Information Locality in the Processing of Classifier-Noun Dependencies in Mandarin Chinese

Introduction Both expectation and locality have been established as key factors in characterizing incremental processing difficulty in natural language [1-4]. A complete theory of sentence processing must therefore take into consideration both factors and model their potential interactions. To address this, lossy-context surprisal (LCS) [5-6] proposed that a word's processing is proportional to its surprisal given a lossy memory representation of the context (formula in 1), instead of a perfectly retained context assumed by standard Surprisal theory [2]. One key prediction LCS makes is the so-called *information locality*, whereby expectation effects become weaker with stronger locality constraints. However, this prediction was not always borne out, and some studies found the opposite pattern of interaction, as expectations played a bigger role under strong locality constraints [7-8]. To resolve this, we present three experiments and a computational model on the processing of Mandarin Chinese (MC) classifier-noun dependencies.

In MC, a noun in certain contexts must take a classifier. While different nouns are compatible with different SPECIFIC classifiers, there is a GENERAL classifier GE that almost all countable nouns can take. Although the two options can be used interchangeably [9], specific classifiers render it easier to predict the upcoming noun. This contrast allowed us to create varying levels of expectation. We in addition manipulated dependency length to create different locality constraints.

Experiments We crossed Classifier (specific vs. general) and Distance (local vs. 1RC vs. 2RC) in a 2X3 design (ex. in 3). *Expectation*-based theories predict shorter RTs with specific classifiers, while locality-based theories longer RTs with longer distance. LCS additionally predicts that the processing facilitation from specific classifiers gets weaker as distance increases. For all statistical analysis, we fitted maximal Bayesian hierarchical RTs on log-transformed RTs of the noun. **Exp1** (n=96) was conducted in the lab using SPR. Its results (**Fig.** 1) provided evidence for locality effects, whereby the local and 1RC conditions combined were read faster than the 2RC condition (β=-0.113, CrI=[-0.157, -0.067]). Surprisingly, there was no evidence for effects of classifier. Follow-up analysis suggests that specific classifiers only facilitated processing in the 2RC condition, but not in local or 1RC conditions. We suspected that this was due to spillover from specific classifiers, as they could be harder to process due to lower frequency than the general one. We therefore conducted **Exp2** (n=80) online using A-Maze (illustrated in **Fig.** 2; [10]), since it has been shown to provide more localized measures than SPR [11]. Results (**Fig.** 3) again showed support for locality effects (β=-0.058, CrI=[-0.112, -0.003]). Moreover, we found a main effect of classifier, whereby the specific classifier condition was read faster (β=-0.113, CrI=[-0.157, -0.067]), suggesting that A-Maze successfully avoided spillover. Most importantly, there was evidence for an interaction effect (β=0.058, CrI=[0.018, 0.098]), suggesting that the processing facilitation from specific classifiers became weaker in the 2RC condition, consistent with *information locality*. **Exp3** (n=96; **Fig**. 4) replicated all the basic patterns of Exp2 in the lab.

Computational Model To capture these results, we implemented the resource-rational LCS (RR-LCS) model of [6] to MC. The lossy context representations were optimized to minimize the average model surprisal over large-scale text data, while constraining the overall fraction of deleted words among a context window of 20 words. We then estimated P(w|c) using a Chinese GPT-2 model [12]. The RR-LCS model showed similar results (**Fig**. 5) as in human experiments, whereby the processing differences between the two types of classifiers become smaller under stronger locality constraints, as the representation of classifiers may get deleted as distance increases. This tendency is strengthened as the rate of deleted words gets higher.

Conclusion Echoing LCS, we found experimental support that the processing of classifier-noun dependencies in MC is subject to information locality constraints. Our newly implemented RR-LCS model successfully captured these results. Overall, we show that probabilistic expectations are constrained by memory limitations and that future theory building in sentence processing should take this into consideration. However, we do note that expectation and locality can interact differently in different languages. We leave this for future work.

(1) processing difficulty $\propto -\log P(w|c') = -\log\sum_c P(w|c)P(c|c')$, where $P(w|c)$ is the probability of a word given a lossy (i.e., non-veridical) context, and $P(c|c')$ is the posterior probability of the context c given the lossy representation c'.

(2) An example set of items (total n = 36):

Local conditions (the classifier and the noun are adjacent to each other; ZHANG is a specific classifier that matches the critical word desk, and GE is the general classifier):

Mali tingshuo na-liang-ZHANG/GE **zhuozhi** duzhu-le lu.
Mary hear that-two-ZHANG/GE **desk** block-PER road
"Mary heard that those two desks blocked the road."

1RC conditions (the classifier and the noun are intervened by one prenominal RC):

Mali tingshuo na-liang-ZHANG/GE bei Anna nuozou de **zhuozhi** du-le lu
Mary hear that-two-ZHANG/GE PASS Anna move REL **desk** block-PER road
Mary heard that those two desks that LiuNa moved blocked the road."

2RC conditions (the classifier and the noun are intervened by two prenominal RCs):

Mali tingshuo na-liang-ZHANG/GE bei culude ganzou-le Kaite de Anna nuozou
Mary hear that-two-ZHANG/GE PASS rudely chase-PERF Katie REL Anna move
de **zhuozi** du-le lu
REL **desk** block-PER road
"Mary heard that those two desks that Anna that rudely chased away Katie moved blocked the road."

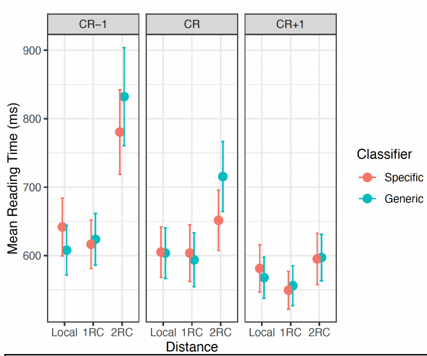

Fig. 1: Results for **Exp1** at the critical region, one region before after

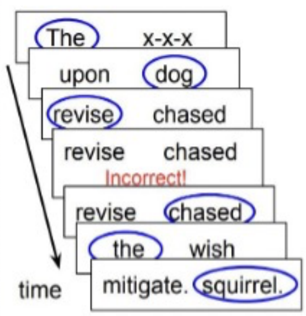

Fig. 2: An illustration of the Maze task [10]

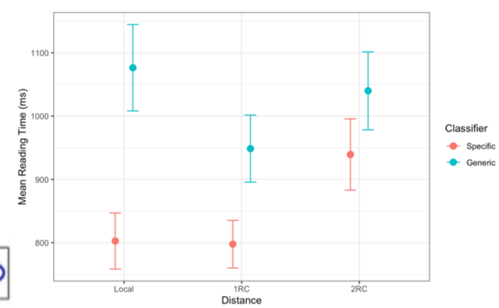

Fig. 3: Results for **Exp2** at the critical region

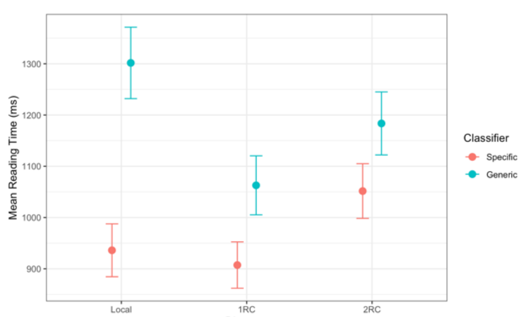

Fig. 4: Results for **Exp3** at the critical region

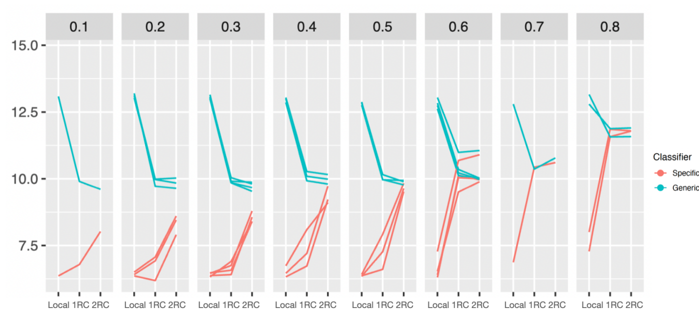

Fig. 5: Model surprisal calculated from Chinese GPT2 (top panel shows different deletion rates)

References [1] Hale (2001) NAACL. [2] Levy (2008) Cognition. [3] Gibson (1998) Cognition. [4] Lewis & Vasishth (2005) Cogn. Sci. [5] Futrell et al. (2020) Cogn. Sci. [6] Hahn et al. (2022) PNAS. [7] Husain et al. (2014) PLoS One. [8] Schwab et al. (2022) JEP. [9] Ma (2015) CSS Press. [10] Boyce et al. (2020) JML. [11] Witzel et al. (2012) J. Psycholinguist. Res. [12] Zhao et al. (2019) EMNLP.