Article

Not peer-reviewed version

The Acceptability and Validity of AI-Generated Psycholinguistic Stimuli

Alaa Alzahrani *

Posted Date: 22 April 2024

doi: 10.20944/preprints202404.1074.v2

Keywords: AI-generated stimuli; psycholinguistic stimuli; ChatGPT; AI for research; auditory stimuli; sentence stimuli

Article

The Acceptability and Validity of AI-Generated Psycholinguistic Stimuli

Alaa Alzahrani

alzahrani.alaaa@gmail.com

Abstract: Sentence stimuli pervade psycholinguistics research. Yet limited attention has been paid to the automatic construction of sentence stimuli. Given their linguistic capabilities, this study investigated the efficacy of ChatGPT/AI tools in generating sentence stimuli. In three psycholinguistic experiments, this study examined the acceptability and validity of AI-formulated auditory sentences and written sentences in two languages: English and Arabic. Participants gave English AI-generated stimuli similar to or higher acceptability ratings than human-composed stimuli, but the opposite was observed for Arabic AI-generated stimuli. The validity of AI-developed stimuli relied on the study design, with only Experiment 2 demonstrating the target psycholinguistic effect (L1/L2 morphosyntactic prediction). These results highlight the promising role of AI as a stimuli developer which could facilitate psycholinguistic research and increase its diversity. Implications for psycholinguistic research were discussed.

Keywords: AI-generated stimuli; psycholinguistic stimuli; ChatGPT; AI for research; auditory stimuli; sentence stimuli

Introduction

Harnessing generative AI for research purposes has garnered widespread attention since the introduction of ChatGPT at the end of 2022 (Dwivedi et al., 2023; Kuteeva & Andersson, 2024; Li et al., 2023). In second language and bilingual research, most of the applications of AI have been focused on data analysis (Anthony, 2023; Curry et al., 2024; Lin, 2023; Uchida, 2024; Zappavigna, 2023) and generating norming data (Heyman & Heyman, 2023; Trott, 2024). A far less examined application of AI is in stimuli development (e.g., Bae, 2024). In the psycholinguistic literature, several methods have been proposed for the generation of word stimuli (e.g., C. Gao et al., 2023; Taylor et al., 2020). However, the automatic formulation of sentence stimuli remains underexplored despite the wide adoption of sentence items in the field (e.g., Bae, 2024). This study examined the acceptability and validity of two types of AI-generated sentence stimuli (auditory, written) across three psycholinguistic experiments conducted in two languages (English, Arabic). Results would enhance our understanding of the current capabilities of AI in facilitating psycholinguistic research.

Literature Review

Types of Psycholinguistic Stimuli

Two main linguistic stimuli formats are used in psycholinguistic research: auditory and visual. The auditory format might present the auditory form of a word (e.g., Cheng et al., 2014), a sentence (e.g., Koch et al., 2023), or a short text (e.g., Rodd et al., 2016) carrying the target feature. Visual stimuli may consist of written language (a word, sentence(s), a text) (e.g., Contemori, 2021) or pictures representing a target object/event (e.g., Garrido Rodriguez et al., 2023). All types of stimuli are controlled in some way to allow a more robust investigation of the effect of interest. Crucially, auditory sentences are typically manually modified using specialized software (e.g., Praat, Audacity) to adjust the presence and duration of pauses (e.g., Bosch et al., 2022; Ito et al., 2023). Additionally, the voice actor may regulate the speech rate by articulating a specific number of syllables per second (e.g., Koch et al., 2023). The current study examined AI-generated auditory and written sentence stimuli due to their frequent use across different psycholinguistic designs (Jegerski & VanPatten, 2013).

The Challenge of Designing Psycholinguistic Stimuli

Any researcher who designed a psycholinguistic study recognizes that developing appropriate stimuli requires extensive effort and considerable money. For instance, auditory stimuli in some cases need to be recorded by a professional voice actor, which can incur substantial fees even for a relatively small number of items. The high cost of stimuli design can be particularly detrimental in non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) contexts (Henrich et al., 2010). Limited research funding in these regions (Petersen, 2021) might discourage early-career researchers and graduate students from conducting psycholinguistic research, exacerbating the existing bias toward WEIRD samples in the field (Blasi et al., 2022; Plonsky, 2023). So far, two methods have been utilized in the literature to facilitate the construction of psycholinguistic stimuli.

The first method is the use of large-scale psycholinguistic databases, which provide key psycholinguistic properties (e.g., semantic, phonological, morphological) of thousands of words (C. Gao et al., 2023). Such databases have been developed for several languages such as English (e.g., Scott et al., 2019), German (e.g., Võ et al., 2009), Portuguese (e.g., Soares et al., 2017), Italian (e.g., Barca et al., 2002), Chinese (e.g., Chang et al., 2016), and Hindi (e.g., Verma et al., 2022). A second approach that could assist psycholinguistic researchers in constructing stimuli is automated stimuli generation tools such as the R package LexOPS (Taylor et al., 2020). This package allows users to generate word stimuli and customize factorial designs based on data from psycholinguistic databases.

One advantage of these freely available psycholinguistic databases and tools is that they could promote consistency and replicability in research through the standardization of word properties (C. Gao et al., 2023). However, one of their potential limitations is that they provide individual words, making them ideal for tasks involving single-word presentation, such as semantic priming (McDonough & Trofimovich, 2009). Meanwhile, other psycholinguistic designs, such as the Visual-World Paradigm (VWP) and Self-Paced Reading (SPR), involve the presentation of complete sentences (Huettig et al., 2011; Marsden et al., 2018). In these cases, researchers face the challenge of constructing grammatically sound and meaningful sentences around the individual database words. This additional task can be time-consuming.

AI and Psycholinguistic Stimuli Development

One promising method to reduce the costs associated with developing psycholinguistic stimuli is to use ChatGPT/AI tools. To the best of our knowledge, only one study has explored AI-augmented stimuli. In a descriptive study, Bae (2024) reported that ChatGPT-4 can construct appropriate psycholinguistics items for an SPR experiment. Although this study highlighted the valuable potential of AI in generating psycholinguistic stimuli, it did not empirically examine the perceived acceptability and validity of the items. Further, the acceptability and validity of AI-generated auditory stimuli remain underexplored despite their prevalence in psycholinguistics (Huettig et al., 2011). As mentioned above, auditory sentences in VWP experiments are required to have the same speech rate and are typically manually modified to add/remove pauses (e.g., Garrido Rodriguez et al., 2023; Ito et al., 2023; Karaca et al., 2023; Koch et al., 2023). These steps can be more effectively accomplished using AI tools.

The Acceptability of AI Linguistic Production

The term acceptability may be used in slightly different ways (e.g., Lau et al., 2017). Here, acceptability is defined as the perception of AI-generated linguistic stimuli as human-like, leading to acceptability ratings that rival or surpass those attained by human-generated stimuli (e.g., Lee & Kim, 2024). Research has shown that LLMs perform well in linguistics tasks that probe judgments of syntactic and semantic knowledge (Goldberg, 2019; Hu & Levy, 2023), and psycholinguistic

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2024 doi:10.20944/preprints202404.1074.v2

properties of words (Trott, 2024), sometimes approaching human performance (c.f., Marvin & Linzen, 2018). Beyond language classification, AI technologies can generate well-crafted abstracts that are often indistinguishable from human-written ones (C. A. Gao et al., 2023), even for experienced linguists (Casal & Kessler, 2023). One crucial limitation of prior research is its sole focus on the English language, which may not represent the acceptability of low-resource languages. Taken together, previous findings underscore the human-like linguistic capabilities of LLMs and suggest the potential acceptability of AI-designed English stimuli.

The Validity of AI Linguistic Production

A common way of assessing the validity of AI-generated content is by comparing it against comparable human data (Curry et al., 2024; Trott, 2024; Uchida, 2024). Likewise, validity in the current study refers to the extent to which AI-generated stimuli could replicate well-known effects in the psycholinguistic field. The validity of AI tools for linguistic research has been examined from two perspectives: AI as a participant in pilot studies (Heyman & Heyman, 2023; Trott, 2024) and a data analyzer (Anthony, 2023; Curry et al., 2024; Lin, 2023; Uchida, 2024; Zappavigna, 2023).

Existing research suggests that there is a strong agreement between the performance of ChatGPT in pilot norming studies and human data. For example, ChatGPT's judgments in a typicality rating task correlated significantly with Dutch and English human ratings (correlation coefficient range = .35, .64) (Heyman & Heyman, 2023). Similarly, GPT-4 judged 24 psycholinguistic properties of English words (e.g., concreteness, imageability, iconicity), showing positive correlations with human judgments (range = .47, .86) (Trott, 2024).

Meanwhile, studies that leveraged AI as a data analysis tool reported less favorable results (Anthony, 2023; Curry et al., 2024; Lin, 2023; Uchida, 2024; Zappavigna, 2023). Most of these studies found that ChatGPT/AI can generate human-like quantitative and qualitative analyses for a small subset of linguistic tasks (e.g., semantic classification, identification of lexico-grammatical patterns), but not for a wider range of tasks (e.g., genre identification, concordance analysis). Overall, prior studies have revealed that the validity of AI in linguistic research may be dependent on the investigated domain. As such, more research is needed to investigate the validity of AI in the psycholinguistic domain.

The Present Study

This study investigated the acceptability and validity of AI-generated sentence stimuli in three experiments (Table 1). The three experiments were designed to cover a wider number of sentence formats (auditory, written), languages (English, Arabic), and psycholinguistic effects (semantic prediction, morphosyntactic prediction, syntactic priming). This allowed a broader investigation of the efficacy of AI-formulated psycholinguistic stimuli, increasing the generalizability of the results.

Table 1. Summary of the three experiments.

Exp	Original/related study	Examined effect	Target language
1	Altmann and Kamide (1991)	Semantic prediction	English
2	Koch et al. (2023)	Morphosyntactic prediction	Arabic
3	Wei et al. (2023)	Syntactic priming	English

Experiment 1: Anticipating L2 Semantic Information

This experiment (Exp 1) replicated Altmann and Kamide's (1991) well-cited study to examine the effect of semantic prediction among L2 English speakers. Theoretical (Pickering & Gambi, 2018) and empirical (Carroll et al., 2024) Evidence suggests that L2 speakers tend to easily generate semantic predictions during sentence comprehension. As such, we expected significant L2 semantic prediction effects.

doi:10.20944/preprints202404.1074.v2

Methods

Procedure and Materials

AI-Speech Acceptability Task

Participants

Twenty L2 English speakers were recruited from Prolific. They came from 12 L1 backgrounds. All participants gave their informed consent before the task and were monetarily compensated for their time.

Materials

Sixteen experimental sentences were adapted from a previous VWP study (Altmann & Kamide, 1999). All sentences were recorded by an AI tool and a human native speaker. This study used the text-to-speech (TTS) AI commercial tool: Lovo AI (https://lovo.ai/) to generate AI recordings. This tool leverages a cutting-edge TTS technique and provides granular control over voice delivery. In the current study, the AI stimuli were recorded using a young male US English voice. A 3-second pause was added after the verb in each sentence to give participants more time to engage in predictive processing (e.g., Karaca et al., 2023). Additionally, a young male native speaker of American English recorded all 16 sentences. He was instructed to maintain a natural speaking pace and to pause for 3 seconds after the verb in each sentence.

This study used the Mean Opinion Scale-Expanded Version 2 (MOS-X2) (Lewis, 2018) to assess the acceptability of AI speech. This standardized questionnaire showed significant correlations with a previous MOS version ($r >= .30$), indicating good concurrent validity. MOS-X2 contains four items that evaluate the perceived intelligibility, naturalness, prosody, and social impression of AI-generated speech. Each item is rated on an 11-point scale ranging from 0 (lowest) to 10 (the highest). In the present study, MOS-X2 achieved somewhat acceptable internal reliability ($r = 0.68$, 95% CI [0.62, 0.74]), nearing the 0.70 threshold (Tavakol & Dennick, 2011).

Two counterbalanced lists were created, with each list presenting one condition of each audio. Items were also randomized within lists. Participants were randomly assigned to one of the lists.

Procedures

Participants completed the experiment remotely via Gorilla.sc (Anwyl-Irvine et al., 2020). They were instructed that they would listen to 16 short English auditory sentences and that they would rate each audio according to four criteria. In each trial, the audio file was played automatically and only once at the beginning of the trial to mimic the one-time presentation of auditory stimuli in VWP experiments (Huettig et al., 2011). Additionally, progression to the next trial was disabled until the auditory sentence finished playing to ensure thorough task completion. During and after listening to the audio, participants saw four questions stacked vertically in the center of the screen, each with an 11-point scale below. The participants used their mouse to select only one value of the scale. The questions remained on screen until participants clicked on the “Next” button to advance to a new trial. The task was self-paced. Following the acceptability task, participants completed a background questionnaire. The experiment took approximately 6-10 minutes to complete.

Statistical Analysis

MOS-X2 was scored following prior research (Herrmann, 2023; Lewis, 2018). Each rating was multiplied by 10, and the average score for each audio was computed by aggregating ratings per condition and participant and dividing them by 4. Thus, acceptability ratings could range from 0 to 100.

A mixed-effects linear model was constructed using lme4 (Bates et al., 2015) in R. The dependent variable was raw ratings; the fixed effect was speech condition (sum coded: Human = -.5, AI = .5).

The model included the maximal random structure that converged (Barr et al., 2013): a by-participant random intercept. The model did not include a by-item intercept because the scoring method required aggregation across items.

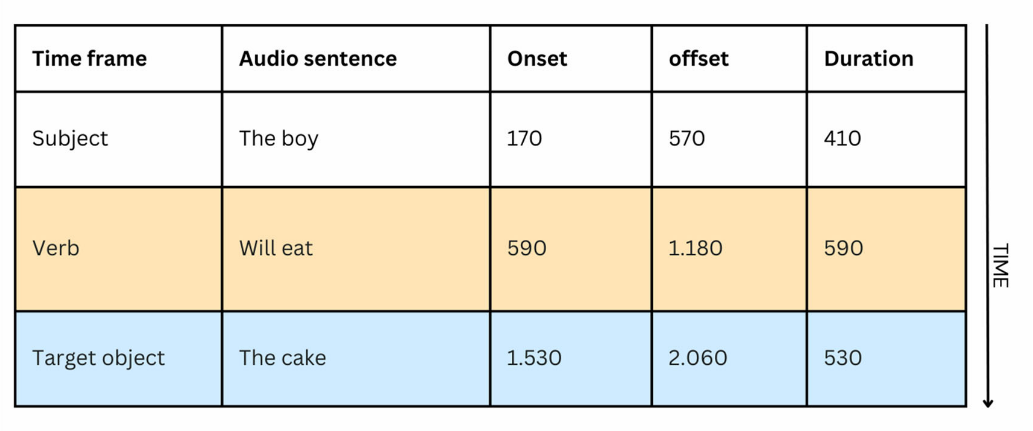
AI-Informed VWP Experiment

Participants

A new group of 30 L2 English speakers were recruited from Prolific. Data from four participants were excluded due to low sampling rates (>5) (Prystauka et al., 2023). The remaining 26 participants came from eight L1 backgrounds. All participants gave their informed consent prior to the experiment and were compensated monetarily for their participation.

Materials

Sixteen experimental sentences were adapted from a previous VWP study (Altmann & Kamide, 1999). Figure 1 illustrates one of these sentences. Eight sentences included a semantically constraining verb (e.g., drink the milk) in prediction trials, and the other eight sentences included non-semantically constraining verbs (e.g., drop the milk) in baseline trials. All sentences were accompanied by a visual display of four objects. In prediction trials, the visual display consisted of one target and three distractors. In baseline trials, the visual presentation included three competitors and one distractor. Sixteen filler sentences were adopted from a prior study (Prystauka et al., 2023). Fillers mimicked the structure of the experimental sentences, with the exception that fillers always included four or three competitor objects. Half of the filler trials were immediately followed by a yes/no comprehension question.



Time frame	Audio sentence	Onset	offset	Duration
Subject	The boy	170	570	410
Verb	Will eat	590	1.180	590
Target object	The cake	1.530	2.060	530

Figure 1. Components in auditory sentence stimuli and their duration in Exp 1.

Two counterbalanced lists were created, which presented the items in a different order. Trial order was pseudo-randomized such that adjacent trials did not display the target picture in the same region and did not repeat the trial type. Participants were randomly assigned to one of the lists.

Procedures

Participants first completed a two-minute calibration task to ensure the accuracy of the web-cam eye-tracker. Then, they completed two practice trials followed by the VWP experiment.

In the experiment, each trial began with a centrally-located fixation point, which lasted for 500 ms. Then, participants heard the auditory stimuli through their headphones while looking at the visual display. After listening to the auditory sentence, participants were asked to click on the object mentioned in the audio to proceed to the next trial. The visual display stayed on the screen until participants clicked on an object. Following object selection, a fixation point appeared in the center of the screen to indicate the start of the next trial. In filler trials, participants always saw a yes/no comprehension question after selecting an object. The question asked about the subject of the previous trial. Participants had to click on the "yes" or "no" box to advance to the next trial. Participants took an optional break after completing half of the trials. After the break, participants retook the calibration task to verify the accuracy of gaze estimation, then completed the remaining half of the experiment, followed by the background questionnaire. The experiment lasted around 10-15 minutes.

Statistical Analysis

Eye-tracking data was cleaned following procedures from an online VWP study (Prystauka et al., 2023). The final participant pool had acceptable sampling rates (> 5 Hz), with a mean of 23.89 Hz (*SD* = 4.95, range = 14.28, 30.39 Hz).

Two time windows were selected for analysis (e.g., Garrido Rodriguez et al., 2023). The first window (TW1) started 200 ms from the verb onset until 200 ms after the auditory presentation of the verb. A 200 ms was added to account for the time to initiate saccades (Saslow, 1967). Increased target fixations in prediction trials during TW1 indicate a predictive processing effect. The second window (TW2) spanned from the object onset to the end of the sentence. TW2 was included to examine whether participants in baseline trials fixated on the target object only after they heard it.

Two statistical techniques were used to capture more accurate results. Cluster-based permutation analysis (CPA) tests were generated using the R package permutes (Voeten, 2023). CPA is well-suited for analyzing eye-tracking data as it can handle autocorrelation in data points (Ito & Knoeferle, 2022). However, one limitation of CPA is that it is computationally extensive and may not be suitable for examining interactions. As such, mixed-effects logistic regression models were built using the R package lme4 to examine the interaction between condition and time. The models' structures are described in the supplementary materials.

Results

AI-Speech Acceptability Task

Descriptive statistics for the acceptability task are presented in Table 2. A linear mixed effects model indicated that L2 English speakers perceived AI speech as being of higher quality than human speech (β = 3.50, SE = 0.21, t = 16.61, p < .001, 95% CI [3.08, 3.91]), suggesting the acceptability of AI-generated auditory stimuli.

Table 2. Descriptive summary of L2 English speakers' acceptability ratings in Exp 1.

Condition	Mean (SD)	Median	Range
Human native speaker	65.81 (13.06)	66.25	37-95
AI	69.31 (15.17)	68.28	39-99

AI-Informed VWP Experiment

Comprehension Task

The percentage accuracy for the task of clicking the mentioned object was 99.66% (*SD* = .003), and 96.53% for the comprehension questions (*SD* = .011).

Statistical Modeling

CPA showed that L2 speakers significantly directed more looks to the target object picture before hearing the object during prediction trials than baseline trials (Figure 2). Logistic regression results did not support this CPA finding. The verb region (TW1) model revealed that the L2 participants did not significantly increase their looks at the target picture over time during prediction trials compared to baseline trials (β = 0.01, *SE* = 0.16, 95% CI [-0.30, 0.31]). Likewise, the L2 English speakers did not significantly increase their fixations to the target picture over time in the object region (TW2) ($\beta = 0.28$, $SE = 0.35$, 95% CI [-0.40, 0.97]).

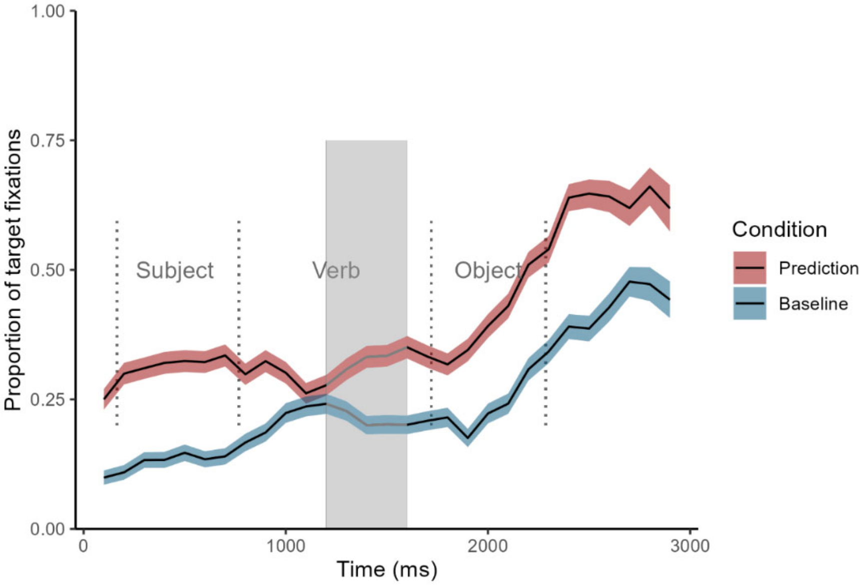


Figure 2. Time course of fixation proportions for target in the prediction (red lines) and baseline conditions (blue lines). Ribbons indicate the standard error. Dotted lines indicate the mean onset and offset of word durations in the sentences. The grey-shaded area indicates a significant CPA cluster.

Discussion

Exp 1 indicated that L2 speakers perceived the quality of AI-generated English speech to be higher than that of human speech, indicating the acceptability of AI speech. However, the AI-generated audio stimuli did not trigger an L2 semantic prediction effect, suggesting the limited validity of AI audio recordings. These findings will be elaborated on and explained in the General Discussion.

Experiment 2: Anticipating L1 and L2 Grammatical Number Information

This experiment (Exp 2) was set out to examine whether L1 and L2 Arabic speakers could predict the number of the next noun based on the verb suffix and whether there were L1-L2 differences in anticipatory processing. A related study by Koch et al. (2023) indicated that although both speaker groups can predict number information, L1 German speakers showed faster prediction effects than L2 German speakers. We expected to find the same results.

Methods

Procedure and Materials

AI-Speech Acceptability Task

Participants

Twenty-seven native Arabic participants were recruited from Prolific. All participants gave their informed consent before completing the task and were monetarily awarded for their participation.

doi:10.20944/preprints202404.1074.v2

Materials

Twelve experimental sentences were randomly selected from the full stimuli list in Exp 2 (n = 32) and used in the acceptability task. Half of the selected sentences (n = 6) were generated by a commercial AI voice generation tool (https://voicemaker.in/), and the other half by a human native speaker (n = 6). The Voice Maker AI tool was thought to offer the best natural-sounding Arabic speech synthesis at the time. The AI stimuli were recorded using a female voice in Modern Standard Arabic (MSA), and a 3-second pause was inserted after each verb. Further, a female native Arabic speaker recorded the sentences in MSA at a natural pace and without pauses due to an instruction error. Items were randomized in the task so that each participant encountered a distinct order of items.

The MOS-X2 was used to evaluate the perceived quality of AI MSA voice (Lewis, 2018). Details about this questionnaire are provided in Exp 1. The researcher translated MOS-X2 into Arabic. In the current study, MOS-X2 demonstrated good internal reliability (r = 0.87, 95% CI [0.85, 0.90]).

Procedures

The procedures in Exp 2 were identical to Exp 1.

Statistical Analysis

Exp 2 performed the same statistical analyses used in Exp 1.

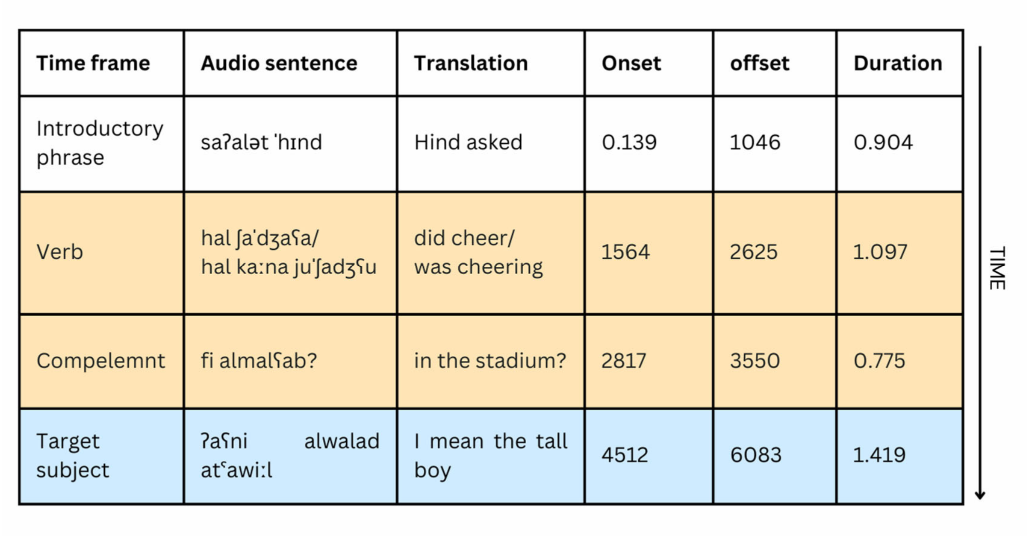
AI-Informed VWP Experiment

Participants

Twenty-one L1 Arabic speakers and 43 L2 Arabic speakers completed Exp 2. The L1 participants were recruited from Prolific and social media, and L2 speakers from King Saud University. The L2 participants came from 19 different L1 backgrounds. All participants provided their informed consent before the experiment and received monetary compensation for participation.

Materials

This study constructed two sets of experimental items, each with 16 sentences. One set used singular masculine verbs in the simple past tense (Study 1), and the other set used the same verbs in the past progressive tense (Study 2). Figure 3 includes one example of the experimental items. Each sentence was paired with a display containing four pictures. In prediction trials, the pictures included one target object and three distractors. In baseline trials, the pictures displayed two competitors and two distractors. Each set included 16 filler items. Fillers had an identical syntactic structure to the experimental sentence, with the exception that all fillers presented feminine verbs/objects to deemphasize the target predictive cue.



Time frame	Audio sentence	Translation	Onset	offset	Duration
Introductory phrase	saʔalət ˈhɪnd	Hind asked	0.139	1046	0.904
Verb	hal ʃaˈdʒaʕa/ hal kaːna juˈʃadʒʕu	did cheer/ was cheering	1564	2625	1.097
Compelemnt	fi almalʕab?	in the stadium?	2817	3550	0.775
Target subject	ʔaʕni alwalad atˤawi:l	I mean the tall boy	4512	6083	1.419

Figure 3. Components in auditory sentence stimuli and their duration in Exp 2.

Procedures

The procedures in Exp 2 were identical to Exp 1 except for one difference. In Exp 2, L1 and L2 Arabic speakers completed only one study: the simple past verb study (Study 1) or the past progressive verb study (Study 2).

Statistical Analysis

Exp 2 performed the same statistical analyses used in Exp 1. The models' structures are described in the supplementary materials. All participants had acceptable sampling rates (> 5 Hz), with a mean of 20.48 Hz (*SD* = 6.69, range = 5.32, 30.05 Hz).

Results

AI-Speech Acceptability Task

Descriptive statistics for the acceptability task are summarized in Table 3. A linear mixed effects model revealed that Arabic native speakers rated AI speech quality substantially lower than human speech (β = -12.65, SE = 0.62, t = -20.26, p < .001, 95% CI [-13.87, -11.43]), indicating the limited acceptability of AI-generated Arabic auditory stimuli.

Table 3. Descriptive summary of Native Arabic speakers' acceptability ratings in Exp 2.

Condition	Mean (SD)	Median	Range
Human native speaker	81.26 (15.13)	81.25	25-96
AI	68.61 (15.62)	65.00	37-100

AI-Informed VWP Experiment

Comprehension Task

The percentage accuracy for the task of clicking the mentioned object was 96% (*SD* = .19), and 88% for the comprehension questions (*SD* = .32).

Statistical Modeling

In both studies, CPA results indicated that L1 Arabic speakers significantly directed more attention to the target picture in prediction trials than in baseline trials before the auditory presentation of the object. In contrast, CPA findings showed that L2 Arabic speakers directed either later (Study 1) or limited target fixations (Study 2) before hearing the object relative to L1 participants.

Results from the verb region (TW1) model were as follows. A significant Trial x Time x Group interaction (β = -0.71, *SE* = 0.09, 95% CI [-0.90, -0.53]) found that L1 speakers were more likely to fixate on the target picture in prediction trials than in baseline trials over time across both Study 1 and 2. This result shows that Arabic L1 participants predictively used the verbal number marking to anticipate the target object before it was mentioned during the verb region (Figure 4). In contrast, when L2 speakers heard a singular verb, they did not utilize the verbal number information as a predictive cue to anticipate the target object. Instead, they directed more attention to the target picture over time in baseline trials compared to prediction trials.

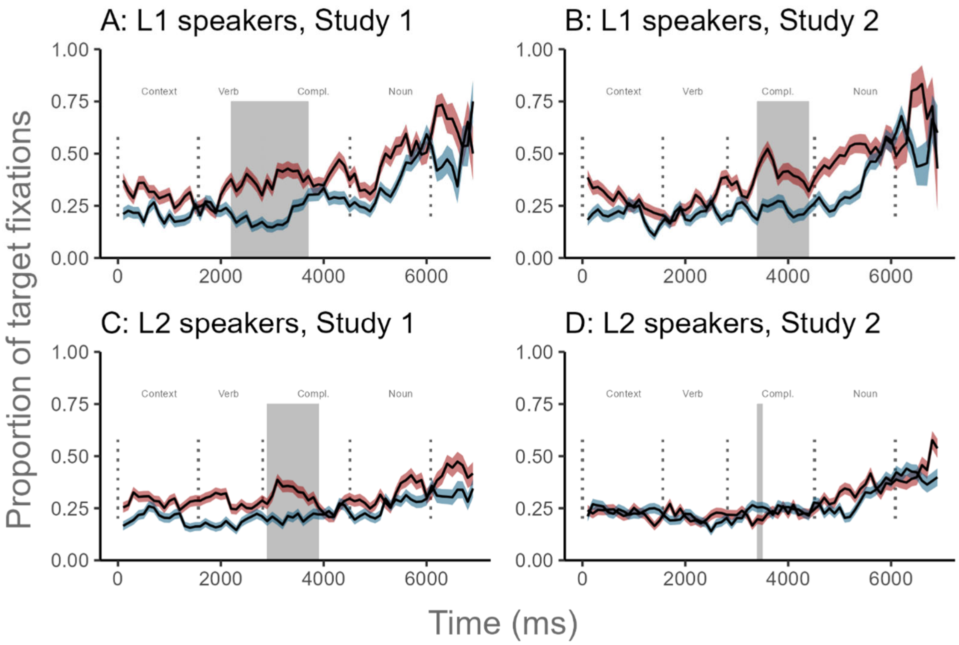


Figure 4. Time course of fixation proportions for target in the prediction (red lines) and baseline conditions (blue lines) per Group (L1 speakers vs. L2 speakers) and Study (Study 1 = simple past verb; Study 2 = past progressive verb). Ribbons indicate the standard error. Dotted lines indicate the mean onset and offset of word durations in the sentences. The grey-shaded area indicates a significant CPA cluster. Compl.: complement.

Results from the noun region (TW2) model were as follows. The Trial x Time x Group interaction (β = 0.83, *SE* = 0.26, 95% CI [0.32, 1.33]) emerged as significant, suggesting that over time L1 participants directed more attention to the target picture in the baseline trials than prediction trials during the noun region. Meanwhile, L2 speakers did not increase their target fixations in baseline trials compared to prediction trials as the sentence unfolded.

Discussion

AI-generated Arabic auditory stimuli were perceived as significantly less acceptable than human-generated recordings. However, AI-formulated stimuli triggered the target psycholinguistic effect (Koch et al., 2023). L1 Arabic speakers showed predictive number processing across both verb types, while L2 Arabic speakers exhibited either delayed or limited prediction.

Experiment 3: Priming Coordination in Comprehension

This experiment (Exp 3) replicated Wei et al.'s (2023) study to investigate the effects of L2 syntactic priming in sentence comprehension. Wei et al.'s study was selected since it used the SPR design and publicly shared its stimuli and analysis scripts. Based on prior research, we expected that L2 participants would show faster reading times (RTs) when reading sentences with similar coordinated subjects compared to those with different coordinated subjects.

Procedure and Materials

AI-Stimuli Acceptability Task

Participants

Ten English native speakers and 10 L2 English speakers were recruited from Prolific to assess the grammaticality of AI sentences from diverse perspectives. All participants provided their informed consent before completing the task and received financial compensation. The L2 participants came from seven L1 backgrounds.

Materials

Sixteen AI-generated sentences were developed using ChatGPT-3. The prompt provided instructions about SPR, and the target structure outlined three criteria for developing an appropriate target stimulus, and presented only one example sentence per condition. Several prompts were tested to obtain well-designed stimuli that follow prior practices (Wei et al., 2023), and the prompt that yielded the most accurate responses was retained. Additionally, 16 human-designed items with an identical structure were adopted from previous research (Wei et al., 2023). The target structure across all sentences was Noun Phrase (NP) coordination. The acceptability task adopted a 5-point Likert scale (1 = totally unacceptable; 5 = totally acceptable) from Wei et al. (2023).

The 32 sentences were assigned to four lists. Each list presented each item in one condition and included 16 fillers with unrelated structures. Half of the fillers contained a grammatical error to make the task reasonable. Each experimental item was followed by a filler. Experimental items were randomized per list. Participants were randomly assigned to one of the lists. Twelve of the experimental and filler trials were followed by a simple yes/no comprehension question. Both L1 and L2 Speakers demonstrated high accuracy in the comprehension task (99% (SD = .11), 98% (SD = .13), respectively).

Procedures

Participants completed the task online via Gorilla.sc. First, they were given instructions and examples of what constitutes a grammatically correct and incorrect sentence. In each trial, participants saw a sentence in the center of the screen, accompanied by a 5-point Likert scale positioned below. Participants used their mouse to select one value from the scale. After selecting a value, they clicked on the "Next" button to advance to a new trial. No time limit was imposed.

Statistical Analysis

Ratings data were analyzed using Bayesian mixed-effects ordinal logistic regression via the R package brms (Bürkner, 2017) to account for the repeated measures ordinal data. The dependent variable was ordinal ratings, and the fixed effects included item condition (sum coded: different NP = -.5, similar NP = .5), speech condition (sum coded: Human = -.5, AI = .5), and group (sum coded: L1 speaker = -.5, L2 speaker = .5). The model included the maximal random structure that converged (Barr et al., 2013): random intercepts by item and participant. The model was fitted with weakly informative priors to regularize parameter estimates (Lemoine, 2019). The Lazerhawk R package (Clark, 2024) was used to calculate the coefficients' standard error.

doi:10.20944/preprints202404.1074.v2

AI-Informed SPR Task

Participants

A total of 31 L2 English speakers were recruited from Prolific and completed Exp 3. Participants came from 10 L1 backgrounds. All participants gave their informed consent before the study and were rewarded financially.

Materials

Sixteen experimental sentences were constructed, which contained syntactic coordination at the beginning of the sentence. Half of the experimental sentences contained two conjoined Adjective Phrases (AdjPs) as the subject (i.e., a kind teacher and a diligent student), and the other half contained a relative clause and an AdjP (i.e., a teacher who is kind and a diligent student). This study created 32 fillers using ChatGPT3, which contained unrelated structures (i.e., active and passive constructions, which did not include coordination). A simple yes/no comprehension question was developed for all experimental and filler sentences. Two lists were created, with each presenting one condition of each item. Each experimental item was followed by two fillers. Participants were randomly assigned to one of the lists.

Procedures

The procedure steps closely followed Wei et al. (2023). Each sentence was presented in four segments. First, a fixation point appeared for 500 ms at the center of the screen to signal the place of the sentence segments. Then, participants saw the first segment of the sentence. This segment was displayed until participants pressed the space bar, which was replaced by the second segment. Each segment was presented in the middle of the screen. After presenting each sentence, participants answered a yes/no question by clicking on the "yes" or "no" box and received no feedback. Participants started with four practice trials to familiarize themselves with the procedure. Then, they completed the main SPR experiment. After completing half of the SPR experiment, participants were given a short optional break. After finishing the experiment, participants completed a background questionnaire. Overall, this took around 10 minutes.

Statistical Analysis

Two regions of interest were analyzed. The first region covered the second noun phrase (NP2) and was analyzed as the critical region (e.g., Dubey et al., 2005; Sturt et al., 2010; Wei et al., 2023). The second region consists of the verb and the complement (spillover region). This region was included to examine potential spillover effects. Following prior research (Wei et al., 2023), the RT data was cleaned prior to the analysis. First, trials that had RTs lower than 50 ms or higher than 10,000 ms were removed. Second, trials that we answered incorrectly were also excluded. Overall, approximately 4.59% of data was lost after the cleaning process.

Two linear mixed-effects regression models (LMER) were built per analysis region using the R package lme4. Raw RTs were log-transformed to reduce data skew and meet the normality assumption of LMER (Winter, 2019). The dependent variable was log-transformed RTs, and the fixed effects were item condition (sum coded: different NP = -.5, similar NP = .5), and phrase length (continuous), and the random structure that converged (Barr et al., 2013): random intercepts by item and participant. The model assumptions (linearity, homoscedasticity, normality, and independence) were checked via the R package performance (Lüdecke et al., 2021) and found to be appropriate.

Results

AI-Stimuli Acceptability Task

The descriptive statistics for the acceptability ratings are presented in Figure 5. The Bayesian ordinal logistic regression revealed significant main effects for item condition ($\beta = 0.50$, SD = 0.19, 95%

doi:10.20944/preprints202404.1074.v2

CI [0.11, 0.87]) and speech condition (β = 0.94, SD = 0.20, 95% CI [0.55, 1.35]). This indicates that L1 and L2 speakers gave higher ratings for similar NP sentences compared to different NP and perceived AI-generated sentences as more acceptable than human-generated ones. There was no significant effect for group (L1 speaker vs. L2 speaker) on acceptability ratings. Overall, these results suggest that both L1 and L2 English speakers perceived AI-generated sentences as acceptable as human-composed sentences.

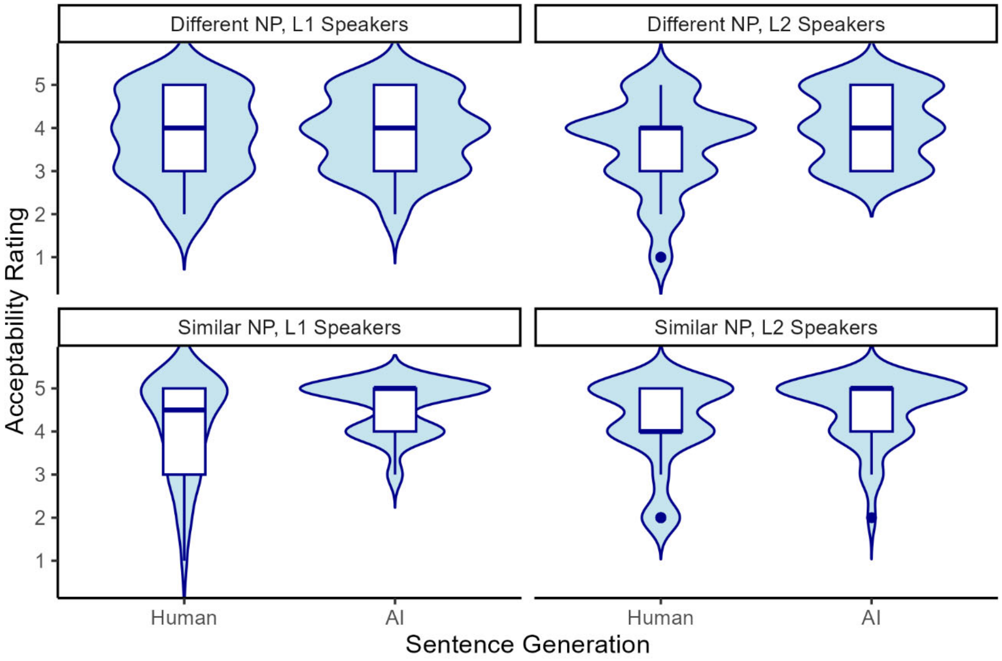


Figure 5. Distribution of L1 and L2 English speakers' ratings by item condition (Different NP vs. Similar NP) and speech condition (Human vs. AI).

AI-Informed SPR Task

Comprehension Task

All participants scored above 80% on the comprehension task. The percentage accuracy for the comprehension questions was 95% (SD = .052).

Statistical Modeling

A descriptive summary of RT data is presented in Table 4 and Figure 6. Numerical results tentatively suggested that L2 participants showed faster RTs when reading sentences with similar NP subjects compared to sentences with different NP subjects during the critical region. However, LMER results revealed no significant difference in RTs between the two item conditions during the critical region (NP2) ((β = -0.03, SD = 0.07, t = -0.41, p = .682, 95% CI [-0.186, 0.123]) and the spillover region (the verb + complement region) (β = 0.00, SD = 0.07, t = 0.03, p = .978, 95% CI [-0.143, 0.147]).

Table 4. Mean reaction times (raw RTs, in ms) by condition for the critical region (NP2).

NP1	NP2	n	RT	SD	SE
AdjP	AdjP	248	982.41	594.11	37.73
AdjP	RC	248	1017.86	572.02	36.32

Note. NP = noun phrase; AdjP = adjective phrase; RC = relative clause.

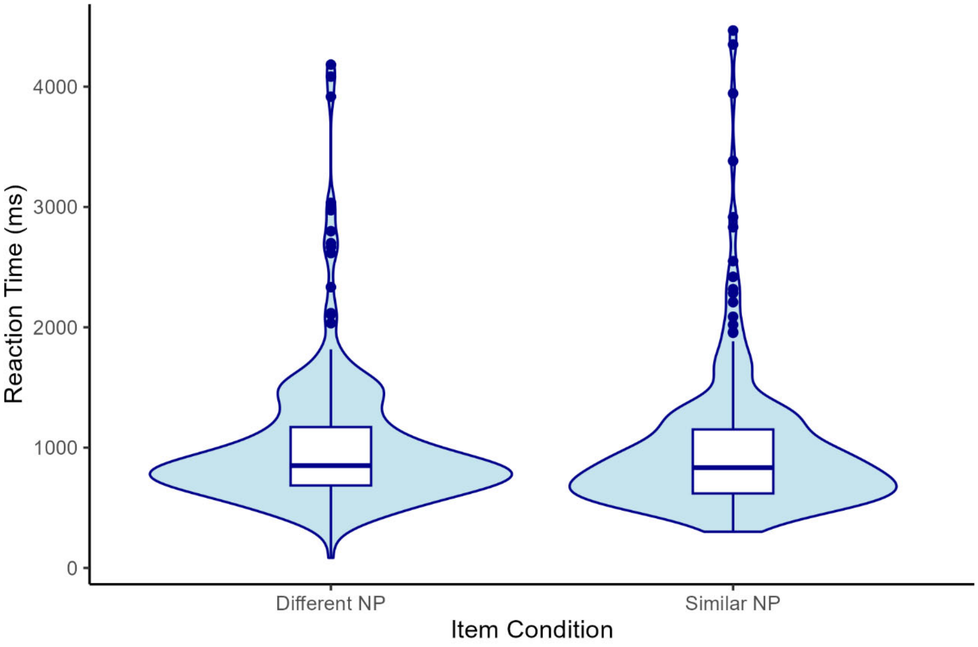


Figure 6. Distribution of raw RTs by condition for the critical region (NP2).

Discussion

Findings from Exp 3 indicate that AI-generated experimental sentence stimuli can be viewed as acceptable as human-constructed sentences. Nevertheless, AI-constructed stimuli did not yield a significant L2 syntactic priming effect, possibly due to the small target effect, which could be better captured in the presence of an interaction (Wei et al., 2023) or by using a larger number of items/participants (Marsden et al., 2018).

General Discussion

Despite its well-attested capabilities, the use of ChatGPT/AI as a stimuli developer has been relatively less examined in the bilingual literature. Sentence stimuli are widely utilized in the psycholinguistic field (Jegerski & VanPatten, 2013), yet existing psycholinguistic stimuli generation methods mainly output individual words (e.g., C. Gao et al., 2023; Taylor et al., 2020). A potentially cost-effective approach for developing sentence stimuli is via the application of AI technologies. As such, this study examined the acceptability and validity of AI-generated auditory and written stimuli in three psycholinguistic experiments. Across the experiments, participants perceived the quality of AI-created stimuli as equal to or surpassing those developed by experienced humans, with variation by the target language. AI-constructed stimuli triggered morphosyntactic prediction effects (Exp 2), but not semantic prediction (Exp 1) or syntactic priming (Exp 3) effects. These results indicate the acceptability of AI-developed psycholinguistic items and provide some support for their validity. We elaborate on these findings below and discuss their implications for research.

Acceptability of AI-Produced Stimuli

This study provided new evidence that L1/L2 English speakers are more likely to perceive AI-composed auditory stimuli (Exp 1) and written sentences (Exp 3) as human-like. This corroborates linguists' judgment that AI outputs appropriate psycholinguistic stimuli tailored to specific linguistic features (Bae, 2024). However, unlike previous research, the present study collected acceptability ratings from a group of L1 and/or L2 speakers to provide a more objective and comprehensive assessment of acceptability.

The incorporation of AI into the stimuli design process might thus be a welcome move, potentially saving researchers' time and financial expenditure. The benefits of AI-augmented design could be more pronounced during the construction of auditory stimuli, which require a great deal of care so that each component across numerous sentences has the same duration. In VWP, auditory stimuli often require multiple re-recordings to achieve the desired level of audio quality (e.g., Chun, 2018). This process can significantly increase the time and financial costs associated with stimuli construction. In addition, VWP experiments may include a sizable number of items ranging from 16 (Altmann & Kamide, 1999) to 56 items (Corps et al., 2022), which can further double with counterbalanced designs. The integration of AI-speech synthesis tools in VWP offers three benefits. First, it reduces the time spent on audio recording, thereby giving researchers more time to attend to the soundness of the experiment. Second, it reduces the financial costs of conducting research as AI speech generation tools tend to be more affordable than professional voice actor services. Finally, it facilitates the inclusion of a sizable number of experimental items within a VWM study, leading to a more robust investigation of the effect of interest (Prystauka et al., 2023).

Another important finding from the current study is that the perceived acceptability of AI-generated auditory stimuli diverged across two experiments. English AI auditory stimuli were perceived as acceptable (Exp 1) but not Arabic AI auditory stimuli (Exp 2). Even though the two experiments used the same speech naturalness task, they are distinct in four ways. The two experiments did not only utilize two different target languages (Exp 1: English; Exp 2: Arabic), but also different AI-speech synthesis tools, participant groups, and number of items. The TTS technique employed in Exp 2 had lower quality than that of Exp 1, suggesting that the quality of AI speech varied across the experiments. Further, two different groups of speakers completed the naturalness speech task: L2 speakers in Exp 1 and L1 speakers in Exp 2. These two groups differ in their language experience, which likely influenced their definition of what constitutes a naturally sounding speech, resulting in variations in their ratings. Additionally, a larger number of AI-produced items was included in Exp 1 (N = 16) than in Exp 2 (N = 6), which could have impacted the results. As there are many potential differences between Exp 1 and Exp 2, it is difficult to pinpoint the actual reason for the observed divergence in AI-speech acceptability ratings across the two experiments.

Validity of AI-Produced Stimuli

The AI-developed stimuli in the present study successfully replicated one known effect in the psycholinguistic literature: morphosyntactic prediction (Exp 2), but not the other two well-studied effects of semantic prediction (Exp 1) and syntactic priming (Exp 3). Likewise, research on corpus linguistics and discourse analysis reported mixed findings regarding the validity of ChatGPT/AI as a data analysis tool (Anthony, 2023; Curry et al., 2024; Lin, 2023; Uchida, 2024; Zappavigna, 2023). These studies found that AI tools replicated human results in a few tasks, (e.g., semantic classification, identification of lexico-grammatical patterns), but not in others (e.g., genre identification, concordance analysis). However, our findings cannot be directly compared to prior studies due to differences in the role of AI across the studies. Instead, we compared results within our current study's experiments or contrasted our findings with the original studies when feasible.

As both Exp 1 and 2 share many similarities (e.g., VWP design, AI auditory stimuli, adult participants), it is possible to compare their results to understand the non-significant prediction effect in Exp 1. The current pattern of results does not fit with theoretical (Pickering & Gambi, 2018) and empirical (Carroll et al., 2024) proposals, which posit that L2 speakers might find it easier to engage in semantic prediction (Exp 1) than morphosyntactic prediction (Exp 2). In order to explain the results, we hypothesize that the limited prediction effect in Exp 1 may be attributable to three potential factors. These factors include: the characteristics of the recruited speaker group, the presence of a preview window for the stimuli, and the presentation speed of the auditory stimuli.

First, Exp 1 focused on L2 English speakers, while Exp 2 investigated both L1 and L2 Arabic speakers. The L2 speakers across Exp 1 and 2 demonstrated delayed or limited prediction effects (See Figures 2 and 4), whereas the L1 speakers in Exp 2 always exhibited significant effects (See Figure 4). Thus, it is likely that AI-generated auditory stimuli did not yield the expected semantic prediction effect in Exp 1 because it mainly involved L2 speakers. Support for this idea comes from the observation that the original study by Altmann and Kamide (1999) recruited only L1 English speakers. However, this explanation is not consistent with the well-observed finding that L2 speakers could use semantic cues from the verb to pre-activate the upcoming noun (e.g., Carroll et al., 2024; Chun et al., 2021; Ito et al., 2018; Schlenter, 2019).

Therefore, a second potential factor that could have reduced the prediction effects in Exp 1 is the absence of a preview time. Some VWP studies include a non-related constituent at the beginning of the auditory stimuli (Koch et al., 2021) or present the visual display for 500-1000 ms before the onset of the auditory sentence (Prystauka et al., 2023) so that participants could have more time to process the visual input and consequently show the expected predictive behavior. In the present study, all auditory stimuli in Exp 2 started with an unrelated phrase, which gave participants time to preview the picture set (Figure 3), while Exp 1 did not integrate any form of visual preview. A third crucial factor that might have mitigated prediction effects in Exp 1 is the speed of the auditory stimuli. On average, the English AI-formulated auditory stimuli in Exp 1 included four syllables per second, while the Arabic stimuli in Exp 2 presented three syllables per second. Speed of speech rate has been shown to influence linguistic prediction effects (Huettig & Guerra, 2019).

Findings from Exp 3 could be interpreted in light of the original study (Wei et al., 2023). Exp 3 differed from Wei et al.'s study in two crucial ways, which could have contributed to the observed null syntactic priming effect in Exp 3. First, the L2 English participants were primed with only one structure (AdjPs) in Exp 3, while they were primed with two structures (AdjPs, relative clause) in Wei et al.'s study. In Wei et al., a significant priming effect emerged only when considering the interaction between the two priming conditions, not when examining the main effects of each condition. Therefore, the inclusion of only one priming condition in Exp 3 might have made it difficult to replicate the syntactic priming effect observed in the original study. Second, the L2 participants in Exp 3 came from varied L2 backgrounds, whereas those in Wei et al. were all L1 Chinese speakers. Although the influence of L1 on L2 syntactic priming may not be robust (Flett et al., 2013; Shin & Christianson, 2012; Wei et al., 2023), it is possible that L1 experience mediates priming, a factor that was not controlled in Exp 3.

Implications for Psycholinguistic Research

The current study highlighted the good acceptability and partial validity of AI-generated psycholinguistic stimuli, suggesting that the role of AI as a stimuli developer holds some promise. The following will discuss the research implications of these findings.

First, by leveraging AI, the stimuli design process can be optimized, resulting in enhanced efficiency and reduced costs. The incorporation of AI in research can enhance efficiency by automating time-consuming tasks such as recording sentences and constructing a large number of filler sentences. Another advantage of AI technologies is their relative affordability, which could cut the costs of the stimuli creation process. In the current study, a TTS AI speech tool priced at approximately 24 USD per month offered unlimited sentence generation, whereas the least expensive freelance voice actor charged 28 USD for recording exactly 16 sentences. This difference will likely increase with an increasing number of items.

Second, AI-generated auditory stimuli offer increased control, leading to more consistent and replicable experiments. Note this advantage cannot be said about AI-generated written stimuli due to the inherent variability in the text generated by LLMs (Megahed et al., 2024). Human-generated auditory stimuli from the same speaker are likely to be prone to variation due to the type of recording device, background noise, and room acoustics. AI tools reduce this variability as they can be configured with pre-defined criteria to specify pitch, intonation, background noise, and duration of pauses. This in turn has the potential of increasing standardization across studies. Different researchers may use the same AI TTS tool, which could foster consistency and replicability of studies conducted in the same language.

Third, the use of AI in stimuli development can arguably increase psycholinguistic research diversity. Utilizing AI as a stimuli developer can open the doors for researchers from different backgrounds, including those with limited funding or access to trained professionals, to participate in building scientific knowledge. Specifically, early career researchers, individual researchers, and graduate students might benefit from the integration of AI into their research process as it minimizes the time and money needed to carry out a psycholinguistic experiment. AI-supported stimuli creation could help bridge the existing gap between research output from WEIRD and non- WEIRD contexts (Blasi et al., 2022; Plonsky, 2023), facilitating a more nuanced understanding of language processing and acquisition.

Fourth, as AI is rapidly transforming research methodologies, it becomes increasingly important for journal editors to recognize its impact and consider providing guidance to researchers on how to effectively incorporate AI in stimuli design. For example, journals may recommend which AI tools are appropriate for stimuli creation, provide the needed documentation steps for formulating AI-generated stimuli to allow future replications, and address potential ethical considerations associated with AI-supported stimuli generation (e.g., voice cloning) to ensure fair research practices.

Limitations and Future Directions

Although this is one of the few studies examining the efficacy of AI in the development of psycholinguistic stimuli, it has several limitations. First, the present results are constrained by the currently available AI technologies, and it is quite possible that newer AI technologies will generate more human-like stimuli in a wider number of languages beyond English. A promising topic for future research is to explore whether current limitations in the generation of auditory stimuli for less researched languages (e.g., Arabic) can be improved by utilizing voice cloning technology (Arik et al., 2018). Second, L1 speakers were not included in all experiments, and they may show different behavior patterns than L2 speakers (see Exp 2 results). Future research may consider testing AI-generated stimuli on both L1 and L2 speaker groups to capture a more accurate picture of the validity of AI-formulated sentences. Second, unlike the original study, Exp 3 did not examine the syntactic priming effect using an interaction design due to limited funding.

Fourth, this study used AI tools to generate only single sentences. The current results may not apply to common stimuli with different lengths, such as multiple sentences (Brothers et al., 2017) and short stories (Rodd et al., 2016), and more research is needed to address this gap. Fifth, AI was used to generate written sentences in English only, and the current findings may not translate to AI-generated sentences in other languages. It is important to test the acceptability and validity of AI-composed stimuli across various languages to enhance our understanding of its limitations and advantages. Sixth, as VWP incorporates a visual display of several objects, it remains unknown whether AI-generated pictures could equally substitute human-generated drawings. Finally, exact replication of written sentence stimuli cannot be guaranteed with the current versions of ChatGPT even when the same prompt is provided (Megahed et al., 2024). This raises the issue of replicability for AI-augmented studies, which should be addressed in the field as the use of AI in research will likely increase over time.

Conclusions

This study examined the acceptability and validity of AI-created stimuli for two common psycholinguistic designs, VWP and SPR, across three experiments and in two languages (English, Arabic). Participants viewed English AI-generated stimuli as human-like but not for Arabic. The validity of AI-formulated stimuli is likely dependent on the study design, with only Exp 2 showing the expected psycholinguistic effect but not Exp 1 and 3. These findings underscore the promising role of AI as a stimuli developer, which could enhance efficiency and diversity in psycholinguistic research. Future research may explore how AI-generated stimuli could be improved (e.g., incorporating data from psycholinguistic databases) to maximize the advantages of AI technologies and ultimately facilitate academic research.

Statement Concerning Research Involving Human Participants: This research was conducted in compliance with the ethical regulations governing research involving human subjects in the Kingdom of Saudi Arabia, issued by the National Committee for Bioethics Research on Human Subjects in Saudi Arabia (KSU-REC). In accordance with KSU-REC and the Helsinki Declaration, this study ensured the protection of participants' rights, welfare, and confidentiality. Participants were informed about the experiment procedures and voluntarily agreed to participate. The participants indicated that they understood their right to withdraw at any time, and their written consent was obtained before the experiment. The primary research protocol, which took the form of an eye-tracking task or a self-paced reading task, did not cause any harm to the participants. Participants were identified by random ID numbers to ensure data confidentiality.

Data Availability Statement: All materials, data and R codes are available at:

https://osf.io/6us4d/?view_only=37ab6e17dcc7436b9b21d685e8690c70

Conflict of interest: The author declares that there is no conflict of interest.

References

1. Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/s0010-0277(99)00059-1
2. Anthony, L. (2023). Corpus AI: Integrating Large Language Models (LLMs) into a Corpus Analysis Toolkit. *The 49th Annual Convention of the Japan Association for English Corpus Studies (JAECS)*.
3. Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x
4. Arik, S. O., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). *Neural Voice Cloning with a Few Samples*.
5. Bae, H. (2024). *Chatgpt as a Research Assistant in Experimental Linguistics. SSRN.* https://doi.org/http://dx.doi.org/10.2139/ssrn.458554
6. Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers*, *34*, 424–434. https://doi.org/https://doi.org/10.3758/BF03195471
7. Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.
8. Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting-linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01
9. Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/https://doi.org/10.1016/j.tics.2022.09.015
10. Bosch, J. E., Chailleux, M., Yee, J. E., Guasti, M. T., Arosio, F., & Ayoun, D. (2022). Prediction on the basis of gender and number in Mandarin-Italian bilingual children. In D. Ayoun (Ed.), *Studies in Bilingualism* (pp. 243–271). John Benjamins.
11. Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. https://doi.org/10.1016/j.jml.2016.10.002
12. Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01
13. Carroll, R., Patarroyo, A., & Hopp, H. (2024). Predictive L2 sentence processing in noise: Differential effects across linguistic cues. *15th Speech in Noise Workshop*. https://doi.org/10.5281/zenodo.10512307
14. Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, *2*(3).
15. Chang, Y. N., Hsu, C. H., Tsai, J. L., Chen, C. L., & Lee, C. Y. (2016). A psycholinguistic database for traditional Chinese character naming. *Behavior Research Methods*, *48*, 112–122. https://doi.org/https://doi.org/10.3758/s13428-014-0559-7
16. Cheng, X., Schafer, G., & Riddell, P. M. (2014). Immediate auditory repetition of words and nonwords: an ERP study of lexical and sublexical processing. *PloS One*, *9*(3). https://doi.org/10.1371/journal.pone.0091988
17. Chun, E. (2018). The role of prediction in adaptation: An evaluation of error-based learning accounts. [Unpublished doctoral dissertation, University of Florida].

18. Chun, E., Chen, S., Liu, S., & Chan, A. (2021). Influence of syntactic complexity on second language prediction. In E. Kaan & T. Grüter (Eds.), *Prediction in Second Language Processing and Learning* (pp. 70–89). John Benjamins. https://doi.org/https://doi.org/10.1075/bpa.12.04chu
19. Clark, M. (2024). *lazerhawk: Miscellaneous functions mostly inspired by synthwave* (Version 0.3.0). [R package]. https://github.com/m-clark/lazerhawk.
20. Contemori, C. (2021). Changing comprehenders' pronoun interpretations: Immediate and cumulative priming at the discourse level in L2 and native speakers of English. *Second Language Research*, *37*(4), 573–586. https://doi.org/10.1177/0267658319886644
21. Corps, R. E., Brooke, C., & Pickering, M. J. (2022). Prediction involves two stages: Evidence from visual-world eye-tracking. *Journal of Memory and Language*, *122*.
22. Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1).
23. Dubey, A., Sturt, P., & Keller, F. (2005). Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. *Proceedings of the Human Language Technology Conference Conference on Empirical Methods in Natural Language*, 827–834.
24. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ..., & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2023.102642
25. Flett, S., Branigan, H. P., & Pickering, M. J. (2013). Are non-native structural preferences affected by native language preferences? *Bilingualism: Language and Cognition*, *16*(4), 751–760. https://doi.org/10.1017/S1366728912000594
26. Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, *6*(1). https://doi.org/https://doi.org/10.1038/s41746-023-00819-6
27. Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). Scope: The south carolina psycholinguistic metabase. *Behavior Research Methods*, *55*(6), 2853–2884. https://doi.org/https://doi.org/10.3758/s13428-022-01934-0
28. Garrido Rodriguez, G., Norcliffe, E., Brown, P., Huettig, F., & Levinson, S. C. (2023). Anticipatory Processing in a Verb-Initial Mayan Language: Eye-Tracking Evidence During Sentence Comprehension in Tseltal. *Cognitive Science*, *47*(1).
29. Goldberg, Y. (2019). *Assessing BERT's syntactic abilities. ArXiv.* https://doi.org/1901.05287.
30. Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 26–26.
31. Herrmann, B. (2023). The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *International Journal of Speech Technology*, *26*(2), 395–415.
32. Heyman, T., & Heyman, G. (2023). The impact of ChatGPT on human data collection: A case study involving typicality norming data. *Behavior Research Methods*, 1–8. https://doi.org/https://doi.org/10.3758/s13428-023-02235-w
33. Hu, J., & Levy, R. P. (2023). Prompting is not a substitute for probability measurements in large language models. ArXiv. https://doi.org/https://doi.org/10.48550/arXiv.2305.13264
34. Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research*, *1706*, 196–208. https://doi.org/10.1016/j.brainres.2018.11.013
35. Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. , 137(2),. *Acta Psychologica*, 151-171.
36. Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, *21*(2), 251–264. https://doi.org/10.1017/S1366728917000050
37. Ito, A., & Knoeferle, P. (2022). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*, 1–33. https://doi.org/10.3758/s13428-022-01969-3
38. Ito, A., Nguyen, H. T. T., & Knoeferle, P. (2023). German-dominant Vietnamese heritage speakers use semantic constraints of German for anticipation during comprehension in Vietnamese. *Bilingualism: Language and Cognition*, 1–18. https://doi.org/https://doi.org/10.1017/S136672892300041X
39. Jegerski, J., & VanPatten, B. (2013). Research methods in second language psycholinguistics. Routledge.
40. Karaca, F., Brouwer, S., Unsworth, S., & Huettig, F. (2023). Morphosyntactic predictive processing in adult heritage speakers: Effects of cue availability and spoken and written language experience. *Language, Cognition, and Neuroscience.* https://doi.org/https://doi.org/10.1080/23273798.2023.2254424
41. Koch, E., Bulté, B., Housen, A., & Godfroid, A. (2021). Using verb morphology to predict subject number in L1 and L2 sentence processing: A visual-world eye-tracking experiment. *Journal of the European Second Language Association*, *5*(1), 115–132. https://doi.org/10.22599/jesla.79

42. Koch, E., Bulté, B., Housen, A., & Godfroid, A. (2023). The predictive processing of number information in subregular verb morphology in a first and second language. *Applied Psycholinguistics*, 1–34.
43. Kuteeva, M., & Andersson, M. (2024). Diversity and Standards in Writing for Publication in the Age of AI—Between a Rock and a Hard Place. *Applied Linguistics*. https://doi.org/https://doi.org/10.1093/applin/amae025
44. Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241. https://doi.org/https://doi.org/10.1111/cogs.12414
45. Lee, G., & Kim, H. Y. (2024). Human vs. AI: The battle for authenticity in fashion design and consumer response. *Journal of Retailing and Consumer Services*, *77*. https://doi.org/https://doi.org/10.1016/j.jretconser.2023.103690
46. Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*. https://doi.org/10.1111/oik.05985
47. Lewis, J. R. (2018). Investigating MOS-X ratings of synthetic and human voices. *Voice Interaction Design*, *2*(1).
48. Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, *5*(6). https://doi.org/https://doi.org/10.1016/S2589-7500(23)00083-3
49. Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, *3*(3), 1–5. https://doi.org/https://doi.org/10.1016/j.acorp.2023.100065
50. Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60). https://doi.org/https://doi.org/10.21105/joss.03139
51. Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861–904. https://doi.org/https://doi.org/10.1017/S0142716418000036
52. Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. ArXiv.
53. McDonough, K., & Trofimovich, P. (2009). *Using priming methods in second language research*. Routledge.
54. Megahed, F. M., Chen, Y. J., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2024). How generative AI models such as ChatGPT can be (mis) used in SPC practice, education, and research? An exploratory study. *Quality Engineering*, *36*(2), 287–315. https://doi.org/https://doi.org/10.1080/08982112.2023.2206479
55. Petersen, O. H. (2021). Inequality of research funding between different countries and regions is a serious problem for global science. 2(6). https://doi.org/10.1093/function/zqab060
56. Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158
57. Plonsky, L. (2023). Sampling and generalizability in Lx research: A second-order synthesis. *Languages*, *8*(1). https://doi.org/https://doi.org/10.3390/languages8010075
58. Prystauka, Y., Altmann, G. T., & Rothman, J. (2023). Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods*, 1–19.
59. Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, 16–37. https://doi.org/https://doi.org/10.1016/j.jml.2015.10.006
60. Saslow, M. G. (1967). Latency of saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1030–1033. https://doi.org/10.1364/JOSA.57.001030
61. Schlenter, J. (2019). Predictive language processing in late bilinguals: Evidence from visual-world eye-tracking. Doctoral dissertation. Universität Potsdam.
62. Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*, 1258–1270. https://doi.org/https://doi.org/10.3758/s13428-018-1099-3
63. Shin, J. A., & Christianson, K. (2012). Structural priming and second language learning. *Language Learning*, *62*(3), 931–964. https://doi.org/10.1111/j.1467-9922.2011.00657.x
64. Soares, A. P., Costa, A. S., Machado, J., Comesaña, M., & Oliveira, H. M. (2017). The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods*, *49*, 1065–1081. https://doi.org/https://doi.org/10.3758/s13428-016-0767-4
65. Sturt, P., Keller, F., & Dubey, A. (2010). Syntactic priming in comprehension: Parallelism effects with and without coordination. *Journal of Memory and Language*, *62*(4), 333–351. https://doi.org/https://doi.org/10.1016/j.jml.2010.01.001
66. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, Tavakol, M., Dennick, R. https://doi.org/10.5116/ijme.4dfb.8dfd

67. Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*, 2372–2382. https://doi.org/10.3758/s13428-020-01389-1
68. Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, 1–19. https://doi.org/10.3758/s13428-024-02337-z
69. Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, *4*(1). https://doi.org/https://doi.org/10.1016/j.acorp.2024.100089
70. Verma, A., Sikarwar, V., Yadav, H., Jaganathan, R., & Kumar, P. (2022). Shabd: A psycholinguistic database for Hindi. *Behavior Research Methods*, *54*(2), 830–844. https://doi.org/https://doi.org/10.3758/s13428-021-01625-2
71. Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. https://doi.org/https://doi.org/10.3758/BRM.41.2.534
72. Voeten, C. C. (2023). permutes: Permutation Tests for Time Series Data (R package version 2.8).
73. Wei, H., Boland, J. E., Zhang, C., Yang, A., & Yuan, F. (2023). Lexically Independent Structural Priming in Second Language Online Sentence Comprehension. *Language Learning*. https://doi.org/https://doi.org/10.1111/lang.12588
74. Winter, B. (2019). Statistics for linguists: An introduction using R. Routledge.
75. Zappavigna, M. (2023). Hack your corpus analysis: how AI can assist corpus linguists deal with messy social media data. *Applied Corpus Linguistics*, *3*(3). https://doi.org/https://doi.org/10.1016/j.acorp.2023.100067

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.