Embodiment for Spatial Metaphors of Abstract Concepts Differs Across Languages in Chinese-English Bilinguals

Wei, Yufen; Yang, Wenwen; Oppenheim, Gary; Hu, Jiehui; Thierry, Guillaume

Language Learning

DOI:
10.1111/lang.12632

Published: 04/04/2024

Peer reviewed version

EMPIRICAL STUDY

Embodiment for Spatial Metaphors of Abstract Concepts Differs Across Languages in Chinese-English Bilinguals

Abstract: Embodied cognition posits that processing concepts requires sensorimotor activation. Previous research has shown that perceived power is spatially embodied along the vertical axis. However, it is unclear whether such mapping applies equally in the two languages of bilinguals. Here, using event-related potentials, we compared spatial embodiment correlates in participants reporting the source of auditory words as being presented from above or below their sitting position. Chinese-English bilinguals responded faster for congruent presentations of high-power words (presented above) but not of low-power words (presented below) in both languages. Low-power words together also failed to modulate N400 amplitude or interact with language. However, follow-up analyses on high-power words showed congruency effects on N400 amplitude in Chinese, but not in English. Finally, English controls showed no effect. This suggests that spatial embodiment differs across languages in bilinguals, but the roles of culture and proficiency require further research.

Keywords: conceptual representation; embodied cognition; metaphor; bilingualism; ERP

Introduction

Concepts are at the very heart of what it means to be human. Theories of conceptual embodiment (Barsalou, 1999; 2008) have struggled to explain how abstract concepts (e.g., peace) may relate to activation of sensorimotor systems which are more readily associated with the processing of concrete concepts (e.g., daffodil). However, we use strategies to relate abstract concepts to concrete experiences in the world, notably by using metaphors (e.g., "the sweet taste of peace"). One interesting question is whether cognitive embodiment of abstract concepts differs across languages in bilinguals. Indeed, metaphorical references are generally implemented differently in the two languages of bilinguals due to idiosyncrasies of language expressions. In the present study, we thus ask whether metaphorical mapping similarly applies in the native (L1) and second language (L2) of Chinese-English bilinguals when they process abstract concepts. We take the example of spatial metaphors for perceived power ("king" [high-power] – up, "servant" [low-power] – down), to address this question using Event-Related Potentials (ERPs) which provide a high temporal resolution index of implicit semantic processing requiring no conscious evaluation on the part of the participant.

Background Literature

Embodied Theories of Cognition

Embodiment theory posits that accessing concepts in semantic memory systematically involves sensorimotor activation (Barsalou, 1999, 2008; Gallese & Lakoff, 2005; Glenberg & Kaschak, 2002; Pulvermüller, 2005). This proposal has received empirical support from experiments in which participants processed nouns, verbs, words representing actions performed by specific body parts, or sentences describing actions (Hauk et al., 2004; Marino et al., 2014; Santana & De Vega, 2013; Zwaan & Yaxley, 2003). For instance, an fMRI study showed that reading the words lick, pick, and kick differentially activates cortical regions involved when participants perform the bodily movements associated with these words (e.g., moving their tongue, finger, or foot; Hauk et al., 2004). Aravena et al. (2010) presented participants with sentences that described actions either congruent or incongruent with the motor experiences of their hand. They found that incongruent sentences elicited greater N400 ERP amplitudes than congruent sentences, suggesting that language comprehension involves sensorimotor activation (see Grafton & Tipper, 2012 for a review).

Concrete/Abstract Problem

A major difficulty encountered by proponents of embodiment, however, is the representation of abstract concepts. Abstract concepts have been defined as mental representations referring to entities that are neither purely physical nor spatially constrained (Barsalou & Wiemer-Hastings, 2005; see also Borghi et al., 2022). Dual Coding Theory (Paivio, 1990), for instance, assumes that abstract and concrete concepts are stored differently in long-term memory: Abstract concepts would be stored as verbal-symbolic representations, detached for direct bodily experiences, whereas concrete concepts evoke both verbal and visual-perceptual representations. Others have argued that abstract and concrete concepts are similarly grounded in perception and action (Buccino et al., 2016; Kiefer & Pulvermüller, 2012). In addition to sensory-motor experiences, abstract concepts have also been theorized to depend more than concrete concepts on metaphors, situated action, emotion, introspection, and lexical associations (Barsalou et al., 2008, 2018; Borghi, 2020; Gallese & Lakoff, 2005; Vigliocco et al., 2009).

According to Conceptual Metaphor Theory, abstract concepts are grounded in experience through mediation of metaphoric mappings to the concrete domain (e.g., Gallese & Lakoff, 2005; Lakoff & Johnson, 1999). For instance, time, a highly abstract concept, is often conceived in relation to space, a concrete domain, where experience abounds

(Casasanto & Boroditsky, 2008; Li et al., 2019). Studies looking into abstract concept embodiment have exploited the stimulus-response compatibility effect, that is, the fact that participants respond more quickly and accurately in judgement tasks when the nature of response matches some stimulus feature. Such studies have shown that the effect for concrete concepts also applies to abstract ones. For example, Meier and Robinson (2004) asked participants to indicate whether words such as "hero" and "liar" had a positive or negative meaning. They found that participants responded faster when words with a positive connotation appeared above (congruent) rather than below (incongruent) fixation.

Embodiment and Bilingualism

Although there is increasing evidence supporting embodied cognition theory, the representation of concrete and abstract concepts in bilinguals remains underexplored. Some researchers argue that unlike in the L1, sensorimotor information is only partially activated in L2 processing of action-related verbs and nouns (Ahlberg et al., 2018; Bergen et al., 2010; Buccino et al., 2017; Dudschig et al., 2014; Foroni, 2015; see Kogan et al., 2020; Kühne & Gianelli, 2019; Monaco et al., 2019 for review). Buccino et al. (2017) found that the sensorimotor system is involved in the processing of both L1 and L2 nouns denoting graspable items as compared to non-graspable ones using a go/no-go paradigm.

However, other studies have suggested that L2 sensorimotor activations require deeper semantic processing (Bai & He, 2022; Birba et al., 2020; de Grauwe et al., 2014; Tian et al., 2020; Vukovic & Shtyrov, 2014; Xue et al., 2015). For example, de Grauwe et al. (2014) used fMRI to investigate embodiment in L2 using a lexical decision task and found that semantic representations in L2 and L1 can produce activations to simple action words in motor and somatosensory areas. Xue et al. (2015) tested Chinese-English bilinguals in a sentence acceptability task using ERPs, in which participants compared high and low body-object-interaction (BOI) words (i.e., words that imply direct motor control vs. none) embedded in either a rich or poor sensorimotor context. High BOI L2 words were more acceptable to participants, processed faster, and elicited lower N400 amplitudes than low BOI L2 words, especially in a rich context, lending support to the idea that L2 concrete words are embodied.

Evidence for sensorimotor activation by L2 abstract words remains scarce, however, with most research focusing on the processing of emotion words (Dudschig et al., 2014; Vukovic & Shtyrov, 2014). Dudschig et al. (2014) found that L2 input automatically activates sensorimotor representations for implicit location words (e.g., roof) and emotion words (e.g., happy) using an adapted Stroop paradigm, in which participants had to identify word color using upward or downward finger movements to reach target buttons. Vukovic and Shtyrov (2014) measured event-related desynchronization (ERD) in native speakers of German who were also fluent in English. They presented participants with pairs of verbs that were either semantically congruent (related, but not identical action words, e.g., chewing – kissing) or incongruent (action verbs preceded by abstract words, e.g., knowing – running). They found that both languages elicit similar patterns of brain activity, with a stronger effect in German than English. Also, action verbs activated motor areas more than abstract verbs, especially when action verbs were related in meaning. The authors concluded that embodied cognition applies to both concrete and abstract words in the two languages, whilst depending on factors such as proficiency, semantic similarity, and word class.

To our knowledge, only one ERP study to date attempted to capture embodied cognition effects in bilinguals in relation to spatial temporal metaphors: Li et al. (2019) presented participants with dates in the form of days of the week or numerical years and asked them to report the interval duration between the date heard and the date of testing. Hearing the word “Friday” from a loudspeaker situated in front of participants on a

Wednesday created a mental challenge greater that hearing the same word played in their back, an effect the authors had hoped to detect because the metaphor for "the day after tomorrow" in Chinese is *hou tian*, which literally translates as "back day" in English. Chinese-English bilinguals suffered interference between the origin of spoken words in the space around them and the spatiotemporal metaphor used to refer to time within 400 ms of stimulus onset time. Strangely, however, such embodiment interference effect was found only when participants operated in L2 English.

The Case of Perceived Power

Perceived power has recently gained attention because it characterizes relatively abstract concepts (e.g., king) while also maintaining a metaphorical relation to the concrete domain of space. Schubert (2005) demonstrated that processing perceived power in a semantic judgement task spontaneously activates vertical space information. Participants responded faster and more accurately when high-power words are presented in a high relative to low screen position. Zanolie et al. (2012) used a spatial cueing paradigm to investigate whether processing perceived power automatically recruits spatial attention. Participants responded significantly faster when a high-power word was followed by a target letter appearing at an upper location compared to a lower location on the screen. They also found increased N1 amplitude for congruent relative to incongruent spatial positions.

Convergingly, Wu et al. (2016) asked participants to judge whether a target word, displayed in the upper or lower part of the screen, was associated with power or weakness. Participants responded significantly faster and displayed increased N1 amplitude in congruent relative to incongruent trials, but P3 and LPC amplitudes were modulated in the opposite direction. To examine this in L2, Wang (2016) gave the same instruction to Chinese-English bilinguals. The congruency effect on RTs was larger in L1 than L2, and stronger in participants with higher L2 proficiency. Another study conducted by Hu (2021), however, only found the congruency effect in L1.

The current study

Current evidence of perceived power embodiment in bilinguals is limited: (1) Although processing of perceived power is thought to be metaphorically mapped along the vertical axis, only two studies reported differences between languages (Wang, 2016; Hu, 2021); (2) Most of the evidence comes from behavioral data in tasks where power manipulation is explicit, which may differ from the case of natural language processing; (3) N1 and P1 modulations reported in ERP studies (Wu et al., 2016, Zanolie et al., 2012) may be heavily affected by low-level stimulus characteristics.

Here, to tackle embodiment more directly, we used an auditory source localization task in Chinese-English bilinguals. Participants reported whether spoken words in Chinese and English were played from speakers located above or below their sitting position. Words could either relate to human entities varying in power status (e.g., king/servant – "power words"), names of objects with direct spatial reference (e.g., sun/ground – "location words"), or animal names (e.g., dog), serving as fillers. High-power and high-location belonged to the congruent condition when played above and to the incongruent condition when played from below. Note that the ratio of power words, location words, and fillers was unbalanced (by design), with less than half of stimuli being congruent and acting as implicit target (prone to elicit P300-like modulations, Polich, 2007).

To our knowledge, it is the first time that embodiment of an abstract concept is tested based on direct mapping with a sensory modality associated with spatial awareness in an ERP experiment. Indeed, we did not use location descriptors (the words "up" and "down"), a visual property of the stimulus (position on the screen), or a symbolic spatial representation (arrow), but rather a direct reference to bodily experience of space (sound origin). Importantly, participants were asked to report the source of the sound in the space around them as opposed to deliberately making judgement about perceived power, thus shifting the context toward implicit conceptual processing.

We hypothesized (a) effects of embodiment on abstract cognition (i.e., behavioral and ERP effects of congruency) and (b) differences in strength of embodiment across languages (L2 effects reduced compared to L1effects). If the sensorimotor network is involved in processing spatial metaphors of perceived power, we expect shorter RTs and less errors when participants hear a word referring to a high-power word presented from above than from below. ERPs are expected to show larger P3 amplitudes in congruent than incongruent conditions due to the relatively lower frequency of congruent pair across the experiment. We also predicted larger N400 amplitudes for incongruent than congruent conditions, especially when Chinese-English bilinguals are tested in L1 Chinese. This study was pre-registered on AsPredicted.com (https://aspredicted.org/SG4_5P9).

Methods

Participants

Thirty-two Chinese-English bilinguals and 27 English native speakers with self-reported normal hearing and normal or corrected-to-normal vision participated in this experiment. Data from seven bilingual participants and two native speakers of English were excluded due to: a) Accuracy rate falling below 80% considering the simplicity of the task; b) EEG data displaying non-correctable drifting, excessive line noise, and a loss of more than five channels preventing accurate data interpolation. We thus included 25 datasets for each language group in the final analyses (Chinese-English bilinguals: 12 females, mean age = 24.1 ± 3.7; English native speakers: 17 females, mean age = 20.4 ± 3.6).

All participants were students at Bangor University and residents in the UK at the time of testing. Language background was assessed using the Language Experience and Proficiency Questionnaire and (LEAP-Q; Kaushanskaya et al., 2020; summarized in Table 1). Bilingual participants had received an average of 16 ± 2.3 years of education (college level education), had exposure to Chinese from birth, and low-to-average levels of daily exposure to English (4.0 ± 1.6 on a scale from 0 – never to 10 – always). They had an English proficiency score > 6 as measured by the International English Language Testing System (IELTS) and self-reported their language proficiency in Mandarin Chinese as very high and in English as upper intermediate (Fig. 1).

Participants gave written informed consent before taking part in this study and received cash or credits for their participation. The study was approved by the ethics committee of the School of Human and Behavioral Sciences at Bangor University (approval no. 2021-17074).

Table 1 Chinese-English bilingual participants' language background

Measure	Mean	SD
Age of Chinese acquisition	1.8	1.5
Age of English acquisition	8.0	4.1
Daily Chinese use (%)	65.3	23.2
Daily English use (%)	34.7	17.7

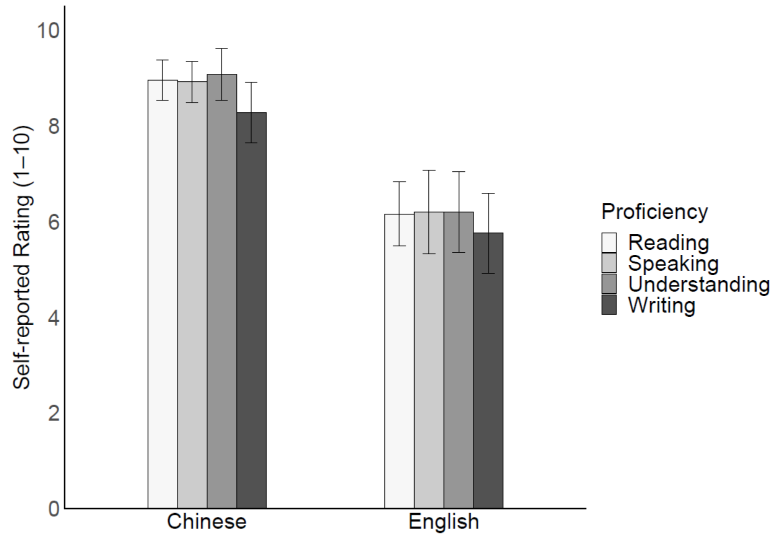

Fig. 1 Chinese-English bilingual participants' self-reported ratings of Chinese and English proficiency (10 point-scale). Error bar represents confidence interval.

Materials

Stimuli consisted of 120 Chinese words and 120 translation equivalents in English (see Appendix S1). The stimuli were further divided into high/low perceived power human words (referring to people, professions, or social status, e.g., king/servant, 60 words), high/low spatial references (denoting objects associated with higher or lower positions in space, e.g., sun/ground, 30 words) and animal names (fillers, e.g., tiger/rabbit, unrelated to space, 30 words). Stimuli were partially derived from the stimulus list compiled by Dudschig et al. (2014). All words were presented auditorily. Word recordings were downloaded from https://easypronunciation.com/zh/practice-chinese-pronunciation-online (for Chinese), and the https://dict.youdao.com/?keyfrom=cidian (for English). All the files were natural-sounding synthesized audio normalized in amplitude and resampled to 44.1 KHz with a 16-bit resolution using Adobe Audition (Version 13.0). Chinese stimuli varied in length from 240–1195 ms (M = 714 ± 17 ms). English stimuli varied in length from 243–1118 ms (M = 687 ± 15 ms).

Thirty volunteers who did not participate in the experiment rated Chinese words for perceived power, familiarity, and valence on a five-point Likert scale. Some of the words had to be changed prior to testing due to low familiarity rating or ambiguous perceived power in the norming, leading to six high-power and three low-power words being replaced. As expected, the perceived power of the high-power words was significantly higher (M = 3.69 ± 0.74) than that of low-power words (M = 1.47 ± 0.18), $t(49) = 15.2$, $p < .001$, Cohen's d = 4.26. Valence ratings for high-power words were significantly higher (M = 3.62 ± 0.44) than for low-power words (M = 2.51 ± 0.61), $t(49) = 7.32$, $p < .001$, Cohen's d = 2.05. Familiarity did not differ significantly between high-power (M = 4.25 ± 0.42) and low-power words (M = 4.19 ± 0.44), $t(49) = .64$, *Cohen's d* = .15. Lexical frequency was controlled using log transformed values from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). No significant difference was found between high-power (M = 2.45 ± 0.53), and low-power words (M = 2.31 ± 0.55), $t(47) = 1.3$, $p = .361$, *Cohen's d* = .26.

Another 30 volunteers rated relation-to-space, familiarity, and valence of words referring to high and low spatial references on a seven-point Likert scale. High-location words were rated as referring to significantly higher location (M = 6.15 ± 0.71) than low-location words (M = 2.93 ± 0.35), $t(28) = 15.75$, $p < .001$, *Cohen's d* = 5.75. As was the case for words associated with perceived power, valence ratings for high-location words were significantly more positive (M = 4.54 ± 0.71) than low-location words (M = 3.71 ± 0.35), $t(28) = 4.08$, $p < .001$, *Cohen's d* = 1.49. No difference was found in terms of familiarity (words with higher spatial location: M = 6.22 ± 0.37; words with lower spatial location: M = 5.94 ± 0.44, $t(28) = 1.89$, $p = .07$, *Cohen's d* = .69), or lexical frequency (words with higher spatial location: M = 2.58 ± 0.42; words with lower spatial location: M = 2.48 ± 0.49, $t(28) = .61$, $p = .55$, *Cohen's d* = .22).

Volunteers who normed the Chinese stimuli also rated English translated equivalents for perceived power and familiarity. The perceived power of the high-power words was significantly higher (M = 3.62 ± 0.59) than that of low-power words (M = 1.45 ± 0.17), $t(49) =18.57$, $p < .001$, *Cohen's d* = 5.21. There was no significant difference in terms of familiarity (high-power words: M = 4.32 ± 0.45; low-power words: M = 4.25 ± 0.46, $t(49) = 0.54$, $p = .59$, *Cohen's d* = .15). Lexical frequency and valence were controlled using values exported from the LexOPS database (Taylor et al., 2020). High-power words had a significantly greater (M = 4.52 ± 0.49) lexical frequency than low-power words (M = 3.88 ± 0.76), $t(56) = 3.81$, $p < .001$, *Cohen's d* = 1.00). There was no significant difference in valence ratings between high-power words (M = 5.28 ± 1.09) and low-power words (M = 5.10 ± 1.45), $t(56) = .55$, $p = .59$, *Cohen's d* = .14.

As regards words referring to high/low spatial locations, there was no significant difference in familiarity (high-location words: M = 6 ± 0.52; low-location words: M = 5.77 ± 0.62, $t(23) = 1.01$, $p = .32$, *Cohen's d* = .41), or lexical frequency (high-location words: M = 4.49 ± 0.5; low-location words: M = 4.43 ± 0.52, $t(28) = .32$, $p = 0.75$, *Cohen's d* = .12). However, valence ratings for high-location were significantly more positive (M = 6.17 ± 0.9) than for low-location words (M = 5.07 ± 1.29), $t(28) = 2.69$, $p =.01$, *Cohen's d* =.98.

Finally, eight Chinese words and eight English words which were not included in the main experimental blocks were used as practice items.

Procedure

After signing the consent form, participants sat in the center of a dimly light and quiet testing booth, 75 cm away from a 19-inch CRT monitor. One speaker was located above participants' head and two speakers were set on the ground on either side of their seat (to simulate a sound origin from below). The distance between speakers and participants ears varied between 1.1–1.4 m depending on participants' height (it was not technically feasible to change the speakers position relative to the participant's head, but we assume that resulting variations in volume would be negligible). Participants were then asked to complete the LEAP-Q questionnaire.

The information sheet presented the study as being focused on understanding differences in the way Chinese-English bilinguals and English native speakers process language. Participants were instructed to report the origin of the stimulus as being above or below their sitting position when it did not refer to an animal by pressing designated buttons on a serial response box. No go trials featuring an animal name were filler trials. Perceived power or location attributes were not mentioned in the instructions. Response buttons arranged along a front-back axis representing the vertical space were counterbalanced between participants. At the end of the experiment, participants were asked why they thought we used loudspeakers at different locations[1] and only one participant reported a potential association between the location of the speakers and the metaphorical meaning of the stimuli. The data from this participant were excluded from analysis.

Chinese-English participants performed the experiment once in Chinese and once in English with order counterbalanced between languages. Stimuli were presented over four blocks (two in Chinese and two in English) preceded by eight practice trials. Each word was presented only once per block, either from above or from below, that is either in a congruent or an incongruent configuration (90 experimental trials and 30 filler trials per block, adding up to 480 trials in total). Native participants performed the experiment only in English. Stimuli were presented once either congruently or incongruently in two blocks. 240 trials were used overall. In the congruent condition, high-power and high-location words (e.g., king and sun) were played from above, and low-power words and low-location words (e.g., servant and ground) were presented from below. Sound source was reversed in the incongruent condition.

The experiment was programmed using E-Prime 2.0 (Psychology Software Tools, Inc., Pittsburgh, PA). Trials started with a fixation cross displayed in the center of the screen (Fig. 2). After a random duration of 450–650 ms (in steps of 20 ms), the auditory stimulus was presented via the speaker located above or those located below until participant's response was recorded within a 3000 ms time window. Participants were asked to respond only after the auditory file had finished playing. Upon responding, they received the fixation turned blue to provide feedback, and after each 8–10 trials, the fixation turned pink for 2000 ms, giving them time to blink should they need to. They were encouraged to blink during the customized 2000 ms pauses to minimize eye blinks and fatigue-related artefacts.

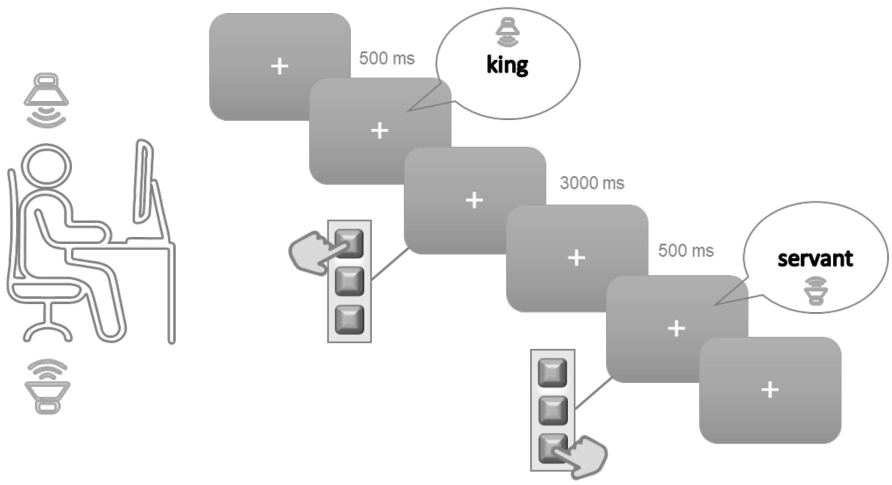

Fig. 2 Experimental setup and time sequence of stimulus presentation

Data Analysis

Behavioral Data Analysis

Response times (RTs) were extracted for correct responses only, that is excluding 7.7% error trials. We also excluded trials with RTs shorter than 200 ms or greater than 2500 ms (0.1%)

as *a priori* improbable values for valid measurements. Then we applied a log transformation after a Box-Cox test to recover a normal distribution for RTs. Trials with log-transformed RTs more than ±2.5 standard deviations from each participant's condition mean were excluded as outliers (1.5%). RTs were then analyzed using linear mixed effects regression (*lmer* function) using the R (R Development Core Team, 2022) package lme4 (Bates et al., 2015). As planned in our pre-registration, RTs were modelled as a function of three fixed effects predictors, Language (Chinese, English), Congruency (congruent, incongruent), and Relation to Space (direct, metaphorical), centered to minimize collinearity. Participants had 43 ± 15 trials per condition on average remaining after data cleaning. Accuracy data were submitted to logistic mixed-effects regression using the *glmer* function of lme4 with a binominal link function as a function of three fixed effects predictors, Language (Chinese, English), Congruency (congruent, incongruent), and Relation to Space (direct, metaphorical). All models included random intercepts for subjects and items and maximal random slopes for each within-subjects and within-items predictors including main effects and interactions, respectively (Barr et al., 2013). Fixed and random effects and interactions that did not significantly contribute to model fit were systematically removed. We have reported unstandardized beta estimates as estimated with the lme4 package (Bates et al., 2015), using β to indicate coefficients. The p values were obtained using the lmerTest package (Kuznetsova et al., 2017) based on Satterthwaite's approximations. An alpha level of .05 were used to establish statistical significance. Pairwise comparisons were performed using the emmeans package (Lenth, 2020) in R to compute the Bonferroni test to correct for multiple comparison in all cases.

EEG Data Recording and Analysis

Electrophysiological data were recorded at a rate of 1 kHz from 64 active Ag/AgCl electrodes placed according to the extended 10-20 convention and referenced to electrode Cz.

Impedances were kept below 7 kΩ. The electroencephalogram (EEG) was filtered using an online bandpass filter (0.05–200 Hz).

EEG data pre-processing was performed offline with scripts in MATLAB (R2021b, The MathWorks, Inc., Natick, MA), using functions included in EEGLAB v2022.0 (Delorme & Makeig, 2004) and ERPLAB v9.0 (Lopez-Calderon & Luck, 2014). The data were resampled to 250 Hz and filtered with separate high-pass and low-pass finite impulse response (FIR) filters (passband edges: 0.1 and 20 Hz, respectively). Line noise at ~50 Hz and harmonics was corrected using the Cleanline (v2.0) procedure before being re-referenced to the global average reference. Ocular artefact correction was performed using Independent Component Analysis (ICA, Stone, 2002) focusing on visual inspection of components associated with blinks and eye movements (ICLabel v1.4). Data were then segmented into epochs ranging from -200–1000 ms from stimulus onset, and baseline correction was applied relative to the EEG signal between -200–0 ms. Epochs containing activity exceeding ± 150 μV at any electrode site apart from the electrooculogram channels within each epoch window were discarded, resulting in an average of 35 ± 8 trials per condition in each participant.

Mean amplitudes of the P3 were computed between 250–350 ms after stimulus onset at 6 electrodes (i.e., CP1, CP2, CPz, P1, Pz, P2;Polich, 2007) and mean amplitudes of the N400 were computed between 350–500 ms after stimulus onset at 9 electrodes of predicted maximal sensitivity (i.e., FC1, FCz, FC2, C1, Cz, C2, CP1, CP2, CPz; Kutas et al., 1984; Kutas & Hillyard, 1980; Schendan & Kutas, 2007). Mean P3 and N400 amplitudes were analyzed by means of a 2 (Language: Chinese, English) × 2 (Relation-to-Space: direct, metaphorical) × 2 (Congruency: congruent, incongruent) repeated-measures Analysis of Variance (ANOVA). We computed Type III ANOVA for main effects and interactions using the *aov_car* function of the Afex package in R (Singmann et al., 2023). A Greenhouse-Geisser correction for non-sphericity was applied when required and Bonferroni correction for multiple comparisons via two-tailed *t*-test was implemented in *post-hoc* pairwise comparisons. Cohen's d was calculated as a measure of effect size. All materials, aggregated results and analysis scripts are available on the OSF server (osf.io/hm2c5).

Results

Results Based on Pre-Registration

Behavioral Results

The linear mixed model that we conducted on RTs revealed a main effect of Language, such that participants responded faster in Chinese than English overall, $\beta = .087$, SE = .025, 95% CI = [0.05, 0.50], $t = 3.513$, $p = .001$. However, they were not faster when responding to congruent than incongruent items, $\beta = .015$, SE = .010, 95% CI = [0.00, 0.03], $t = 1.479$, $p = .140$. Two-way and three interactions involving Congruency as a factor were not significant (*p*s >.1; see Appendix S2 for summary tables).

Participants had a higher accuracy for congruent items than incongruent items, $\beta = -0.384$, SE = .177, 95% CI = [-0.73, -0.03], $z = -2.170$, $p = .030$. They also had fewer errors for direct relationship to space (location words) than metaphorical relationship to space (power words), $\beta = -0.452$, SE = .187, 95% CI = [-0.82, -0.09], $z = -2.424$, $p = .015$. Furthermore, we found a significant interaction between Language and Congruency, $\beta = .897$, SE = .339, 95% CI = [0.23, 1.56], $z = 2.647$, $p = .008$, such that bilingual participants tested in Chinese registered more correct responses to congruent than incongruent items, $\beta = .832$, SE = .241, 95% CI = [0.36, 1.30], $z = 3.459$, $p < .001$, whilst no such difference was found in English, $\beta = -.064$, SE = .249, 95% CI = [-0.55, 0.42], $z = -0.258$, $p = .796$ (Fig. 3a). No other two-way or three-way interactions was significant (*p*s >.1).

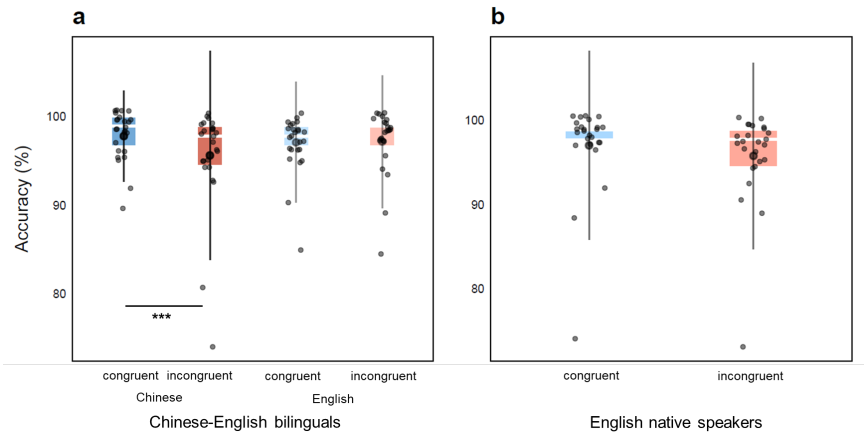

Fig. 3 Accuracy for (a) Language × Congruency interactions in Chinese-English bilinguals and (b) English native speakers. *** p < .001.

We also compared bilinguals with English native participants, with Congruency and Relation-to-Space as within-subject factors and Group (Chinese-English bilinguals, English native speakers) as the between-subjects factor. The logistic mixed effects modelling of accuracy revealed no fixed effect of Congruency, β = .011, SE = .013, 95% CI = [-0.01, 0.04], t = .880, p = .381, or Group, β = -.064, SE = .041, 95% CI = [-0.15, 0.02], t = -1.565, p = 0.124, or Relation-to-Space, β = .004, SE = .014, 95% CI = [-0.02, 0.03], t = .309, p = .756, and no significant two- or three-way interactions (ps > .1; Fig. 3b). In the corresponding accuracy analysis, no effect reached significance (ps > .1; Fig. 3b).

P3 Event-Related Potentials (250–350 ms)

In Chinese-English bilinguals, repeated measures ANOVA on P3 mean amplitudes showed a main effect of Language, F(1,24) = 5.50, p = .028, $\eta_p^2 = .186$ (large effect, according to Cohen, 1988), such that P3 amplitude was significantly more positive in English than in Chinese (Fig. 4). There was no main effect of Congruency, F(1,24) = 1.26, p = .272, $\eta_p^2 =$ .05 (small effect) or Relation-to-Space, F(1,24) = .27, p = .611, $\eta_p^2 = .011$ (small effect) and no two-way or three-way interactions reached significance (ps > .1).

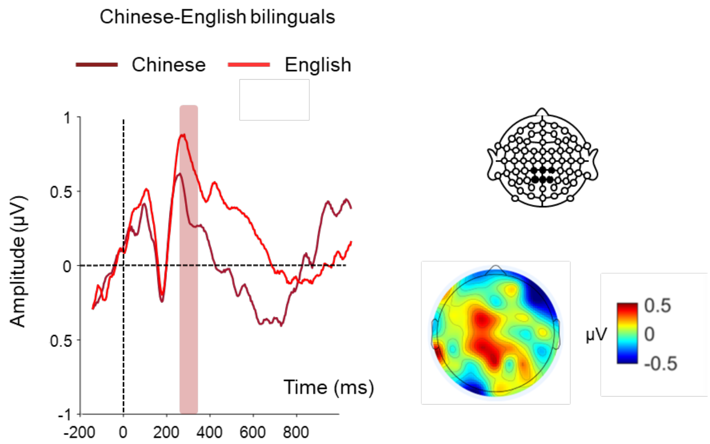

Fig. 4 ERP plots depicting the Language main effect. ERPs are computed from a linear derivation of six electrodes (CP1, CPz, CP2, P1, Pz, P2). The analysis window is highlighted in pink.

The corresponding analyses conducted between groups failed to show a main effect of Group, $F(1,48) = .90$, $p = .349$, $\eta_p^2 = .018$ (small effect), Relation-to-Space, $F(1,48) = .68$, $p = .936$, $\eta_p^2 < .001$ (negligible effect) or Congruency, $F(1,48) = .49$, $p = 0.507$, $\eta_p^2 = .009$ (negligible effect). There was no other two-way or three-way interaction (*p*s>.2).

N400 Event-Related Potentials (350–500 ms)

In Chinese-English bilinguals, the three-way repeated measures ANOVA on N400 mean amplitudes revealed a main effect of Language, $F(1,24) = 9.85$, $p = .004$, $\eta_p^2 = .291$ (large effect), such that N400 amplitude was significantly more negative in Chinese than English (Fig. 5a). There was no main effect of Congruency, $F(1,24) = 1.33$, $p = .259$, $\eta_p^2 = .053$ (small effect), but there was a significant main effect of Relation-to-Space, $F(1,24) = 12.48$, $p = .002$, $\eta_p^2 = .342$ (large effect), such that N400 amplitudes were more negative for direct (location words) than metaphorical (power words) references to space (Fig. 5a).

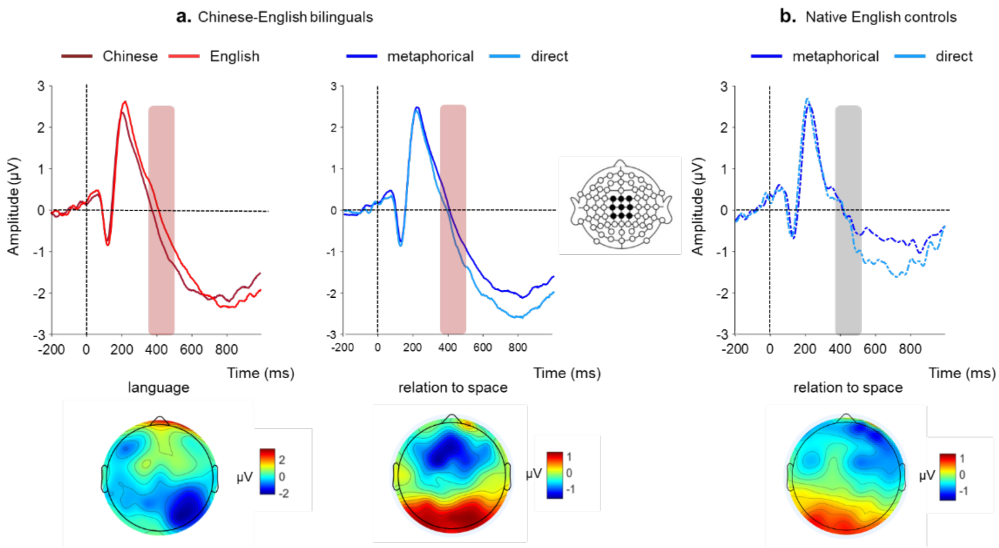

Fig. 5 (a) ERP plots depicting the Language and Relation-to-Space main effects. (b) ERPs elicited by stimuli with a direct and metaphorical reference to space in native English controls for comparison. ERPs were computed from a linear derivation of nine central electrodes (FC1, FCZ, FC2, C1, Cz, C2, CP1, CPz, CP2). The analysis window is highlighted in pink.

The interaction between Language and Relation-to-Space was close to reaching significance, $F(1,24) = 3.79$, $p = .063$, $\eta_p^2 = .136$ (medium effect). We found no interaction between Language and Congruency, $F(1,24) = .75$, $p = .394$, $\eta_p^2 = .03$ (small effect); or Congruency and Relation-to-Space, $F(1,24) = .05$, $p = .823$, $\eta_p^2 = .002$ (negligible effect), and there was no interaction between the three factors, $F(1,24) = .22$, $p = .645$, $\eta_p^2 = .009$ (negligible effect).

We also conducted a between-within repeated measures ANOVA on N400 mean amplitude, with Congruency and Relation-to-Space as within-subjects factors and Group (Chinese-English bilingual, English native speaker) as a between-subjects factor. There was no main effect of Group, $F(1,48) = .01$, $p = .936$, $\eta_p^2 < .001$ (negligible effect), Congruency, $F(1,48) = .86$, $p = .360$, $\eta_p^2 = .018$ (small effect) or Relation-to-Space, $F(1,48) = 3.03$, $p = .088$, $\eta_p^2 = .059$ (medium effect). None of the two-way interactions were significant

(ps >.1), and the three-way interaction was not significant either, $F(1,48) = .06$, $p = .807$, $\eta_p^2 = .001$ (negligible effect).

Summary of pre-registered analyses and interim discussion

Our pre-registered analyses failed to detect any significant effects of semantic/spatial congruency in either the behavioral or the ERP data. One explanation for this lack of effect might relate to the fact that two previous studies reporting power-based congruency effects only detected them for high power words (Zanolie et al., 2012; Wu et al., 2016). Although we did not consider this possibility in our pre-registration, stronger congruency effects could be expected for high-power words, and we will address these further in the general discussion. For now, we consider the possibility that differences between high-power and low-power words may have shadowed the interactions of interest. We therefore provide additional (not pre-registered) analyses that distinguish between high-power and low-power roles. Because this distinction only makes sense for the stimuli with a metaphorical relation to space (i.e., the human words, e.g., king/servant), we restrict our analyses to these items. As in our previous analyses, we started by addressing within-subjects effects in our bilingual group, and then address between-subjects differences by comparing their performance to matched English native controls. Note that these analyses still address our core research question, namely whether processing metaphors of power in relation to space differs between languages in bilinguals.

Exploratory Analyses distinguishing between high- and low-power words

Behavioral Results

Linear mixed effects regressions modelled RTs as a function of three within-subject factors, Language (Chinese, English), Power (high-power, low-power), and Congruency (congruent, incongruent). Participants had 29 ± 1 trials per condition on average remaining after data cleaning. They responded faster when tested in Chinese as compared to English, β = .085, SE = .026, 95% CI = [0.04, 0.44], t = 3.261, p = .002, and faster to high-power words than to low-power words β = .054, SE = .012, 95% CI = [0.02, 0.15], t = 4.393, p < .001. Furthermore, Power interacted with Congruency, β = -0.074, SE = .027, 95% CI = [0.00, 0.14], t = -2.746, p = .007 (Fig. 6a). Pairwise comparison showed longer responses time to incongruent than congruent stimuli for high-power words, β = -0.051, SE = .018, 95% CI = [-0.44, -0.08], t = -2.807, p = .006, but corresponding analyses for low-power words showed no such effect, β = .023, SE = .018, 95% CI [-0.07, 0.29], t = 1.250, p = .211. No other main effects or interactions approached significance (ps > .2).

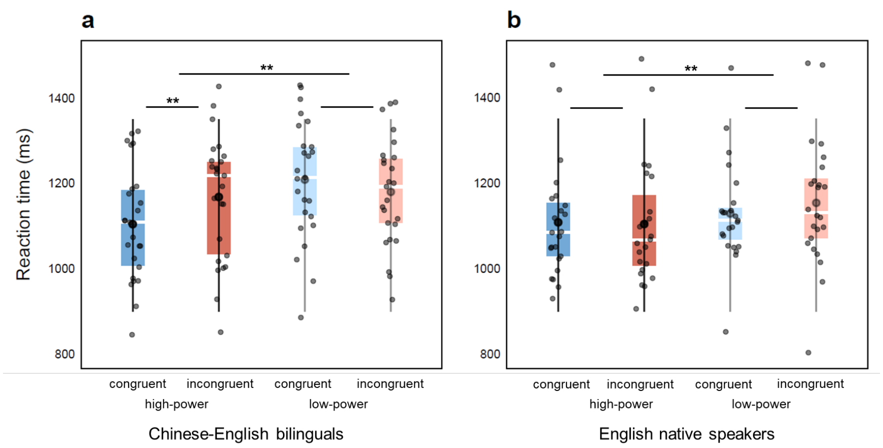

Fig. 6 Individual mean Reaction times results when limited to power words (a) Power × Congruency interaction in Chinese-English bilinguals and (b) English native speakers. ** p < .05.

To exclude the possibility that valence might have driven the Power × Congruency interaction, we entered Valence as another predictor into our linear mixed effect model. The model revealed fixed effects of Language, β = .089, SE = .026, 95% CI = [0.04, 0.45], t = 3.442, p = .001, and Power, β = .064, SE = .019, 95% CI = [0.01, 0.13], t = 3.439, p < .001. We detected no fixed effect of Valence, β = .001, SE = .015, 95% CI = [0.00, 0.01], t = .093, p = .926. Critically, the two-way interaction between Power and Congruency remained significant, β = -0.078, SE = .039, 95% CI = [0.00, 0.08], t = -2.007, p = .046. Pairwise comparison showed that the congruency effect was larger when participants responded to high-power words, *β* = -0.055, SE = .024, 95% CI = [0, 0.10], *z* = -2.276, *p* = .024, than when they responded to low-power words, *β* = .022, SE = .023, 95% CI = [-0.07, 0.03], *z* = .993, *p* = .322. No two-or-three-way interactions involving Valence as a predictor approached significance (*ps* > .8).

Mixed effects logistic regression of accuracy data detected no fixed effects of Language, *β* = .085, SE = .337, 95% CI = [-0.57, 0.75], *z* = .254, *p* = .8; or Power, *β* = -0.117, SE =.195, 95% CI = [-0.50, 0.27], *z* = -0.598, *p* =.550, but bilingual participants had higher accuracy in the congruent than the incongruent condition, *β* = -0.448, SE = .194, 95% CI = [-0.83, -0.07], *z* = -2.302, *p* = .021.We also found a significant interaction between Language and Congruency, *β* = 0.851, SE = .433, 95% CI = [0.00, 1.70], *z* = 1.965, *p* = .049. Planned comparisons showed that when bilingual participants were tested in Chinese, they had less errors for congruent stimuli than the incongruent stimuli, *β* = .873, SE = .303, 95% CI = [0.28, 1.47], *z* = 2.879, *p* = .004, whilst this was not the case when they were tested in English, *β* = .022, SE = .278, 95% CI = [-0.52, 0.57], *z* = .080, *p* = .913(Fig. 7a).

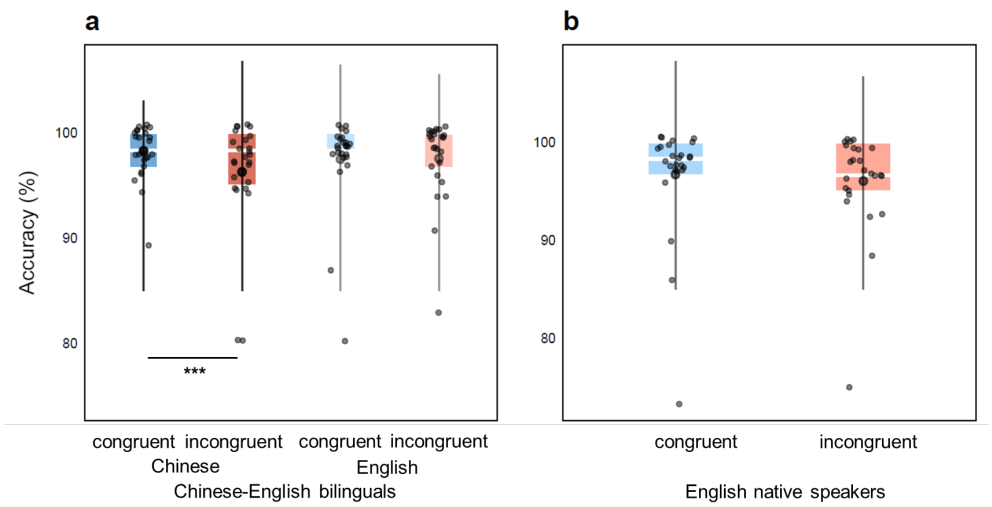

Fig. 7 Accuracy results when limited to power words (a) Language × Congruency interactions in Chinese-English bilinguals and (b) English native speakers. ** *p* < .05.

Moreover, the analysis revealed a significant interaction between Power and Congruency, *β* = 1.567, SE = .491, 95% CI = [0.60, 2.53], *z* = 3.192, *p* = .001 (Fig. 8a). Pairwise analysis indicated that when bilingual participants responded to high-power words, they had more errors for incongruent than congruent stimuli, *β* = 1.231, SE = .331, 95% CI = [0.58, 1.88], *z* = 3.718, *p* < .001. No such effect was found when participants responded to low-power words, *β* = -.336, SE = .294, 95% CI = [-0.91, 0.24], *z* = -1.142, *p* = .254. There was no interaction between Language and Power, *β* = .178, SE = .442, 95% CI = [-0.69, 1.04], *z* = .402, *p* = .688, and the interaction between three factors was not significant either, *β* = -1.087, SE = .873, 95% CI = [-2.80, 0.62], *z* = -1.245, *p* = .213.

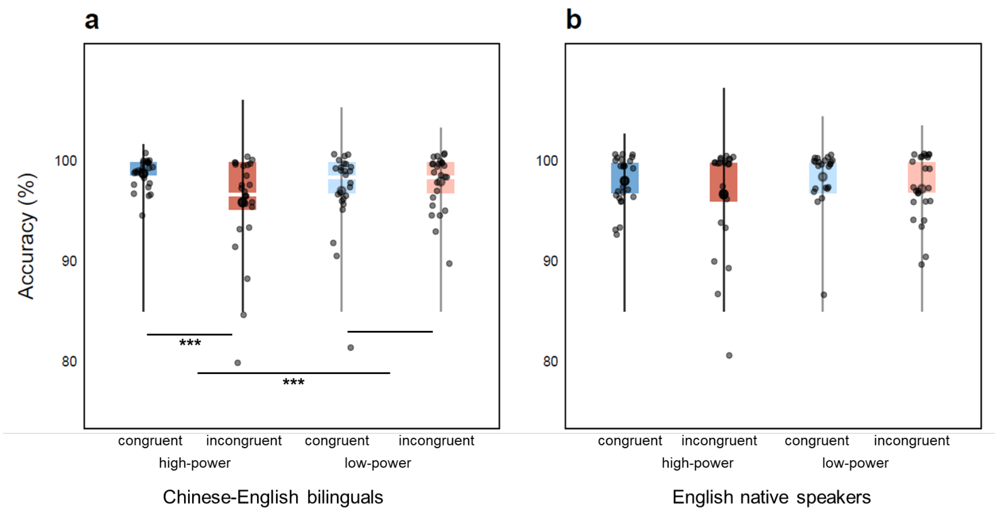

Fig. 8 Accuracy results when limited to power words : (a) Power × Congruency interaction in Chinese-English bilinguals and (b) English native speakers. *** *p* < .001.

When compared with English native controls, the lineal mixed model on RTs revealed a main effect of Power, *β* = .055, SE = .016, 95% CI = [0.02, 0.09], *t* = 3.388, *p* < .001, such that both groups of participants responded significantly faster to high than low-power words (Fig. 8b). Linear mixed effects regressions detected no fixed effects of Group, *β* = -.061, SE = .041, 95% CI = [-0.14, 0.02], *t* = -1.497, *p* = .141, or Congruency, *β* = .009, SE = .016, 95% CI = [-0.02, 0.04], *t* = .571, *p* = .569, or two-way interactions involving Congruency as a factor (*p*s > .1), but a significant interaction between the three factors emerged, β = .076, SE = .024, 95% CI = [0.03, 0.12], *t* = 3.211, *p* = .002. Bonferroni corrected multiple comparisons showed that when Chinese-English bilinguals were tested in English, they had significantly faster RTs for high-power words in a congruent as compared to an incongruent condition, β = -0.051, SE = .025, 95% CI = [-0.49, -0.01], *t* = -2.076, *p* = .040, whereas no such effect was found when participants responded to low-power words, β = .020, SE = .024, 95% CI = [-0.14, 0.34], *t* = .821, *p* = .413. There were no such differences in English native controls (*p*s >.1). In the corresponding accuracy analysis, no main effect or interaction were significant (*p*s > .1).

P3 Event-related Potentials (250–350 ms)

In the exploratory analysis with Power as a factor, ERP amplitudes were analyzed by means of a 2 (Language: Chinese, English) × 2 (Power: high-power, low-power) × 2 (Congruency: congruent, incongruent) repeated-measures ANOVA.

This ANOVA showed no main effect of Language, *F*(1,24) = 1.8, *p* = .192, $\eta_p^2 = .07$ (medium effect), Power, *F*(1,24) = 3.19, *p* = .087, $\eta_p^2 = .117$ (medium effect) or Congruency, *F*(1,24) = 1.64, *p* = .213, $\eta_p^2 = .064$ (medium effect). There was no two-way interaction (*p*s > .1), but we found a significant three-way interaction between Language, Power, and Congruency, *F*(1,24) =5.84, *p* = .024, $\eta_p^2 = .196$ (large effect). Bonferroni corrected paired sample *t*-tests showed that for high-power words, ERP amplitude were significantly more positive in the congruent than the incongruent condition when bilingual participants were tested in Chinese, *t*(24) = 2.673, *p* = .013, Cohen's *d* = 0.44 (small effect) whereas no such difference was found when they were tested in English, *t*(24) = -0.767, *p* = .451, Cohen's *d* = -0.06 (negligible effect) (Fig. 9). We did not find such difference for the low-power words (*p*s > .5).

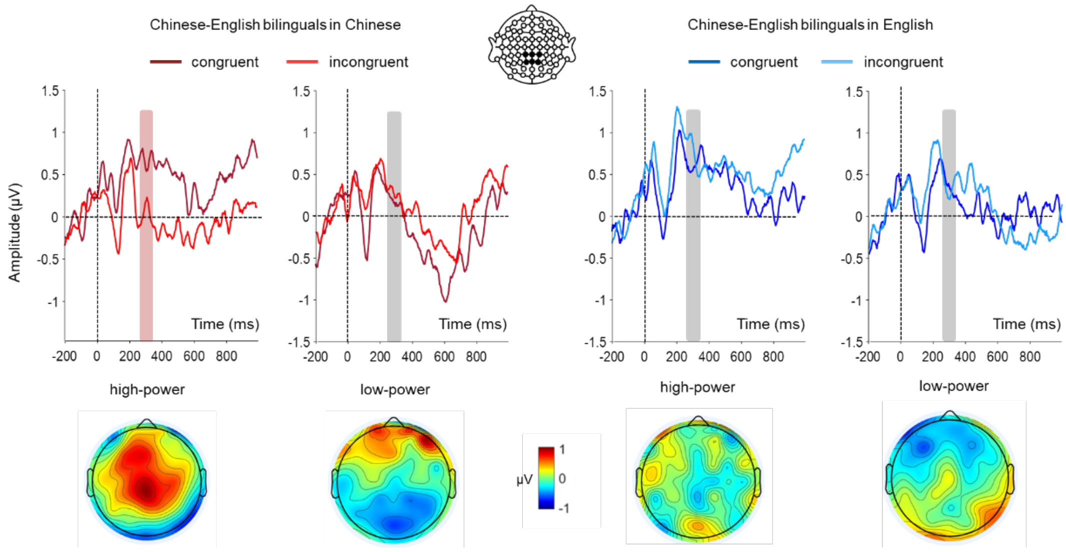

Fig. 9 ERPs plots showing the Language × Power × Congruency interaction in Chinese-English bilinguals. ERP waves depict potential variation from a linear derivation of six electrodes (CP1, CPz, CP2, P1, Pz, P2). The analysis window is shaded in pink to highlight significant differences and framed in gray when differences are not significant.

The between-within repeated measures ANOVA on P3 mean amplitude revealed no main effect of Group, $F(1,48) = .74$, $p = .393$, $\eta_p^2 = .015$ (small effect) or Congruency, $F(1,48) = .01$, $p = .932$, $\eta_p^2 < .001$ (negligible effect) but showed a marginal effect of Power, $F(1,48) = 3.94$, $p = .053$, $\eta_p^2 = .076$ (medium effect). The interaction between Power and Congruency was not significant over Group, $F(1,48) = 3.24$, $p = .078$, $\eta_p^2 = .063$ (medium effect) and there were no two-way or three-way interactions (*p*s > .2).

N400 Event-related Potentials (350–500 ms)

The within-subject repeated measures ANOVA data revealed a marginal effect of Language on mean N400 amplitude in bilingual participants, $F(1,24) = 4.10$, $p = .054$, $\eta_p^2 = .146$ (large effect), such that N400 amplitudes were more negative when participants were tested in Chinese as compared to English. There was no main effect of Power, $F(1,24) = .60$, $p = .444$, $\eta_p^2 = .025$ (small effect) or Congruency, $F(1,24) = .23$, $p = .636$, $\eta_p^2 = .01$ (small effect).

There was no interaction between Language and Power, $F(1,24) = 1.91$, $p = .180$, $\eta_p^2 = .074$ (medium effect) or between Language and Congruency, $F(1,24) = 1.06$, $p = .313$, $\eta_p^2 = .042$ (small effect) but the interaction between Power and Congruency was close to reaching significance, $F(1,24) = 3.75$, $p = .065$, $\eta_p^2 = .135$ (medium effect). Critically, we found a significant three-way interaction between Language, Power and Congruency, $F(1,24) = 5.37$, $p = .029$, $\eta_p^2 = .183$ (large effect).

To unpack this interaction, we conducted a 2 (Language) × 2 (Congruency) ANOVA separately for high-power words. We found a marginal effect of Congruency, $F(1,24) = 3.91$, $p = .059$, $\eta_p^2 = .140$ (large effect), and a significant interaction between Language and Congruency, $F(1,24) = 4.62$, $p = .042$, $\eta_p^2 = .161$ (large effect). Pairwise comparisons corrected for multiple comparison showed that mean ERP amplitude were significantly more negative in the incongruent than the congruent condition when bilingual participants were tested in Chinese, $t(24) = 3.009$, $p = .006$, Cohen's $d = 0.45$ (small effect) whereas no such difference was found when they were tested in English, $t(24) = -.406$, $p = .689$, Cohen's $d = -0.12$ (negligible effect) (Fig. 10a).

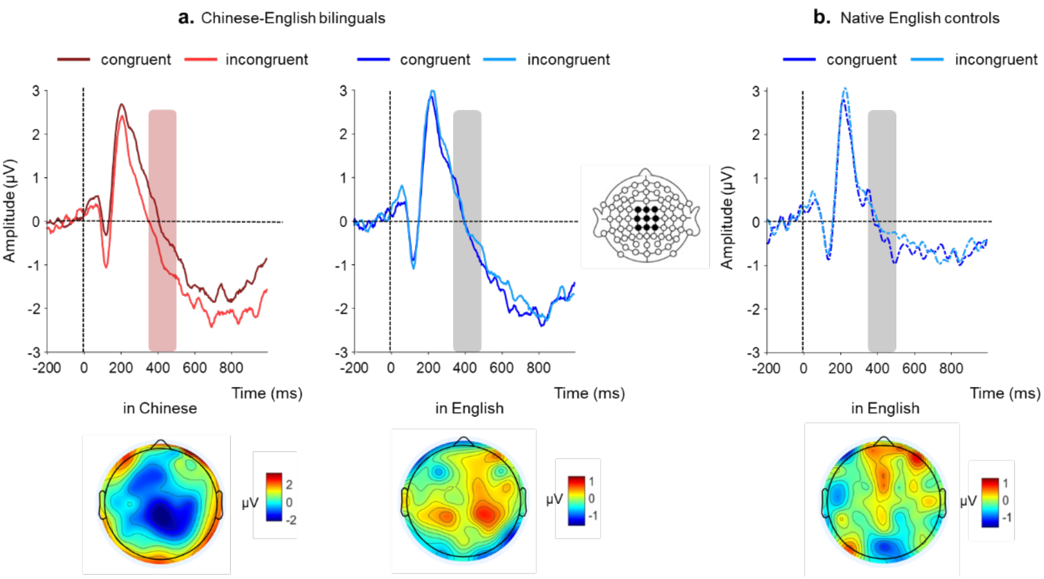

Fig. 10 (a) ERPs plots showing the Language × Congruency interaction in the follow-up ANOVA conducted on power words only in Chinese-English bilinguals; (b) For visual comparison, ERPs elicited in native English controls by congruent and incongruent power words for comparison. ERP waves depict potential variation from a linear derivation of nine central electrodes (FC1, FCZ, FC2, C1, Cz, C2, CP1, CPz, CP2). The analysis window is shaded in pink.

The corresponding analyses conducted between groups failed to show a main effect of Group, $F(1,24) = .00$, $p = .960$, $\eta_p^2 < .001$ (negligible effect), Power, $F(1,24) = 2.25$, $p = .140$, $\eta_p^2 = .045$ (small effect) or Congruency, $F(1,24) = 2.52$, $p = .119$, $\eta_p^2 = .050$ (small effect) (Fig. 10b). There was no other two-way or three-way interaction (*p*s>.1).

Discussion

The present study compared manifestations of spatial embodiment of abstract concepts in Chinese-English bilinguals and native English speakers using ERPs. More specifically, we tested whether patterns of P3 and N400 modulations elicited by conceptual processing of perceived power differently rely on embodied representations in the two languages of bilinguals.

Behavioral data failed to show the expected main effect of congruency between sound origin and metaphorical relation to space of perceived power. Bilingual participants responded faster in Chinese than in English overall for words either with a direct (location words) or metaphorical (power words) relation to space. They made fewer errors when responding to location words than power words. Furthermore, bilingual participants tested in Chinese registered more correct responses to congruent than incongruent trials, whereas no such difference was found in English. ERP results based on pre-registered analyses also failed to show the expected congruency effect in either the P3 or N400 range. However, exploratory analyses restricted to power words and excluding location words showed a congruency effect on RTs for high-power words, but not for low-power words. Participants made less errors for congruent than incongruent stimuli when responding to high- as compared to low-power words and when they were tested in Chinese relative to English (no effect was found in English native controls). Critically, analyses focusing on high-power words showed that bilinguals elicited the originally predicted pattern of response: increased P3 amplitudes in the congruent than incongruent condition, and more negative N400 amplitudes in the reverse comparison. This, however, applied when they were tested in Chinese, not in English.

This study attempted to test embodiment theory based on a direct mapping using auditory stimuli presented in space around participants. Bilingual participants made less errors in the congruent than incongruent condition when tested in Chinese, but not when tested in English. In contrast, participants in Dudschig et al. (2014) made an upward or downward movement to indicate the color of words (e.g., bird, shoes) implicitly associated with spatial locations. They found a congruency effect in both German (L1) and English (L2), but the effect was stronger in L1 than in L2. Differences between results may stem from the fact that such mapping may have been too "direct" for participants to map sound origin with visual semantic properties of words across modalities. For instance, because "sun" and "ground" are not spontaneously associated with sounds, they can be considered visual references, and, thus, it is not clear why such concepts should be associated with above and below location in terms of sound origin. Therefore, in hindsight, our attempt to map concrete references to the perceptual world may have failed due to a lack of "natural" mapping of location across sensory modalities (visual-auditory). As regards the processing of perceived power, we further contend that grouping high and low-power words together may have masked the congruency effect given hints in previous studies that high-power words are more effective stimuli to elicit metaphorical spatial references (Schubert, 2005; Zanolie et al., 2012).

In the exploratory analysis, we found that participants were slower responding to low- than high-power words, which is consistent with previous studies (Schubert, 2005; Wu et al., 2016; Zanolie et al., 2012). For instance, Schubert (2005) presented power words either at the top or at the bottom of the screen and instructed participants to decide whether the word referred to an individual perceived as powerful or powerless. Participants were slower when responding to low-power words when compared to high-power words, as interpreted as a general tendency to look for powerful entities in the environment.

Participants also responded significantly faster and more accurately in congruent than incongruent trials for high-power words but not for low-power words. Arguably, these results could reflect an effect of valence in stimulus selection, with high-power words having more positive valence than low-power words. Valence has been associated with metaphorical spatial reference previously. Using the same color Stroop task as above, Dudschig et al., (2015) showed that emotional states associated with upright or downward bodily experiences (e.g., happy – upwards; depression – downwards) directly affect motor responses along the vertical axis. Participants responded significantly faster when their movement was congruent with the spatial orientation associated with emotion words. This being said, in our experiment, the interaction between Power and Congruency remained significant after adding Valence as a predictor in linear mixed effects modelling, meaning that valence on its own could not account for metaphorical spatial mapping.

Participants registered less errors for congruent than incongruent stimuli when tested in Chinese. This is compatible with results from behavioral studies suggesting that metaphorical mapping works differently in the two languages of bilinguals (Hu, 2021; but see

Wang, 2016; Bai & He, 2022). Wang (2016) found stronger congruency effects in L1 than in L2, which were again stronger in bilinguals with a higher proficiency in L2. We found no congruency effect in L2, however. It should be noted that Wang 1) presented stimuli higher or lower on the screen, which potentially lead to a shift in spatial attention; and 2) used an overt power judgement task, making it more likely that participants became aware of the aim of the spatial manipulation to some extent, whereas in our study participants reported the source of the sound in space and did not make explicit judgements about perceived power.

The ERP analyses focused on P3 and N400 instead of initial components such as P1 and N1 because we aimed to investigate the processing of perceived power at a semantic level rather than a perceptual-attentional one. In addition, early components in previous studies may have been affected by low-level stimulus characteristics such as position on the screen, which is likely to have triggered eye movement artefacts and externally driven –as opposed to endogenously generated– modulations of brain responses. However, this is not to say that early components recorded in the current study were necessarily unaffected by low-level properties of stimuli, namely sound origin, and differences in intensity between sounds coming from above and below the participant’s sitting position. Indeed, in our study, sound level was not perfectly balanced between lower and upper source for technical reasons discussed below under limitations.

When high-power words were analyzed separately, we found the predicted congruency effect on P3 and N400 amplitudes. This variation was found in Chinese but not English in bilinguals, and not found at all in English controls. This suggests that metaphorical mapping for perceived power differs across languages in bilinguals and that it is weaker or possibly absent in native speaker of English. The P3 component is known to reflect orientation of attention, stimulus evaluation, and target detection (Polich, 2007). Due to the unbalanced proportion of power, location, and filler words in the experiment (only one fourth of stimuli were effectively congruent), participants likely detected congruent stimuli as standing out. Furthermore, participants will have found it easier and less cognitively taxing to process high-power words played from above than below, leading to N400 amplitude reduction. Indeed, the N400 is a well-established index of semantic processing, which has shown similar modulation in spatiotemporal metaphor processing (Li et al., 2019).

We thus found that manifestations of embodiment of an abstract concept varies across languages of bilinguals. This could be due to relatively lower proficiency in L2, associated with limited spreading of activation from L2 words. However, the fact that we did not find embodiment effects in our native group of English suggests lower general reliance on embodiment in English. This may have a conceptual origin or illustrate cultural differences between the two languages.

However, our findings are also partly inconsistent with previous studies (Dudschig et al., 2014; Vukovic & Shtyrov, 2014; Wang, 2016), suggesting that processing abstract concepts is grounded in sensorimotor representations, albeit to a weaker extent. For instance, Vukovic and Shtyrov (2014), investigating motor cortex activity in German-English bilinguals as they responded to L1 and L2 abstract and action prime-probe verb pairs, found motor activations in L1 (stronger) and L2 (weaker). They concluded that embodied cognition applies to both concrete and abstract words in both languages of bilinguals. Note that participants in this study were late but highly proficient L2 English speakers, which may explain why their L2 successfully triggered access to semantic (and therefore embodied) representations (see also Kroll & Stewart, 1994). Kroll & Stewart (1994) proposed that lexical and the conceptual are two separate levels of representation, the latter being shared by both languages. As L2 proficiency increases, the link between L2 and conceptual representation strengthens. Since our participants reported their English proficiency as upper intermediate, the connection between L2 words and their conceptual representation may have been too weak to prompt full access to embodied representations.

Another possible interpretation of the observed pattern of results could relate to the implicit nature of the task used: Participants were asked to report the source of a sound rather than make deliberate judgements about perceived power. This potentially reduced the connection between words and conceptual representations when tested in L2. Indeed, at debriefing, participants reported no explicit knowledge of hidden manipulations between the sound origin and word meaning or awareness of the purpose of this study (with the exception of one participant, excluded). This is partly consistent with results from previous studies suggesting that the L2 sensorimotor activation depends on task demands and depth of semantic processing (Bai & He, 2022; Vukovic et al., 2017). For example, Bai and He (2022) found a congruency effect only in late L2 learners performing a semantic categorization task, but not in a lexical decision task, suggesting that motor circuit recruitment in low proficiency bilinguals depends upon semantic task demands.

It remains to be discussed why we observed P3 and N400 amplitude modulations for high- but not low-power words. One possible explanation for such asymmetric embodiment of perceived power is that high-power words entertain stronger metaphorical association with higher positions (i.e., “king” – above) than low-power words with lower positions (i.e., “servant” – below). Our participants were university students who likely perceived professors or supervisors as having a higher status than themselves whereas employees or interns would not be considered as having a lower social status. It remains, however, that the congruency effect found for high-power words disappeared when bilingual participants were tested in L2 and that it was not found in control participants tested in their L1 English. It is thus possible that the asymmetry between high and low-power words is supplemented by cultural differences. However, evidence regarding potential variations between cultures concerning perceived power is inconsistent (Bond & Hwang, 1986; Yang et al., 2021), and thus it is not possible at this point to determine the origin of differences in embodiment manifestation across languages.

Limitations and Future Directions

Our study has several limitations that should be addressed in the future. The number of stimuli was not balanced across experimental and control conditions (60 power words vs. 30 location words). Whilst having the same number of stimuli in all conditions would be ideal, identifying location words that were clearly associated with higher and lower positions proved to be very challenging, especially for the lower positions. But as discussed above, the lack of a congruency effect for location words might also be attributed to the indirect mapping between the visual and auditory domain concerning spatial localization.

Secondly, the sound amplitude was not perfectly matched between lower and upper sound sources, potentially increasing the noise in the measurements, and rendering early ERP components less reliable. Indeed, due to the inability to position a speaker underneath the participants' chair to provide clear audio output from below, we resorted to employing two speakers on either side of the participants' chair to simulate sound from below. Despite efforts to minimize volume disparities between higher and lower positions by muffling the floor speakers, some differences remained. Additionally, we were unable to adjust the system's position according to each participants' height, introducing another albeit weak source of variance.

Although we recruited Chinese-English bilinguals with an IELTS score of 6 or higher and assessed their fluency in the native and second language based on self-reports (LEAP-Q), fluency measures may have lacked precision, since the timing of IELTS testing was not controlled. Note however that fluency in English would only have improved after our bilingual participants started their studies in the UK. In addition, it is noteworthy that self-report measures such as those obtained using the LEAP-Q questionnaire are considered a reliable and valid indicator of language ability that reflect competency measured by other means (Kaushanskaya et al., 2020; Marian et al., 2007; Marian & Hayakawa, 2021). Also, and finally, despite the sample size (25) being on the upper end of the range used in previous, comparable ERP studies (e.g., Li et al., 2019; 2022; Wu et al., 2016; Zanolie et al., 2011), we acknowledge that statistical power would be improved with a larger participant sample, especially when looking at higher order interactions.

Future studies could address the limitations listed above and use electrophysiological or neuroimaging methods to explore how linguistic experience (e.g., age of acquisition, L2 proficiency, L2 acquisition stages) and task demands influence L2 embodiment of abstract processing and how such effects can be distinguished from cultural effects.

Conclusion

Altogether, we found that the embodiment of perceived power – a relatively abstract concept, is asymmetric for high and low-power words and differs across languages in bilinguals. We found embodiment for high-power L1 words in Chinese-English bilinguals but failed to observe such effect in the L2, and such embodiment effect was undetectable in control speakers of English tested in the L1. These results point to either a difference in embodiment representation across languages in bilinguals, a difference in embodiment across cultures, or both. The current dataset, however, does not allow us to determine which factor is driving the differences. In any case, we contend that weaker embodiment in L2 is linked to L2 proficiency in bilinguals. Our study provides the first electrophysiological evidence for the involvement of direct sensorial (auditory) representation in the spatial mapping of perceived power in bilinguals as an index of embodiment.

Notes

1 Participants' answers for post-experiment questions can be found on OFS: osf.io/hm2c5.

References

Ahlberg, D. K., Bischoff, H., Kaup, B., Bryant, D., & Strozyk, J. V. (2018). Grounded cognition: Comparing Language × Space interactions in first language and second language. *Applied Psycholinguistics*, **39**(2), 437–459. https://doi.org/10.1017/S014271641700042X

Aravena, P., Hurtado, E., Riveros, R., Cardona, J. F., Manes, F., & Ibáñez, A. (2010). Applauding with closed hands: Neural signature of action-sentence compatibility effects. *PLoS ONE*, **5**(7). https://doi.org/10.1371/journal.pone.0011751

Bai, Y., & He, W. (2022). Involvement of the sensorimotor system in less advanced L2 processing: Evidence from a semantic category decision task. *Frontiers in Psychology*, **13**. https://doi.org/10.3389/fpsyg.2022.980967

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

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, **22**(4), 577–609. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, **59**, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **373**(1752). https://doi.org/10.1098/rstb.2017.0144

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In Manuel de Vega, Arthur Glenberg, and Arthur Graesser (eds), *Symbols and Embodiment Debates on meaning and cognition* (pp. 245–284). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0013

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 129–163). Cambridge University Press. https://doi.org/10.1017/CBO9780511499968.007

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1– 48. https://doi.org/10.18637/jss.v067.i01

Bergen, B., Lau, T. T. C., Narayan, S., Sto Janovic, D., & Wheeler, K. (2010). Body part representations in verbal semantics. *Memory and Cognition*, **38**(7), 969–981. https://doi.org/10.3758/MC.38.7.969

Birba, A., Beltrán, D., Martorell Caro, M., Trevisan, P., Kogan, B., Sedeño, L., Ibáñez, A., & García, A. M. (2020). Motor-system dynamics during naturalistic reading of action narratives in first and second language. *NeuroImage*, **216**. https://doi.org/10.1016/j.neuroimage.2020.116820

Bond, M. H., & Hwang, K. (1986). The social psychology of Chinese people. In *The psychology of the Chinese people.* (pp. 213–266). Oxford University Press.

Borghi, A. M. (2020). A future of words: Language and the challenge of abstract concepts. *Journal of Cognition*, **3**(1): 42. https://doi.org/10.5334/joc.134

Borghi, A. M., Shaki, S., & Fischer, M. H. (2022). Abstract concepts: external influences, internal constraints, and methodological issues. *Psychological Research*, **86**(8), 2370–2388. https://doi.org/10.1007/s00426-022-01698-4

Buccino, G., Colagè, I., Gobbi, N., & Bonaccorso, G. (2016). Grounding meaning in experience: A broad perspective on embodied language. *Neuroscience and Biobehavioral Reviews*, **69**, 69–78. https://doi.org/10.1016/j.neubiorev.2016.07.033

Buccino, G., Marino, B. F., Bulgarelli, C., & Mezzadri, M. (2017). Fluent speakers of a second language process graspable nouns expressed in L2 like in their native language. *Frontiers in Psychology*, **8**. https://doi.org/10.3389/fpsyg.2017.01306

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE*, **5**(6), Article e10729. https://doi.org/10.1371/journal.pone.0010729

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, **106**(2), 579–593. https://doi.org/10.1016/j.cognition.2007.03.004

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

de Grauwe, S., Willems, R. M., Rueschemeyer, S. A., Lemhöfer, K., & Schriefers, H. (2014). Embodied language in first- and second-language speakers: Neural correlates of processing motor verbs. *Neuropsychologia*, **56**(1), 334–349. https://doi.org/10.1016/j.neuropsychologia.2014.02.003

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, **134**(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dudschig, C., de la Vega, I., & Kaup, B. (2014). Embodiment and second-language: Automatic activation of motor responses during processing spatially associated L2 words and emotion L2 words in a vertical Stroop paradigm. *Brain and Language*, **132**, 14–21. https://doi.org/10.1016/j.bandl.2014.02.002

Foroni, F. (2015). Do we embody second language? Evidence for "partial" simulation during processing of a second language. *Brain and Cognition*, **99**, 8–16. https://doi.org/10.1016/j.bandc.2015.06.006

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, **22**(3), 455–479. https://doi.org/10.1080/02643290442000310

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, **9**(3), 558–565. https://doi.org/10.3758/BF03196313

Grafton, S. T., & Tipper, C. M. (2012). Decoding intention: A neuroergonomic perspective. *NeuroImage*, **59**(1), 14–24. https://doi.org/10.1016/J.NEUROIMAGE.2011.05.064

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, **41**(2), 301-307. https://doi.org/10.1016/S0896-6273(03)00838-9

Hu J.H. (2021). 中国英语学习者语言模式与权力空间隐喻认知的交互作用研究 [A study on the interaction between language and metaphorical processing of perceived power in Chinese EFL speakers]. 解放军外国语学院学报, **44**(1), 10–18.

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism*, **23**(5), 945–950. https://doi.org/10.1017/S1366728919000038

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, **48**(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Kogan, B., Muñoz, E., Ibáñez, A., & García, A. M. (2020). Too late to be grounded? Motor resonance for action words acquired after middle childhood. *Brain and Cognition*, **138**. https://doi.org/10.1016/j.bandc.2019.105509

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language*, **33**(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kühne, K., & Gianelli, C. (2019). Is embodied cognition bilingual? Current evidence and perspectives of the embodied cognition approach to bilingual language processing. *Frontiers in Psychology*, **10**. https://doi.org/10.3389/fpsyg.2019.00108

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, **207**(4427), 203–205. https://doi.org/10.1126/science.7350657

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word Expectancy and Event-Related Brain Potentials During Sentence Processing. In S. Kornblum & J. Requin (Eds.), *Preparatory States and Processes* (pp. 217–237). Hillsdale Lawrence Erlbaum.

Kuznetsova A., Brockhoff P.B., Christensen R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, **82**(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakoff George, & Johnson Mark. (1999). *Philosophy In The Flesh: The Embodied Mind And Its Challenge To Western Thought*. Basic Books.

Li, Y., Casaponsa, A., Wu, Y. J., & Thierry, G. (2019). Back to the future? How Chinese-English bilinguals switch between front and back orientation for time. *NeuroImage*, **203**. https://doi.org/10.1016/j.neuroimage.2019.116180

Li, Y., Oppenheim, G., & Thierry, G. (2023). Time flows vertically in Chinese. *Brain and cognition*, **170,** 106057. https://doi.org/10.1016/j.bandc.2023.106057

Lenth, R. V. (2020). emmeans: Estimated marginal means, aka least-squares means (R package, Version 1.4.7) [Computer software]. https://CRAN.R-project.org/package=emmeans

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, **8**(1). https://doi.org/10.3389/fnhum.2014.00213

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *Journal of speech, language, and hearing research: JSLHR*, **50**(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Marian, V., & Hayakawa, S. (2021). Measuring Bilingualism: The Quest for a "Bilingualism Quotient". *Applied psycholinguistics*, **42**(Suppl 2), 527–548. https://doi.org/10.1017/s0142716420000533

Marino, B. F. M., Sirianni, M., Volta, R. D., Magliocco, F., Silipo, F., Quattrone, A., & Buccino, G. (2014). Viewing photos and reading nouns of natural graspable objects similarly modulate motor responses. *Frontiers in Human Neuroscience*, **8**(12). https://doi.org/10.3389/fnhum.2014.00968

Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up: association between affect and vertical position. *Psychological science*, **15**(4), 243–247. https://doi.org/10.1111/j.0956-7976.2004.00659.x

Monaco, E., Jost, L. B., Gygax, P. M., & Annoni, J. M. (2019). Embodied semantics in a second language: Critical review and clinical implications. *Frontiers in Human Neuroscience*, **13**. https://doi.org/10.3389/fnhum.2019.00110

Paivio, A. (1990). *Mental Representations*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, **118**(10), 2128–2148. https://doi.org/10.1016/J.CLINPH.2007.04.019

Pulvermüller F. (2005). Brain mechanisms linking language and action. *Nature reviews. Neuroscience*, **6**(7), 576–582. https://doi.org/10.1038/nrn1706

Santana, E. J., & De Vega, M. (2013). An ERP study of motor compatibility effects in action language. *Brain Research*, **1526**, 71–83. https://doi.org/10.1016/j.brainres.2013.06.020

Schubert T. W. (2005). Your highness: vertical positions as perceptual symbols of power. *Journal of personality and social psychology*, **89**(1), 1–21. https://doi.org/10.1037/0022-3514.89.1.1

Singmann H., Bolker B., Westfall J., Aust F., Ben-Shachar M. (2023). _afex: Analysis of Factorial Experiments_. (R package version 1.2-1) [Computer software]. https://CRAN.R-project.org/package=afex.

Stone, J. V. (2002). Independent component analysis: an introduction. *Trends in Cognitive Sciences*, **6**(2), 59–64. https://doi.org/10.1016/S1364-6613(00)01813-1

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**(6), 2372–2382. https://doi.org/10.3758/s13428-020-01389-1

Tian, L., Chen, H., Zhao, W., Wu, J., Zhang, Q., De, A., Leppänen, P., Cong, F., & Parviainen, T. (2020). The role of motor system in action-related language comprehension in L1 and L2: An fMRI study. *Brain and Language*, **201**, 104714. https://doi.org/10.1016/J.BANDL.2019.104714

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, **1**(2), 219–247. https://doi.org/10.1515/langcog.2009.011

Vukovic, N., Feurra, M., Shpektor, A., Myachykov, A., & Shtyrov, Y. (2017). Primary motor cortex functionally contributes to language comprehension: An online rTMS study. *Neuropsychologia*, **96**, 222–229. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2017.01.025

Vukovic, N., & Shtyrov, Y. (2014). Cortical motor systems are involved in second-language comprehension: Evidence from rapid mu-rhythm desynchronisation. *NeuroImage*, **102**(P2), 695–703. https://doi.org/10.1016/j.neuroimage.2014.08.039

Wang, Q. (2016). Embodied Cognition Processing and Representation of Power Words by Second Language Learners with Different Proficiency levels. *Chinese Journal of Applied Linguistics*, **39**(4). https://doi.org/10.1515/cjal-2016-0030

Wu, X., Jia, H., Wang, E., Du, C., Wu, X., & Dang, C. (2016). Vertical position of Chinese power words influences power judgments: Evidence from spatial compatibility task and event-related Potentials. *International Journal of Psychophysiology*, **102**, 55–61. https://doi.org/10.1016/j.ijpsycho.2016.03.005

Xue, J., Marmolejo-Ramos, F., & Pei, X. (2015). The linguistic context effects on the processing of body-object interaction words: An ERP study on second language learners. *Brain Research*, **1613**, 37–48. https://doi.org/10.1016/j.brainres.2015.03.050

Yang, H., Nick Reid, J., Katz, A. N., & Li, D. (2021). The Embodiment of Power as Forward/Backward Movement in Chinese and English Speakers. *Metaphor and Symbol*, **36**(3), 181–193. https://doi.org/10.1080/10926488.2021.1907185

Zanolie, K., Dantzig, S. van, Boot, I., Wijnen, J., Schubert, T. W., Giessner, S. R., & Pecher, D. (2012). Mighty metaphors: Behavioral and ERP evidence that power shifts attention on a vertical dimension. *Brain and Cognition*, **78**(1), 50–58. https://doi.org/10.1016/j.bandc.2011.10.006

Zwaan, R. A., & Yaxley, R. H. (2003). Spatial iconicity affects semantic relatedness judgments. *Psychonomic Bulletin & Review*, **10**(4), 954–958. https://doi.org/10.3758/BF03196557

Supporting Information

Appendix S1 Materials

Appendix S2 Complete output from linear mixed models and ANOVAs (Results based on pre-registration)

Appendix S3: Complete output from linear mixed models and ANOVAs (Exploratory analyses)