Are you asking me or telling me? Using linguistic diversity to explore foundations of perspective taking in language comprehension

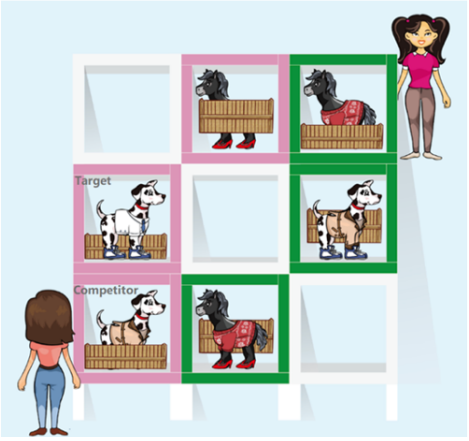
Yingjia Wan[1]; Yipu Wei[2]; Craig Chambers[1] ([1]University of Toronto; [2]Peking University)

Although classic accounts of perspective-taking claimed that **common ground** (shared knowledge) is essential for successful comprehension, contemporary work shows that, not only do core processes draw on multiple perspectives [1], but also that the needed perspective computations vary according to utterance goals: e.g., whereas *imperatives* ("Pick up the big…") relate primarily to **shared** referents, *questions* ("What's above the cow with the…?") require listeners to consider **privileged** referential candidates [2]. Thus, recognizing the relevant speech act (e.g., question, command, assertion…) seems critical for calculating perspective. In languages like English, lexical-syntactic and prosodic cues usually allow listeners to differentiate these speech acts from utterance onset. But are these cues essential, or could perspective be efficiently calculated without them? We address this question by leveraging a feature of Mandarin Chinese, namely that utterances can be ambiguous in their status as a statement until their final constituent. This is because a wh-word such as *shénme* ("what") is expressed in situ, entailing sentence pairs with structures such as **woman-wear-is-what** (Engl: "What is the woman wearing?") and **woman-wear-is-shoes** (Engl: "The woman is wearing shoes"—note that prosody is not an early disambiguator [3]). Thus, if speech act identification drives the framing of perspective, comprehension should be delayed in Mandarin. However, because speech acts are themselves by-products of a broader communicative ecosystem (i.e., the context provides the motivations for interlocutors to provide or seek information), higher-level task goals might independently steer attention to knowledge held by a speaker vs. listener.

To assess this, we used a two-player Visual World task based on [2], where displays contained cartoon animals wearing clothes and footwear. Critically, visual barriers entailed that different clothing and footwear items were occluded for a given player (see Fig. 1). At the end of each trial, players needed to answer verification questions posed by the experimenter (e.g., *How many horses are wearing dresses?*) to ensure they would query the other player about items unknown to them and describe items unknown to the other player. Importantly, one player was a confederate who produced scripted utterances on critical trials. These were initially ambiguous (i.e., in terms of speech act, with their status as a question/ statement being disambiguated by the utterance-final word) and entailed the anticipation of different referents depending on which speech act was incrementally inferred. The measure of interest was the participant player's anticipatory gaze to the target (vs. a competitor), before hearing the utterance-final word. We manipulated task knowledge across two conditions (between participants). In the **structured task** condition, the participant player was tasked with answering verification questions about clothing, whereas the confederate player was responsible for footwear. Thus, each player could expect the kind of information the other player would seek vs. provide. This provides a higher-level cue that, in principle, could be used to infer whether a co-player's utterance is likely a statement or a question based on the first few words. In the **unstructured task** condition, no information of this sort was provided, entailing that there was no way to infer the speech act before hearing the final word of a critical utterance. **RESULTS**: In the structured task condition, participants could (in real time) correctly infer whether unfolding utterances were statements or questions, as reflected in early target identification. In contrast, in the unstructured task condition, referent identification was delayed until hearing the sentence-end disambiguating information. Together, the results: (i) provide evidence that perspective computations in real-time language processing do not rely on 'shallow' syntactic/prosodic cues that signal an utterance's speech act; (ii) show that the broader behavioral context provides a source of perspective information enabling listeners to infer the speech act of an unfolding utterance (and identify relevant referential candidates accordingly); and (iii) help ensure that cognitive models of linguistic perspective-taking generalize beyond a limited set of languages.

Example critical trials. (Note: confederate player views display from the other side.)

Fěnsè gézi lǐ de | gǒu chuān de shì | shénme xié?
Gloss: Pink grid inside MOD | dog wear PTC COP | what footwear?
Translation: What footwear is the dog in the pink square wearing?



Fěnsè gézi lǐ de | gǒu chuān de shì | qúnzi.
Gloss: Pink grid inside MOD | dog wear PTC COP | skirt.
Translation: The dog in the pink square is wearing a skirt.

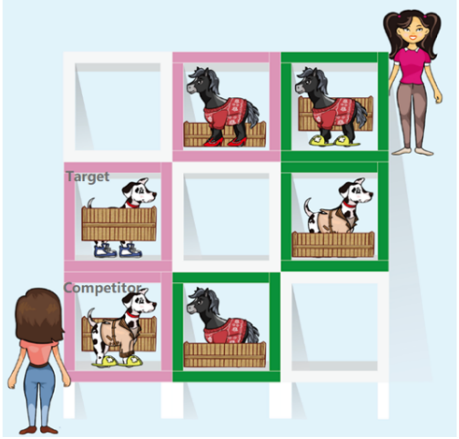
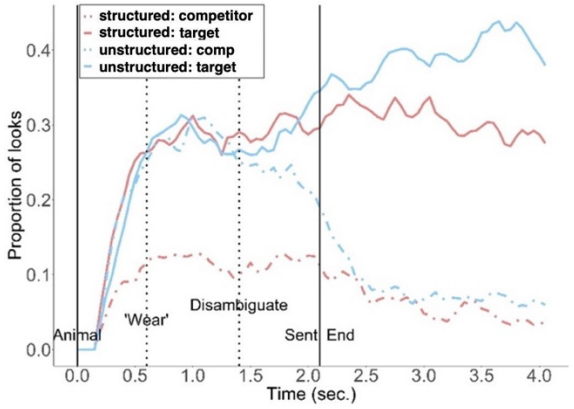


Fig. 1: Critical displays (as viewed by the participant player) where the participant could, in the **structured task** condition, correctly infer the utterance they're hearing is a question (left) or a statement (right), as measured by target anticipation. LEFT: Knowing that the (confederate) partner's focus is on footwear could enable the inference (upon hearing "*gǒu*") that the utterance is a question relating to the (target) dog whose footwear is hidden from the partner (i.e., the partner has no motivation to ask or state anything relating to the other dog in a pink square). RIGHT: The participant believes the partner knows the participant's concern is clothing and thus the participant could infer the unfolding utterance is a statement providing information about the (target) dog in a pink square whose clothing is hidden from the participant, again enabling target identification from "*gǒu*". In the **unstructured task** condition, target identification is not possible for either case above until the sentence-final word is heard. STUDY DETAILS: 48 native Mandarin participants; 48 trials: 24 critical (12 questions & 12 statements) and 24 fillers. Critical utterances were pre-validated to ensure prosodic cues did not disambiguate the speech act.



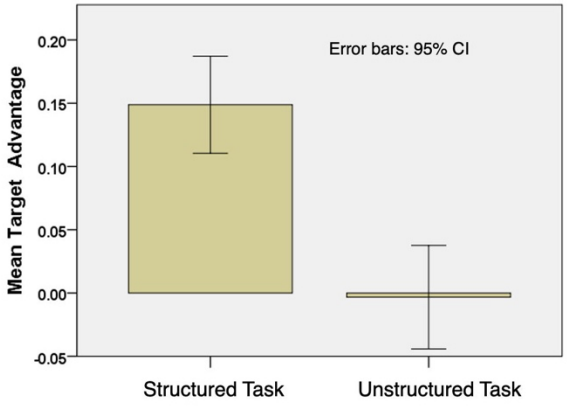


Fig. 2. LEFT: fixations over time (from onset of animal name) show early differentiation of target from competitor in structured but not unstructured task condition. RIGHT: target advantage score (mean target minus mean competitor fixation) across conditions before sentence-final disambiguating information is encountered (mixed effect model, $\beta = 1.050$, $z = 4.312$, $p < .001$).

References: [1] Heller & Brown-Schmidt (2023) *Cognitive Science.* [2] Brown-Schmidt et al. (2008) *Cognition.* [3] Shyu & Tung (2018) *Studies in Prosodic Grammar.*