 Syntactic Locality: predicting asymmetries

We use neural networks as a language model (Hale 2016 and reference therein): the parser, after observing the sequence of words in a sentence, outputs probability to the words that would follow. This probability is assigned on the basis of a training derived from exposure to a large amount of non-annotated (raw) texts. Language models, tested with different architectures, have been proven to essentially capture syntactic properties (Gulordava et al. 2018; Linzen and Baroni 2021 for an overview; but see also Sinha et al. 2021; Wallat et al. 2021; Chaves and Richter 2021, *inter alia*), such as long-distance agreement. Syntactic locality has been also tested (Merlo 2019; Merlo and Ackermann 2018 and related works), although in a small set of languages, mainly English and French.

We will explore the deep multi-layer bidirectional model BERT Devlin et al. (2018). The metrics we explore is surprisal (cf. Levy 2008; Hale 2016; Wilcox et al. 2018), the logarithm of the reciprocal of the probability output to a given word. We mainly test two hypotheses, stated in H_1 and H_2.

H_1: In-situ wh-questions should present lower surprisal than ex-situ wh-questions at the completion of the parsing sentence.

We adopt the notion of *feature match* mutated from Samo and Merlo (2021).

FEATURE MATCH A *feature match*, $match_t(Wh, S)$, is true iff, for a given feature type t (values: XP, PRO), the wh-object Wh and the intervener S are instantiated and have the same value.

We expect that wh-ex-situ in matching configurations (M) should be parsed worse than mismatching ones. Moreover, we expect that mismatching configurations (MM) in which the fronted element is a maximal projection crossing a pronominal element (henceforth MMxp-pro) should be

rated better than pronominal wh-elements crossing a nominal subject (MMpro-xp), as predicted by a theory of intervention locality ((Rizzi, 2004; Friedmann et al., 2009)).

H_2: Mismatching ex-situ configurations should show lower surprisal than matching ex-situ configurations.

3 The study

Experimental stimuli The experimental items are based on 224 stimuli inputed as Chinese characters. Every stimulus is split into four syntactic regions. The first three regions are filled by a syntactic constituent (subject (S), verb (V), wh-object (WH)) in two different orders according to the in-situ properties (S,V,WH; WH,S,V), while the fourth region is dedicated to the question mark. Figure 1 offers an example of a stimulus for every condition.[1]

				Regions			
Type	WH-type	Locality	Intervention	Region 1	Region 2	Region 3	Region 4
In-situ	bare	Matching	No	ni_{Pro}	*kanjianle*	$shenme_{Pro}$	?
In-situ	complex	Mismatching	No	ni_{Pro}	*kanjianle*	$nabenshu_{XP}$	?
In-situ	bare	Matching	No	$laoshi_{XP}$	*kanjianle*	$shenme_{Pro}$	?
In-situ	complex	Mismatching	No	$laoshi_{XP}$	*kanjianle*	$nabenshu_{XP}$	?
Ex-situ	bare	Matching	YES	$shenme_{Pro}$	ni_{Pro}	*kanjianle*	?
Ex-situ	complex	Mismatching	YES	$nabenshu_{XP}$	ni_{Pro}	*kanjianle*	?
Ex-situ	bare	Matching	YES	$shenme_{Pro}$	$laoshi_{XP}$	*kanjianle*	?
Ex-situ	complex	Mismatching	YES	$nabenshu_{XP}$	$laoshi_{XP}$	*kanjianle*	?

Figure 1: Syntactic Regions and one example (romanized in pinyin) for every condition. (Pro = pronoun, XP = maximal projection)

Methods We adopt the pre-trained model of type transformer Chinese BERT (bert-base-chinese).[2] The models output a surprisal measure in a fill-mask task, consisting of hiding target words from a structure (see also the detailed discussion in Renaud 2020)[3]. We test H_1 by calculating the surprisal on the fourth region, the one dedicated to the question mark, by considering the question mark as the locus of full parsing of the sentence. Following the discussion on the intervener in Samo and Merlo, we test H_2 by measuring the surprisal in the intervention locus for ex-situ questions, namely the intervening subject (region 2).[4]

Results and discussion Our data analysis is based on 896 datapoints. H_1 is confirmed as visualized in Figure 2 (left panel). As a matter of fact, wh-in-situ sentences are evaluated with a lower surprisal (M = 0.898, SD = 0.369) than ex-situ wh-questions (M = 4.778, SD = 2.039) in the fourth

[1] The relevant .csv files of the experimental sentences have been processed by python (Python Core Team, 2019). All the processed output has been analyzed with R (R Core Team, 2017) for statistical analysis.

[2] `https://huggingface.co/bert-base-chinese` (03/2022)

[3] We use an edited version of the code discussed in Renaud (2020), available at the following address: `https://github.com/celine-renaud/Memoire` (last accessed 03/2022).

[4] Wh-in-situ cannot be adopted as control group since there is no pure postulated intervention locus.

region/question mark ($t(111) = 19.82179$, $p < 0.0001$). This result is expected from the grammar intuition of speakers creating the bias for the large-scale data as training.

H_2 is also confirmed. Wh-ex-situ in mismatching configurations are overall favoured in terms of lower suprirsal ($M = 9.730$, $SD = 4.313$) than matching configurations ($M = 10.132$, $SD = 4.216$), but these results are not statistically significant ($t(55) = 1.24673$, $p = .215146$). The quality of the typology of syntactic locality emerges, as given in 2 right panel. In line with the native speakers judgements in (1d) and theoretical considerations in syntactic locality (Rizzi, 2004), complex wh-elements crossing prononimal subjets (MMxppro) ($M = 8.378$, $SD = 3.467$) are rated better than prononimal crossing maximal projections (MMproxp) ($M = 14.370$, $SD = 1.918$) in the second (the intervention region) ($t(53) = 8.915$, $p < .00001$). If we take into consideration the four groups (Mxpxp, Mpropro, MMxppro, MMproxp), a one-way ANOVA revealed that there was a statistically significant difference in surprisal ($F(3,107) = [31.979]$, $p < .00001$). Furthermore, a post hoc Tukey test showed that MMxppro differed significantly at $p < .0001$.

Our results can be interpreted in different perspectives. As demonstrated in works on agreement (Linzen et al. 2016; Gulordava et al. 2018; Wilcox et al. 2018), language models seem discriminating between grammatical and ungrammatical/marginal sentences The results for H_1 extend this trend, adding further empirical evidence from Chinese. On the other hand, the results for H_2 should be harvested from syntacticians, especially on the role that lexical elements and their distribution demonstrated from probabilistic models with respect to syntactic locality. Mismatching configurations, marginally favoured by native speakers (see 1d) are also preferred by transformer-based deep neural network language models.

Finally, our findings can also shed light on the non-movement approach to long-distance dependencies, such as HPSG that is modeled with the notion of feature percolation (Pollard and Sag 1994 and much subsequent work), according to which the SLASH feature of the gap is passed up to the filler, a path that may be intervened by the intermediate nodes with various types of features (Chaves, 2021).

4 Conclusions and future studies

Our results can feed novel research questions in formal approaches, with respect to intervention effects in ex-situ questions. For example, we aim to map if the gradient of acceptability judgments by native speakers coincides in some manner with the scale of surprisal encoded by the mechanistic models, along with the structural configurations in both in-situ to ex-situ wh-questions. We leave this matter for future research.

Bibliography

Badan, L., S. Gryllia, and G. Fiorin (2017). Italian echo-questions at the interface. *Studia Linguistica 71*(3), 207–240.

Belletti, A. (2018). Locality in syntax. In *Oxford Research Encyclopedia of Linguistics*, Oxford.

Bonan, C. (2021). *Romance Interrogative Syntax: Formal and typological dimensions of variation*. John Benjamins.

Chaves, R. P. (2021). Island phenomena and related matters. In S. Müller, A. Abeillé, R. D. Borsley, and J.-P. Koenig (Eds.), *Head-Driven Phrase Structure Grammar: The handbook*, pp. 665–723. Berlin: Language Science Press.

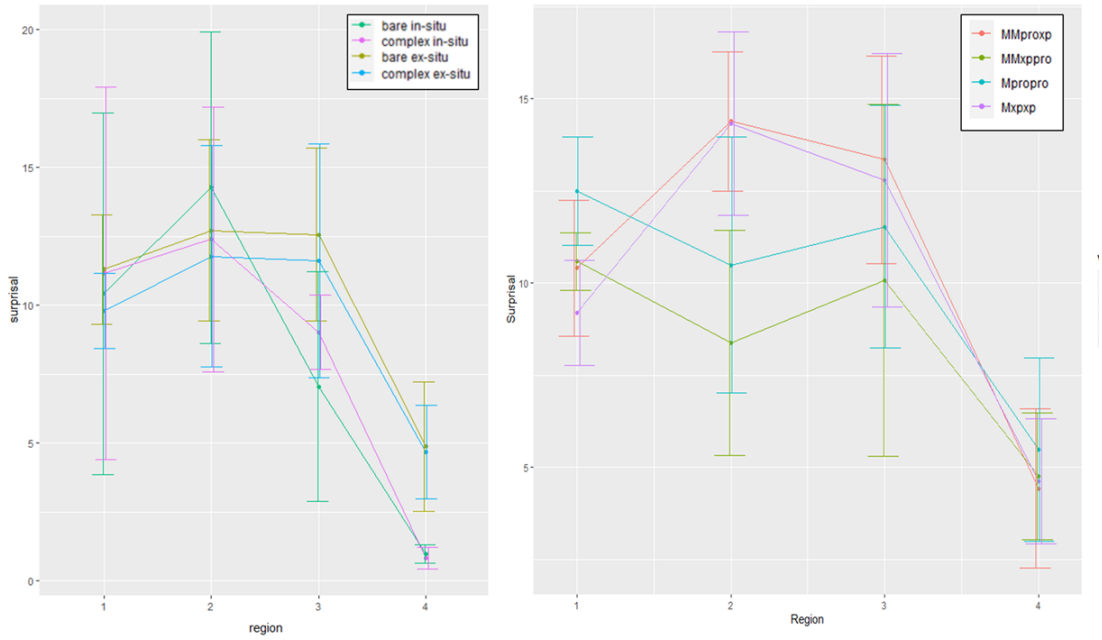

Figure 2: Mean surprisal and syntactic regions for wh-ex-situ and wh-in-situ (left panel) and within-groups of wh-ex-situ (right panel)

Chaves, R. P. and S. N. Richter (2021). Look at that! bert can be easily distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics 4*(1), 28–38.

Cheng, L. L. and J. Rooryck (2000). Licensing wh-in situ. *Syntax 3*, 1–19.

Cheng, L.-S. L. (1991). *On the typology of wh-questions*. Ph. D. thesis, MIT.

Chomsky, N. (1977). On wh-movement. pp. 71–132. New York: Academic Press.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Friedmann, N., A. Belletti, and L. Rizzi (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua 119*(1), 67–88.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen, and M. Baroni (2018). Colorless green recurrent networks dream hierarchically. *CoRR abs/1803.11138*.

Hale, J. (2016). Information-theoretical complexity metrics. *Lang. Linguistics Compass 10*, 397–412.

Huang, C.-T. J. (1982). *Logical relations in Chinese and the theory of grammar*. Ph. D. thesis, Massachusetts Institute of Technology.

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

Linzen, T. and M. Baroni (2021). Syntactic structure from deep learning. *Annual Review of Linguistics 7*, 195–212.

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

Merlo, P. (2019, August). Probing word and sentence embeddings for long-distance dependencies

effects in French and English. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy, pp. 158–172. ACL.

Merlo, P. and F. Ackermann (2018, October). Vectorial semantic spaces do not encode human judgments of intervention similarity. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, Brussels, Belgium, pp. 392–401. Association for Computational Linguistics.

Pan, V. J. (2014). Wh-ex-situ in mandarin chinese: Mapping between information structure and split cp. *Linguistic Analysis 39*(3-4), 371–414.

Pollard, C. and I. A. Sag (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Python Core Team (2019). *Python: A dynamic, open source programming language*. Python Software Foundation. Python version 3.7.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Renaud, C. (2020). Traitement des relations longue distance par les réseaux de neurones en anglais, français et italien. *Mémoire de maîtrise, Université de Genève*.

Rizzi, L. (1997). The fine structure of the Left Periphery. In L. Haegeman (Ed.), *Elements of grammar*, pp. 281–337. Dordrecht: Kluwer Academic publisher.

Rizzi, L. (2004). Locality and left periphery. In A. Belletti (Ed.), *Structures and beyond*, Volume 3 of *The cartography of syntactic structures*, pp. 223–251. Oxford, NY: OUP.

Samo, G. and P. Merlo (2021). Intervention effects in clefts: a study in quantitative computional syntax. *Glossa: a journal of general linguistics 6(*.

Samo, G. and P. Merlo (to appear). Distributed computational models of intervention effects:a study on cleft structures in french. In C. Bonan and A. Ledgeway (Eds.), *It-Clefts: Empirical and Theoretical Surveys and Advances*, pp. pp–pp. Mouton de Gruyter.

Sinha, K., R. Jia, D. Hupkes, J. Pineau, A. Williams, and D. Kiela (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Wallat, J., J. Singh, and A. Anand (2021). Bertnesia: Investigating the capture and forgetting of knowledge in bert. *arXiv preprint arXiv:2106.02902*.

Wilcox, E., R. Levy, T. Morita, and R. Futrell (2018, November). 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*, Brussels, Belgium, pp. 211–221. ACL.

Wu, J. (1999). *Syntax and semantics of quantification in Chinese*. Ph. D. thesis, University of Maryland, College Park.