Effects of instruction on semantic and pragmatic judgment tasks

Ziling Zhu & Dorothy Ahn*

Abstract. Sentence judgment tasks are used often in linguistics studies. However, there is no consensus on how significant the effect of instruction is in such tasks: some argue that instruction is trivial, while others argue that it affects the way participants respond. In this study, we investigate different keywords used in sentence judgment tasks and determine which keyword best teases apart speakers' response to semantically and pragmatically licit and illicit sentences. We test this in English and Mandarin, exploring the possibility of cross-linguistic variation on how speakers respond to different keywords. Our results show that the common keywords used in semantic and pragmatic judgment tasks such as 'natural' do distinguish semantic and pragmatic violations for English speakers, but that the common Mandarin translations of these words fail to distinguish between the two types of violations. Our results highlight the need for language- and study-specific norming procedures in sentence judgment tasks.

Keywords. sentence judgment tasks; instruction type; semantics; pragmatics; psycholinguistics; cross-linguistic variation

1. Introduction. 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: the **instruction** (Schütze 2005; a.o.) and the **response scale** (Schütze & Sprouse 2013; a.o.). Instruction conveys what the researchers ask the participants to do, and response scale determines how the participants communicates back to the researchers. In this project, we focus on the role of instruction in a commonly used experimental paradigm in linguistics: sentence judgment tasks.

Empirical evidence diverges as to whether instruction variation matters in sentence judgment tasks. On the one hand, instruction has been claimed to be trivial in morphological (Aronoff & Schvaneveldt 1978) and syntactic (Cowart 1997) judgment tasks. On the other hand, the effect of instruction was reported in some other syntactic studies, where different keywords lead to significantly different judgments from the participants (Maclay & Sleator 1960, Beltrama & Xiang 2016). Methodological studies on pragmatic judgment tasks have been conducted in Veenstra & Katsos (2018), but the role of instruction variation has received scarce attention.

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 keywords in the instructions, highlighting the need for language and study-specific norming procedures.

This paper is organized as follows: Section 2 introduces the background, discussing the contrasting claims on whether instruction type is a significant factor or not in sentence judgment tasks.

*We would like to thank audience at ELM2 and Rutgers 2022 Open House for their helpful discussions. Funding was generously provided by Meaning Across Languages (MAL) Lab of Rutgers University. Special thanks go to Chaoyi Chen and Joseph Casillas for techinical support on statistics. Authors: Ziling Zhu, Rutgers University (ziling.zhu@rutgers.edu) & Dorothy Ahn, Rutgers University (dorothy.ahn@rutgers.edu).

In Section 3, we report on the results of the study we ran, which suggest that a) instruction type is significant for identifying semantic and pragmatic violations and b) English and Mandarin speakers respond differently to keywords that are often used as translations of each other in empirical studies. Section 4 concludes with a discussion of implications and remaining questions.

2. Background. Instruction variation, as a procedural factor of linguistic experiments, was first investigated by Hill (1961). In Hill's study, ten subjects were instructed to "reject any sentences which were ungrammatical, and to accept those which were grammatical", with no definition of (un)grammaticality. The results showed that participants lacked a consistent understanding of this concept. However, since Cowart's (1997) experiment, to be reviewed next, it has been assumed in experimental linguistics that "the exact nature of instructions matters relatively little" (Schütze & Sprouse 2013), although systematic empirical investigations are not yet in place. In answering this methodological question, as to whether instruction variation is significant in linguistic experiments, two claims have been put forward in the literature.

2.1. INSTRUCTION VARIATION IS TRIVIAL. Instruction variation was claimed to be trivial for morphological (Aronoff & Schvaneveldt 1978) and syntactic (Cowart 1997, Schütze & Sprouse 2013) judgment tasks. We review the following studies that support this view.

To test the morphological productivity of Word Formation Rule with affixes, Aronoff & Schvaneveldt (1978) designed some possible but non-occurring English words and asked participants to evaluate them. Crucially, they manipulated the instruction as in (1).

(1) a. Is the item in your vocabulary?
b. Is the item an English word?
c. Is the item a meaningful word?

Each group of participants only saw one question. The percentage of affirmative responses were calculated and reported. The researchers reported that although the percentages vary as the instructions vary (42% for (1-a), 50% for (1-b), 54% for (1-c)), the instruction variation has no significant effect on how morphology influences participants' responses.

In the field of syntax, Cowart (1997) claims that the nature of instructions matters little. Two types of instructions, termed as 'intuitive' and 'prescriptive' by Cowart, were used in this study. The intuitive instructions in (2-a) highlight the participants' perception of these sentences, namely whether the stimuli are fully normal/very odd, or somewhere between the extremes. In contrast, the prescriptive instructions in (2-b) invoke a prescriptive view, pointing the participants to school grammar and language authorities.

(2) a. **Intuitive Instructions:** Please read each of the sentences listed below. For each sentence, we would like you to indicate your reaction to the sentence. Mark your response sheet A, B, C, or D. Use (A) for sentences that seem fully normal, and understandable to you. Use (D) for sentences that seem very odd, awkward, or difficult for you to understand. (Note: DO NOT USE "E".) If your feelings about the sentence are somewhere between these extremes, use one of the middle responses, B or C. THERE ARE NO "RIGHT" OR "WRONG" ANSWERS. Please base your responses solely on your gut reaction, not on rules you may have learned about what is "proper" or "correct" English.
b. **Prescriptive Instructions:** Please read each of the sentences listed below. For each

> sentence, we would like you to indicate whether or not you think the sentence is a well-formed, grammatical sentence of English. Suppose this sentence were included in a term paper submitted for a 400-level English course that is taken only by English majors; would you expect the professor to accept this sentence? Mark your response sheet A, B, C, or D. Use (A) for sentences that seem completely grammatical and well-formed. Use (D) for sentences that you are sure would not be regarded as grammatical English by any appropriately trained person. (Note: DO NOT USE "E".) If your judgment about the sentence is somewhere between these extremes, use one of the middle responses, B or C. Use B for sentences you think probably would be accepted but you are not completely sure. Use C for sentences you think probably would not be accepted.

Participants were asked to evaluate sentences with local/remote antecedents in different syntactic structures. Cowart reports no linguistically meaningful influences of the two types of instructions.

2.2. INSTRUCTION VARIATION IS SIGNIFICANT. In contrast to the previous view, task effect has been observed in other linguistic judgment tasks (Maclay & Sleator 1960, Beltrama & Xiang 2016), suggesting that instruction variation could be a significant factor in sentence judgment tasks.

To explore the nature of sentence judgment tasks in linguistics, Maclay & Sleator (1960) devised six groups of sentence stimuli differing in whether they are 'grammatical', 'meaningful', and 'ordinary'. The instructions vary accordingly, as in (3).

(3) a. Do these words form a grammatical English sentence?
b. Do these words form a meaningful English sentence?
c. Do these words form an ordinary English sentence?

Participants were asked to answer YES or NO, and the proportion of affirmative responses were calculated. The researchers observed that judgments of syntactic well-formedness and those of semantic meaningfulness are independent from each other. For stimuli with semantic violations (grammatical but not meaningful), 42% of the participants judged them to be grammatical, while only 7% judged them to be meaningful. In contrast, for stimuli with syntactic violations (ungrammatical but meaningful), 34% of the participants judged them to be grammatical, while 52% judged them to be meaningful. Participants systematically rejected syntactic violations more with the 'grammatical' instruction in (3-a), and semantic violations more with the 'meaningful' instruction in (3-b). In other words, the instructions 'grammatical' and 'meaningful' could systematically tease apart participants' response to syntactic and semantic violations, respectively.

Similar observations have been made in a syntactic study by Beltrama & Xiang (2016).[1] To examine whether intrusive resumptive pronouns can rescue island violations, they designed an acceptability task and a comprehensibility task, with instructions in (4-a) and (4-b) respectively.

(4) a. How acceptable is the [target sentence]? Please make your judgments based on how good the [target] sentence sounds in English given the context it is in.
b. We want you to judge these sentences based on how easy they are for you to understand.

Whereas resumptive pronouns do not improve island violations in the acceptability task, such rescuing effect was found in the comprehensibility task. This task effect crucially arises from the

[1] We thank Troy Messick for pointing us to this study.

instruction variation in (4), suggesting that instruction is in fact significant in guiding participants' responses in syntactic sentence judgment tasks.

We fill two research gaps in this project. First, we explore whether instruction is a significant factor in semantic and pragmatic sentence judgment tasks. As reviewed above, most linguistic studies that examine instruction variation in sentence judgment tasks have focused on morphology and syntax. Second, we extend our research question to make a cross-linguistic comparison. Specifically, we ask whether English and Mandarin speakers respond differently to keywords in the instructions that are often assumed to be comparable to each other in empirical studies (Hara et al. 2014, Xue et al. 2020, Law & Syrett 2017).

3. Experiment. To investigate the effects of instruction in semantic and pragmatic sentence judgment tasks, we compared participants' responses to different instructions against the same set of sentence stimuli.

3.1. STIMULI. We chose four commonly used instructions in English sentence judgment tasks, shown in (5), varying in the key adjectives.

(5) 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?

In order to test for language-specific effects, we also created a Mandarin version of the English instructions as in (6), using words commonly used as translations of those found in (5). For example, 'ziran (natural)' was used in Hara et al. (2014), 'ke jieshou (acceptable)' in Xue et al. (2020) and Law & Syrett (2017), among many others.

(6) 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?'

A total of 24 syntactically well-formed sentences were tested as the stimuli. We grouped them into three categories based on their semantic and pragmatic felicitousness. The first group contains 8 **semantically odd** stimuli involving lexical contradictions (7-a)–(7-c), logical contradictions

(7-d),(7-e), and thematic mismatch (7-f)–(7-h). Our categorization of these as semantic violations is based on a few assumptions of what falls under semantic knowledge. First, we assume that world knowledge based on lexical meaning (e.g. a bachelor is unmarried) constitutes semantic knowledge of a word. This kind of lexical contradiction is what is often called 'semantic violations' in EEG studies, for example. Second, we assume that logical relations between propositions are part of the semantic knowledge that a speaker has. Note that world-knowledge and logical relations are quite different from each other. For now, we group these together under 'semantic violations', but it would be good to test whether the two kinds of violations elicit different responses.

(7) a. Jake is a married bachelor.
b. Jasmine talked silently.
c. Bantee's yellow hat is blue.
d. It is raining and not raining outside.
e. I'm lying when I say this sentence is true.
f. Zhangsan smelled 3 o'clock.
g. I bought a lamp, and the lamp is drinking water.
h. I washed brightness.

The second group consisted of 8 **pragmatically odd** stimuli with redundant information, including direct repetition of information (8-a)–(8-d) and repetition of scalar implicature (8-e)–(8-h). Our categorization of pragmatic violations is based on two factors. First, the sentences do not meet our definition of semantic violations: they are neither logically nor lexically contradictory. Second, the sentences violate certain pragmatically-motivated constraints such as the Gricean Maxims. Redundant information, for example, is a violation of the Quantity Maxim, while repetition of scalar implicature is a violation of the Manner Maxim.

(8) a. Yuki arrived. Yuki sat down. Yuki turned on her laptop.
b. Mimi jumped onto the bed. Mimi cried. Mimi decided to sleep.
c. It rained yesterday when it rained yesterday.
d. Carolyn went to the park when she went to the park.
e. Not only are all the students sad, some of them are sad.
f. All the students passed the exam and some of them passed.
g. Three engineers came to work today and two engineers came to work today.
h. Becky and Vera went to the party and Becky went to the party.

The third group contained 8 **neutral** stimuli with no identifiable semantic or pragmatic violations as defined above. The sentences presented under this categorization are shown in (9).

(9) a. Yuki decided to go to school today.
b. Mason thinks it's raining outside.
c. Anya was drinking water, because she was thirsty.
d. I bought a beautiful hat three days ago. The hat was yellow.
e. The students turned off their laptops and went outisde.
f. Yesterday it was sunny in Toronto.
g. A bachelor walked up to us and introduced himself.

h. I have a daughter and a son. They are nice to each other.

3.2. PARTICIPANTS AND PROCEDURE. We recruited 81 native English speakers and 81 native Mandarin speakers (18–64; gender-balanced) via Prolific. Participants were redirected to a Qualtrics survey, where they were asked to first provide some demographic and language background information and then complete the sentence judgment task. Participants were compensated $2-3 for their time.

The study was designed to be between-subject, so that each participant would only see one instruction type for all 24 test items. Participants were presented with the sentence stimuli (randomized in order) one at a time and were asked to respond on a 7-point Likert scale based on the instruction they saw, as in Fig. 1.

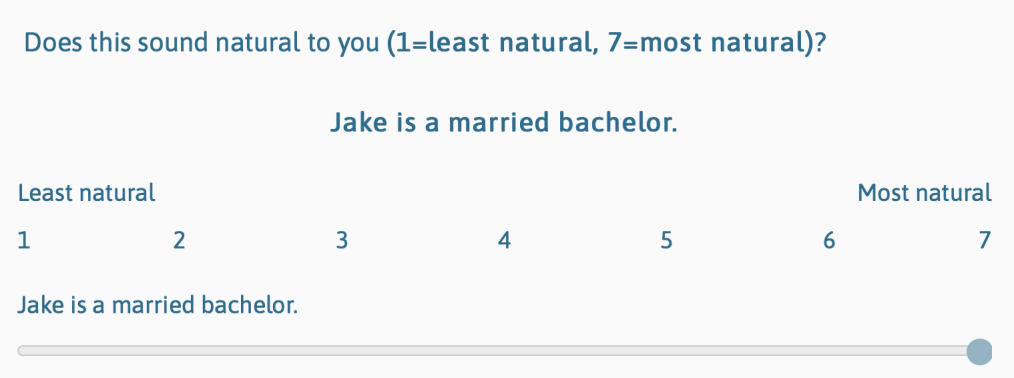

Figure 1: Sample question ('natural' condition with a lexical contradiction)

We collected the ratings for each sentence as the dependent variable and tested for each language whether STIMULI GROUP (NEUTRAL, SEMANTICALLY ODD, PRAGMATICALLY ODD) and/or INSTRUCTION TYPE (NATURAL, ACCEPTABLE, GRAMMATICAL, LIKELY) lead to significant rating differences.

3.3. PREDICTIONS. If instruction variation is trivial for semantic and pragmatic judgment tasks, we would predict that instruction type would not change the rating results for each test sentence. Instead, participants would rate the sentences based on their respective standards. If instruction variation is significant, however, different instructions would lead to different ratings of the same stimuli. If certain keywords are more likely to prime judgments based on certain violations, we would also expect the contrasts between conditions to be consistent across speakers. Finally, if the (in)significance of instruction variation has no cross-linguistic difference, then native English speakers and native Mandarin speakers would show similar contrasts across different instruction types.

3.4. RESULTS. We fit a Cumulative Link Mixed Model in R to compare ratings in different conditions (Fig. 2). We first observed a cross-linguistic variation for INSTRUCTION TYPE: for English, the results showed a main effect of STIMULI GROUP ($p < 0.001$), INSTRUCTION TYPE ($p < 0.001$), and a significant interaction ($p < 0.001$); for Mandarin, we only found a main effect of STIMULI GROUP ($p < 0.001$), but not INSTRUCTION TYPE ($p > 0.1$), and no significant interaction ($p > 0.1$). This suggests that, while English speakers use different instructions to tease apart different

linguistic violations, Mandarin speakers do not make this distinction among the instructions.

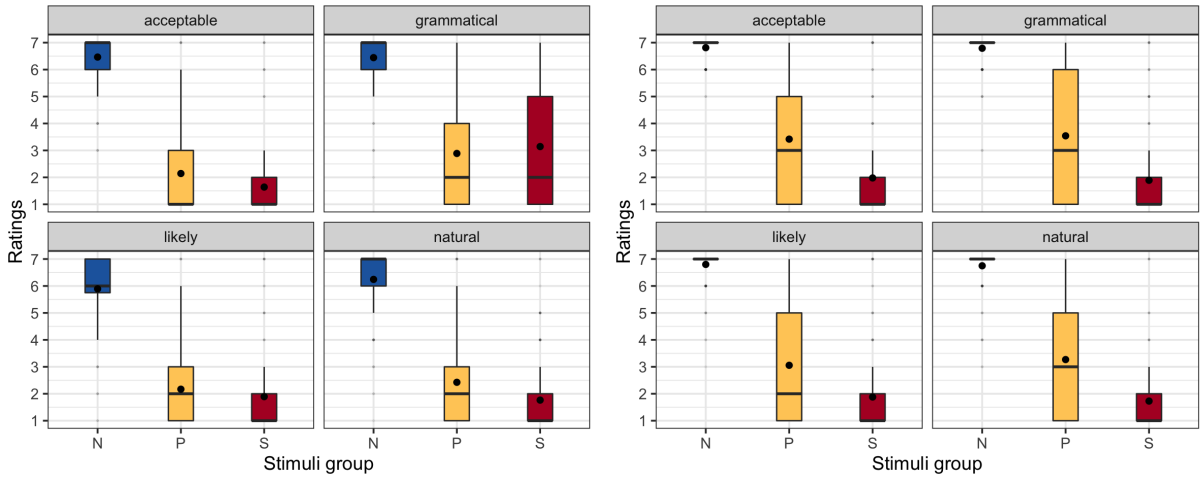

Figure 2: Ratings as function of STIMULI GROUP, grouped by INSTRUCTION TYPE (N: neutral; P: pragmatically odd; S: semantically odd)

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 English and Mandarin.

4. Discussion and conclusions. 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 (Cremers & Chemla 2017, Zlogar & Davidson 2018, Hara et al. 2014; a.o.), which seems to distinguish between semantically and pragmatically odd sentences most clearly. Second, we highlight the need to include control sentences with standard ratings to evaluate semantic and pragmatic violations more accurately. Sprouse et al. (2022) 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. It is important to note that by 'teasing apart semantic and pragmatic violations', we do not mean that there is a clearly delineated diagnostic that determines whether something is a semantic violation or a pragmatic violation. What the stimuli offer instead is a way to compare the rating of a given sentence against the ratings associated with what we know to be logically or pragmatically illicit and indirectly determine the rationale for the participants' response. For example, the stimuli used in the current study has been used as controls to an independent study looking at participants' rating of Mandarin bridging anaphora (Zhu & Ahn in prep). Comparing participants' rating of the target stimuli against the controls allowed us to determine whether participants' rating of bridging anaphora align better with semantically odd sentences or pragmatically odd sentences.

The current study also draws attention to cross-linguistic differences in sentence judgment tasks. First, we see that the range of ratings spreads wider in Mandarin than in English in general. The central tendency bias, where participants avoid the endpoints of a scale, is considered to be one of the most robust biases found in psychology (Stevens 1971). Our data from Mandarin participants, where neutral sentences are rated at the ceiling, suggest that there might be cross-linguistic variation on this tendency as well. This raises a question on the nature of this tendency, i.e. whether this is a general cognitive bias that applies across languages, or something that develops culture-specifically. Another cross-linguistic difference we observe is that instruction type makes a difference in the way participants respond to different stimuli in English, but not in Mandarin. Hence, language-specific norming studies with control sentences seem crucial in order to effectively compare cross-linguistic judgments.

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). In the future, a clustering study could help identify and distinguish between subtypes of semantic and pragmatic violations.

References

Aronoff, Mark & Roger Schvaneveldt. 1978. Testing morphological productivity. *Annals of the New York Academy of Sciences* 318(1). 106–114. https://doi.org/10.1111/j.1749-6632.1978.tb16357.x.

Beltrama, Andrea & Ming Xiang. 2016. Unacceptable but comprehensible: the facilitation effect of resumptive pronouns. *Glossa: a journal of general linguistics* 1(1). https://doi.org/10.5334/gjgl.24.

Cowart, Wayne. 1997. *Experimental syntax*. Sage.

Cremers, Alexandre & Emmanuel Chemla. 2017. Experiments on the acceptability and possible readings of questions embedded under emotive-factives. *Natural Language Semantics* 25(3). 223–261. https://doi.org/10.1007/s11050-017-9135-x.

Hara, Yurie, Shigeto Kawahara & Yuli Feng. 2014. The prosody of enhanced bias in Mandarin and Japanese negative questions. *Lingua* 150. 92–116. https://doi.org/10.1016/j.lingua.2014.07.006.

Hill, Archibald A. 1961. Grammaticality. *Word* 17(1). 1–10. https://doi.org/10.1080/00437956.1961.11659742.

Kirk, Roger. 2012. *Experimental design: Procedures for the behavioral sciences*. Sage Publications.

Law, Jess HK & Kristen Syrett. 2017. Experimental evidence for the discourse potential of bare nouns in mandarin. In *Nels 47: Proceedings of the forty-seventh annual meeting of the north east linguistic society*, vol. 2, 231–40.

Maclay, Howard & Mary D Sleator. 1960. Responses to language: Judgments of grammaticalness. *International Journal of American Linguistics* 26(4). 275–282.

Myers, James. 2017. Acceptability judgments. In *Oxford research encyclopedia of linguistics*,

https://doi.org/10.1093/acrefore/9780199384655.013.333.

Schütze, Carson T. 2005. Thinking about what we are asking speakers to do. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 457–485. Mouton de Gruyter Berlin. https://doi.org/10.1515/9783110197549.

Schütze, Carson T & Jon Sprouse. 2013. Judgment data. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50.

Sprouse, Jon, Troy Messick & Jonathan David Bobaljik. 2022. Gender asymmetries in ellipsis: An experimental comparison of markedness and frequency accounts in English. *Journal of Linguistics* 58(2). 345–379. https://doi.org/10.1017/S0022226721000323.

Stevens, Stanley S. 1971. Issues in psychophysical measurement. *Psychological review* 78(5). 426.

Veenstra, Alma & Napoleon Katsos. 2018. Assessing the comprehension of pragmatic language: Sentence judgment tasks. In Andreas Jucker, Klaus Schneider & Wolfram Bublitz (eds.), *Methods in pragmatics*, vol. 10, 1806. Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9783110424928.

Xue, Wenting, Meichun Liu & Stephen Politzer-Ahles. 2020. A study of complement coercion in Mandarin Chinese: evidence from an acceptability judgment task. In *Workshop on chinese lexical semantics*, 775–784. Springer. https://doi.org/10.1007/978-3-030-81197-6_64.

Zlogar, Christina & Kathryn Davidson. 2018. Effects of linguistic context on the acceptability of co-speech gestures. *Glossa: a journal of general linguistics* 3(1). https://doi.org/10.5334/gjgl.438.