Parallel processing of Chinese words depends on both *n*-gram frequency and structure

Kuan-Jung Huang Psych and Brain Sciences, UMass Amherst kuanjunghuan@umass.edu

In written Chinese, a character may represent either a word or an embedded morpheme, but there are no spaces to mark word boundaries. Chinese readers therefore do not always agree on how a string of characters should be segmented into word units in an offline decision task [1]. Recently, [2] showed fewer/shorter eye fixations for 4-character strings that tend to be considered offline as one word. The current study aims to uncover what linguistic properties determine whether a string of Chinese characters is *processed* as one unit. We examine two factors argued to be critical to phrase (multi-word) processing: *n*-gram frequency, under the usage-based view [3], and syntactic structure, under the composition-based view [4].

Our strings varied widely in *n*-gram frequency (based on which we later constructed subgroups; see below), but group means were matched across structures (Table 1). Mean number of strokes and frequency of the constituents were also matched across structures. Exp. 1 used 2-character strings, with frequency measured from a corpus of 135M words [5]. Critical comparison was between noun-noun (NN) and verb-object (VO) strings, all of which were semantically composable. Also included were verb-verb (VV) and subject-verb (SV) strings, which were non-composable. Exp. 2 used 4-character strings, with frequency measured from a 100M character corpus scraped from *Weibo* and *Huanqiurenwu*. Critical comparison was among NN, VO, and SV strings. Also included were non-composable VNVV and VVNV strings.

We used a part-of-speech judgment task. On each trial, subjects were shown a character string for 17 ms, followed by a post-mask (200 ms). An eye-tracker ensured they always fixated centrally during string presentation. Subjects judged only one of the characters, with its position (varying randomly across trials) indicated by a yellow square (Fig. 1). All the probed characters are morphemes that can stand alone as one word. These characters were separately normed; trials with characters not consistently judged as a noun or a verb (consistency < 75%) were excluded (10%). The critical task manipulation, which was blocked, was the presence/absence of a pre-cue signaling which character would be probed: pre-cued (PC) vs. no-precue (NPC). In the PC condition, subjects could deploy their covert attention to the to-be-probed character, while in the NPC condition, they needed to divide their attention. Using this paradigm, we can determine whether a string's embedded constituents *can* be processed in parallel [6]. If accuracy in NPC is similar to PC (i.e., no divided attention deficit), this means the characters are processed fully in parallel. A substantial drop in accuracy in NPC would reflect serial processing, while a reliable but smaller drop in accuracy would indicate some parallel processing, but with a fixed capacity. The predictions can be quantified with attention-operating-characteristic curves [7] (Fig. 2). Serial processing of the two constituents within a string would suggest a lack of a stored/lexicalized representation of the whole string.

Subjects were native Chinese speakers living in the US with >12 years of previous education in China. In Exp. 1 (N=39), subjects could process NN and VO strings almost fully in parallel (Fig. 3a). In contrast, non-composable VV and SV strings showed less parallelism or full seriality. The greater parallelism in VV than in SV might be due to a POS congruency effect [6]. In Exp. 2 (N=35), with 4 characters, subjects still could process NN and VO in parallel (Fig. 3b). Furthermore, SV were now processed in parallel too. Finally, we split our items into frequently-co-occurring (FCO) and barely-co-occurring (BCO) sets (top and bottom 40 items ranked by *n*-gram frequency). Figure 4 shows a frequency effect interacting with structure: all FCO strings were processed in parallel, BCO SV trended toward seriality, while BCO NN and VO showed modest parallelism.

We use serial vs. parallel processing as evidence of (lack of) lexicalization and investigate what underlies lexicalization. Our *n*-gram frequency effect aligns with a view that multiword strings are stored [3, 8] (note that since our embedded constituents' frequencies were closely matched, *n*-gram frequency necessarily highly correlated with forward and backward transitional probability; we remain agnostic as to which specific statistical information is relevant [3]). Throughout the two experiments, NN and VO resulted in the same pattern, which could be attributed to bare-noun incorporation in Chinese [9]. Notably, even SV, which clearly nests two syntactic units, can pattern like NN, highlighting the role of frequency. However, for BCO strings, the effect of structure emerged, which could reflect either on-the-fly structure building or recruitment of stored structured abstract syntactic objects [10], favoring NN and VO over SV.

References: [1] Liu et al. (2013) *PlosOne* [2] He et al. (2021) *Brit J Psych* [3] McCauley & Christiansen (2019) *PsychRev* [4] Chomsky (1975) *Reflections on Language* [5] Sun et al.

(2018) *BehResMeth* [6] White et al. (2020) *AttPerceptPsych* [7] Sperling & Melchner (1978) *Science* [8] Zang (2019) *Vision* [9] Luo (2022) *J EastAsianLing* [10] Yu et al. (2024) *bioRXiv*

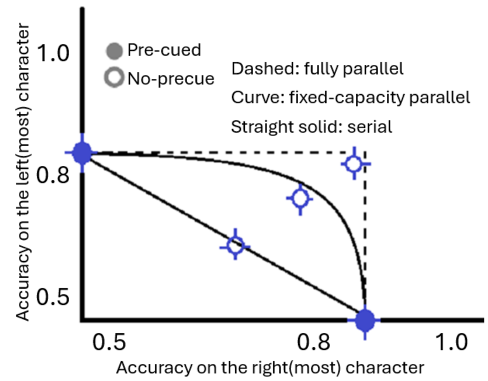

Figure 1. Illustration of the paradigm. Half-way through Exp 2 we changed the pre-cue duration from 750 to 51 ms due to a concern that maintaining central fixation for 750 ms is too taxing. Doing so yielded only a minimal difference in accuracy, with no qualitative changes in the serial/parallel patterns.

Figure 2. Theoretical predictions for NPC based on PC accuracy [7].

Exp. 1 (2-char)	**Example**	**Mean frequency (per million word)**	**# of items**
NN	Night-View	3.18	160
VO	Chop-Wood	3.55	160
SV	Sauce-Chat	< 0.001	64
VV	Buy-Wait	< 0.001	128
AA	Flat-Sharp	< 0.001	32
Exp. 2 (4-char)	**Example**	**Mean frequency (per million char)**	**# of items**
NN (NNNN)	Duke-Name	0.553	112
VO (VVOO)	Repair-Skin	0.523	112
SV (SSVV)	Price-Rise	0.533	112
VNVV	Sell-Word-Accuse	0	56
VVNV	Donate-Shirt-Return	0	56
ANNA	Fierce-Feather-Shiny	0	56

Table 1. Description of items. AA and ANNA not analyzed.

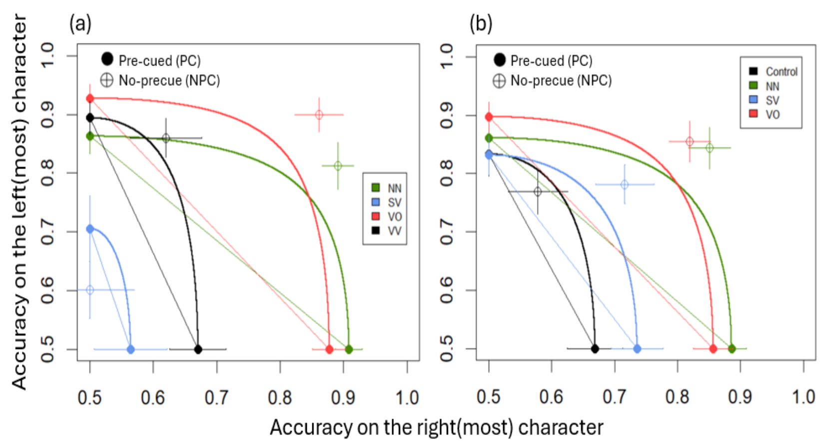

Figure 3. (a) Exp. 1's results (b) Exp. 2's results. Control = VNVV + VVNV. Bars reflect by-subject 95% CIs.

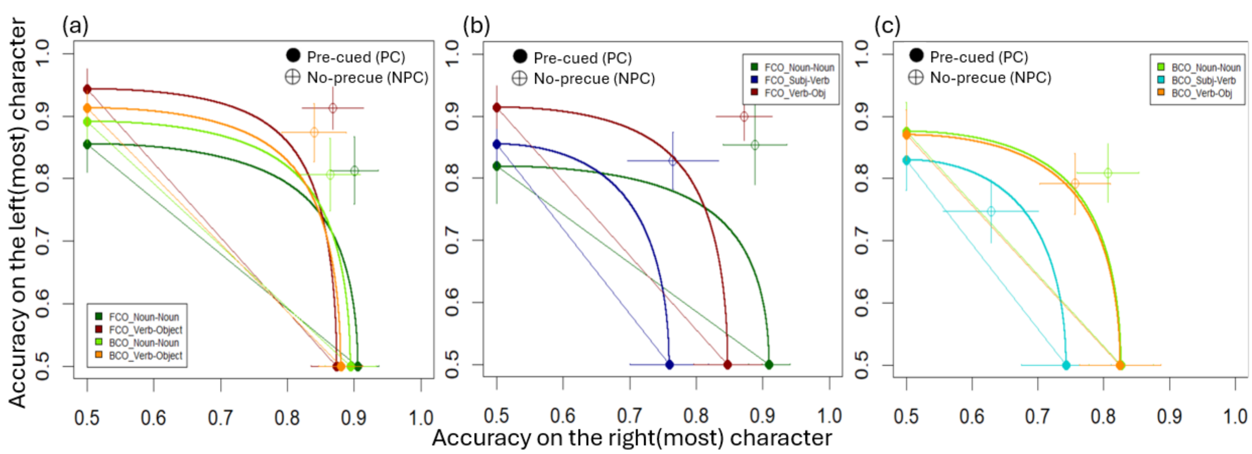

Figure 4. (a) Exp. 1's results after dividing items into FCO and BCO. (b-c) Exp. 2's results after item division, made into 2 separate plots for ease of visualization. The farther the bullseye falls outside the curve, the greater the parallel processing is for that item group.