Effects of instruction on semantic and pragmatic judgment tasks

Ziling Zhu & Dorothy Ahn, Rutgers University

Background. Experimental linguistic work is defined by its design, procedures, and statistical analysis (Kirk, 2012; Myers, 2017). There have recently been more discussions on how to optimize procedures for sentence judgment tasks, featuring two considerations: instruction (Schutze, 2005; a.o.) and response scale (Schutze & Sprouse, 2013; a.o.). Instruction variation was claimed to be trivial for morphological (Aronoff & Schvaneveldt, 1978), syntactic (Schutze & Sprouse, 2013), and pragmatic (Veenstra & Katsos, 2018) judgment tasks. This study fills the research gap for experimental semantics and pragmatics, revealing that instruction is a significant factor in identifying and distinguishing between semantic and pragmatic violations in sentence judgment tasks. Furthermore, we show that English and Mandarin speakers respond differently to different keywords in the instructions, highlighting the need for language and study-specific norming procedures.

Methods. To investigate the effects of instruction in sentence judgment tasks, we compared participants' responses to four commonly used instructions shown in (1) against the same set of sentence stimuli. A total of 24 syntactically well-formed sentences were tested in the stimuli, and we grouped them into three categories based on their semantic and pragmatic felicitousness: (i) 8 `semantically odd` (logical contradiction and thematic mismatch), (ii) 8 `pragmatically odd` (redundant information), and (iii) 8 `neutral`. An example of each sentence type is shown in (2).

(1) a. Does this sound natural to you?
b. Does this sound acceptable to you?
c. Does this sound grammatical to you?
d. How likely is it for a native speaker to say this?

(2) a. Jake is a married bachelor. (`semantically odd`)
b. Yuki arrived. Yuki sat down. Yuki turned on her laptop. (`pragmatically odd`)
c. Mason thinks it's raining outside. (`neutral`)

In order to test for language-specific effects, we also created a Mandarin version of the English study with the instructions as in (3).

(3) a. yixia neirong ting-qilai ziran ma?
following contents hear-impression natural Q-PART?
'Do the following contents sound natural?'
b. yixia neirong ting-qilai fuhe yufa ma?
following contents hear-impression fit grammar Q-PART?
'Do the following contents sound grammatical?'
c. yixia neirong ting-qilai ke jieshou ma?
following contents hear-impression can accept Q-PART?
'Do the following contents sound acceptable?'
d. nin renwei muyu wei hanyu de ren, you duo-da keneng
you think native.language be Mandarin GEN person, have how-big possibility
shuo-chu yixia neirong?
say-out following contents?
'How likely do you think is it for a native speaker of Mandarin to say the following contents?'

We used a between-subject study so that each participant would only see one question type for all 24 test items. Participants were asked to respond on a 7-point Likert scale.

Eighty-one native English speakers and 81 native Mandarin speakers (18-64; gender-balanced) were recruited via Prolific. They were asked to provide some demographic and language background information, and then were presented with the 24 sentence stimuli (randomized in order).

Predictions. If instruction variation is trivial for semantic and pragmatic judgment tasks, we would

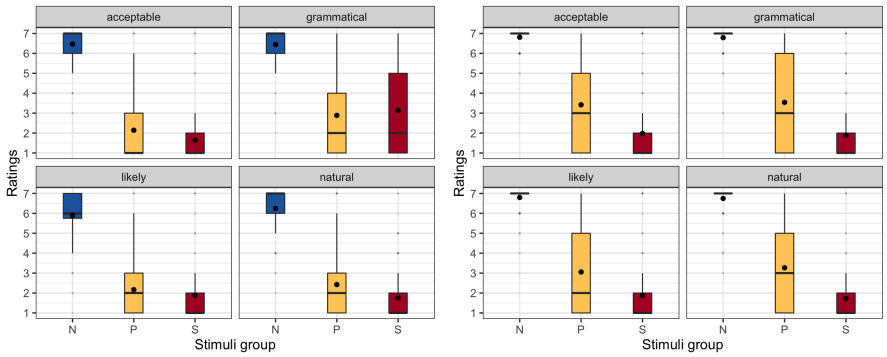

Figure 1:
Ratings as function of stimuli group, grouped by instruction type
N: `neutral`
P: `pragmatically odd`
S: `semantically odd`
English(L), Mandarin(R)

predict that instruction type would not change the rating results for each test sentence. Otherwise, different instructions would lead to different ratings of the same stimuli.

Results. We fit a Cumulative Link Mixed Model in R to compare ratings in different conditions (Fig. 1). For English, the results showed a main effect of stimuli group ($p < 0.001$), instruction type ($p < 0.001$), and significant interaction ($p < 0.001$). For Mandarin, we only found a main effect of stimuli group ($p < 0.001$), and not instruction type ($p > 0.1$), and no significant interaction ($p > 0.1$).

Across the stimuli groups, all instruction types reliably distinguished between `odd` and `neutral` stimuli ($p < 0.001$) for both English and Mandarin. Between `semantically` and `pragmatically odd` sentences, for English, all instruction types led to significantly different responses except for `grammatical` ($p > 0.1$); for Mandarin, all instruction types led to significantly different responses ($p < 0.001$). Moreover, the instruction type `natural` was the most effective in teasing apart the stimuli groups for both languages.

Discussion: Our experiment reveals the significance of instruction type in semantic and pragmatic sentence judgment tasks. First, we confirm the intuitive choice, made by previous researchers, of using 'natural' in the instruction design (Cremers & Chemla, 2017; Zlogar & Davidson, 2018; Hara et al., 2014; a.o.). Second, we highlight the need to include control sentences with standard ratings to evaluate semantic and pragmatic violations more accurately. For instance, Sprouse et al. (2020) use a set of previously-tested sentences as fillers to calibrate newly collected grammaticality judgments in their syntax study. Our preliminary data can serve a similar role in semantic and pragmatic judgment tasks.

The current study also draws attention to cross-linguistic differences in sentence judgment tasks. While `natural` is the best keyword to distinguish between semantic and pragmatic oddness for both languages, the range of responses spreads wider in Mandarin than in English. Hence, language-specific norming studies with control sentences are crucial in order to effectively compare cross-linguistic judgments.

More generally, our study speaks to the general concern on the validity of sentence judgment tasks used for semantic and pragmatic research. The grouping of the stimuli into pragmatically odd, semantically odd, and neutral sentences is not independently motivated and thus potentially theory-internal. However, our results suggest that the paradigm of sentence judgment tasks can identify at least some distinction between logically illicit sentences (`semantically odd`) and sentences that are logical but not discourse-natural (`pragmatically odd`).

Aronoff, M., & Schvaneveldt, R. 1978. Testing morphological productivity. *Annals of the New York Academy of Sciences*, 318(1). **Cremers, A., & Chemla, E. 2017.** Experiments on the acceptability and possible readings of questions embedded under emotive-factives. *Natural Language Semantics*, 25(3). **Hara, Y., Kawahara, S., & Feng, Y. 2014.** The prosody of enhanced bias in Mandarin and Japanese negative questions. *Lingua*, 150. **Kirk, R. 2012.** *Experimental design: Procedures for the behavioral sciences.* **Myers, J. 2017.** Acceptability judgments. *Oxford Research Encyclopedia of Linguistics.* **Schütze, C. T. 2008.** Thinking about what we are asking speakers to do. *Linguistic Evidence.* **Schütze, C. T., & Sprouse, J. 2013.** Judgment data. *Research methods in linguistics.* **Sprouse, J., Messick, T., & Bobaljik, J. 2020.** Gender asymmetries in ellipsis: An experimental comparison of markedness and frequency accounts in English. *Journal of Linguistics.* **Veenstra, A., & Katsos, N. 2018.** Assessing the comprehension of pragmatic language: Sentence judgment tasks. *Methods in Pragmatics.* **Zlogar, C., & Davidson, K. 2018.** Effects of linguistic context on the acceptability of co-speech gestures. *Glossa.*