Timeline blurring in fluent Chinese-English bilinguals

Li, Yang; Jones, Manon; Thierry, Guillaume

Brain Research

DOI:
10.1016/j.brainres.2018.07.008

Published: 15/12/2018

Peer reviewed version

Timeline blurring in fluent Chinese-English bilinguals

Yang Li, Manon Jones, Guillaume Thierry

School of Psychology, Bangor University, Wales, LL59 5HF, UK.

Abstract

Linguistic relativity effects arising from differences in terminology and syntax between languages have now been established in various domains of human cognition. Although metaphors have been shown to affect time conceptualisation, there is little evidence to date that the presence or absence of tense within a given language can affect how one processes temporal sequences of events. Here, we set out to characterise how native speakers of Mandarin Chinese–a tenseless language– deal with reference time misalignment using event-related brain potentials. Fluent Chinese-English participants and native speakers of English made acceptability judgements on sentences in which the adjunct clause started with the connective 'after' and was either temporally aligned or not with the main clause in terms of reference time conveyed by the verb. Native speakers of English failed to overtly report such reference time misalignments between clauses, but significant N400 modulations showed that they nevertheless required additional semantic processing effort. Chinese speakers, however, showed no such N400 modulation suggesting that they did not covertly detect reference time misalignments between clauses in real time. Critically, all participants manifested normal sentence comprehension as shown by a standard N400 semantic violation elicited by incongruent endings. We conclude that Chinese speakers of English experience difficulties locating events on a timeline in relation to one another when temporal information is conveyed by tense.

Keywords: *Reference time; tenselessness; event-related potentials; syntax-semantics interference; linguistic relativity; sentence processing.*

Corresponding author:

G. Thierry | School of Psychology, Bangor University, Wales, LL59 5HF, UK | g.thierry@bangor.ac.uk

1. Introduction

Recent research has provided evidence that language influences cognitive functioning (Athanasopoulos, 2009; Boroditsky, 2001; Boutonnet, Athanasopoulos, & Thierry, 2012; Choi & Bowerman, 1991; Gentner & Goldin-Meadow, 2003; Lantz & Stefflre, 1964; Lucy, 1992; Lupyan & Ward, 2013; Whorf, 1956). Such effects have been demonstrated at the level of elementary visual perception (Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009) and object categorisation (Boutonnet, Athanasopoulos, & Thierry, 2012; Cubelli, Paolieri, Lotto, & Job, 2011; Phillips & Boroditsky, 2003), through to high-level, abstract meaning processing such as event conceptualization (Flecken, Athanasopoulos, Kuipers, & Thierry, 2015) and cultural semantics (Ellis, et al., 2015).

Grammatical variations between languages can influence event conceptualization as shown by studies of motion event categorization both in language tasks (Flecken, 2011; von Stutterheim & Carroll, 2006) and non-verbal tasks (Athanasopoulos, 2009; Athanasopoulos & Bylund, 2013; Flecken, Athanasopoulos, Kuipers, & Thierry, 2015; Flecken, von Stutterheim, & Carroll, 2014). For example, both English and Arabic speakers –whose native languages have aspect markers– spontaneously mention the temporal properties of motion events (e.g., "Two women are walking down a path"). In contrast, native speakers of German –whose mother tongue lacks aspect categorization altogether– describe the same events in more holistic terms, including the mention of a possible endpoint (von Stutterheim & Carroll, 2006).

Here, we set out to examine whether linguistic differences in *tense* marking can affect the representation of temporal relationships between events. Tense is a linguistic device that locates a given situation in time (Declerck, Reed, & Cappell, 2006). It is accepted that English is a tensed language although there is a controversy over the existence of the future tense (Comrie, 1985; Declerck, Reed, & Cappell, 2006; Quirk, Greenbaum, Leech, & Svartvik, 1985). Other languages lack absolute tense altogether. That is, they do not mark either present-future or past-non-past distinctions in their grammar. In Mandarin Chinese, for example, specifying the temporal location of an event is not compulsory (Comrie, 1985; Li & Thompson, 1989). Instead, temporal information is optionally expressed through time adverbials (e.g. *zuó tiān* –'yesterday'; *míng tiān* – 'tomorrow'), modal auxiliaries (e.g. *yào* – 'will'; *jiāng* – 'will'), or through context (Duff & Li, 2002; Smith, 1991), the default position being that the event unfolds in the present (Smith, 2008). Qiu and Zhou (2012), for instance, found that native speakers of Mandarin Chinese are sensitive to the disagreement between a modal auxiliary (e.g., *jiāng yào* – 'will') or an aspectual particle (e.g., the marker of perfect aspect –*guò*) and a temporal noun phrase (e.g., *shàng gè yuè* – 'last month'), as in the following sentence:

1. * 上个月　联合国　将要　派出　特别调查组。[1]

   Last month　UN　jiangyao　Dispatch　Investigation unit

   Last month, the United Nations will dispatch a special investigation unit.

(From Qiu and Zhou, 2012, p. 94) To investigate whether cross-linguistic differences in tense marking can influence readers' perception of time, we tested fluent Mandarin-English bilinguals reading English sentences. According to the approach proposed by Reichenbach (1947), the timeline corresponding to a situation described by an utterance involves three time points: (a) *Speech time* (the time at which the utterance is produced), (b) *reference time* (the perspective from which a situation is perceived), and (c) *event time* (the time at which the event happens). In order to understand the temporal order of events in a given sentence, and therefore its overall meaning, it is necessary to encode on the one hand the relationship between Speech Time and Reference Time (theoretically encoded by tense), and on the other hand the relationship between Reference Time end Event Time (theoretically encoded by aspect). In the case of a tensed language, the three time points and their relationships are coded directly by inflection (Smith, 2008). However, in the case of tenseless Mandarin Chinese, the relationship between Speech Time and Reference Time can remain unspecified because it is not encoded by an inflectional morpheme within the verb (Smith, 2008) and specifying temporal information is not compulsory (Smith, 1991).

We thus created complex English sentences featuring a reference time misalignment (RTM) between their adjunct and main clauses. In all cases, adjunct clauses began with the connective '*after*' and systematically described a first event with perfect aspect –a grammatical category that exists in both English and Chinese. In the RTM conditions (see Figure 1B and 1C) the adjunct clause was in the present or the future tense, whereas the main clause was in the absolute past tense (simple past). Note that such RTM is different from

---

[1] Pin yin version of sentence: Shàng-gè yuè lián hé guó jiāng yào pài chū tèbié diào chá zǔ.

.

tense violation, since the latter entails grammatically incorrect tense forms within a given clause, as in "Yesterday, I sail Diane's Boat to Boston" (from Steinhauer & Ullman, 2002). We also created a semantic violation condition in which the statement was made meaningless by the presence of an incongruent word ending designed to serve as a semantic control, to test participants' understanding of the materials presented **(**see **Table 1**).

< Insert Table 1 about here >

In control (correct) sentences, the adjunct clause was in the past perfect tense (see **Fig. 1A**). It shared its reference time and speech time with the main clause in the simple past tense, and thus was correct according to the rule of temporal connectives (Hornstein, 1990). The RTM conditions were of two kinds: (1) a Present-Past Misalignment (PPM; **Fig. 1B**) and a Future-Past Misalignment (FPM; **Fig. 1C**), in which the tense of adjunct clauses does not share speech time and reference time with the main clause.

< Insert Figure 1 about here >

We tested monolingual speakers of English as controls and Chinese-English speakers, who were fluent in both Mandarin Chinese and English, that is, able to hold a normal, fluid conversation in either of the two languages (Titone,1972; Macnamara, 1967; Grosjean, 1989). We did not expect marked differences between fluent Chinese-English bilinguals and native speakers of English in a sentence acceptability task. However, we expected event-related brain potentials (ERPs) recorded simultaneously to index differences in online processing of temporal information. For native English speakers, we predicted that RTMs would elicit greater negativity in the N400 range compared with correct sentences (see Newland, 2015; Liang et al., 2016), owing to semantic difficulties in reconciling the incongruous timelines presented in the adjunct and main clauses (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984; Van Petten & Kutas, 1990). However, because tense encodes the relationship between reference time and speech time, and since Mandarin Chinese does not encode tense directly, we predicted that native Chinese speakers would be less sensitive to RTMs than their native English peers. This should translate into a relatively weaker N400 modulation in bilinguals and we thus predicted a group x condition interaction in the N400 range.

2. Results

2.1. Behavioural data

Reference Time Misalignment. Collinearity was not an issue in the models used for analysis: Fixed-effects correlations (|*r*|) were less than 0.7 for all predictors. For both accuracy and RT models, the intercept represents the average likelihood that English participants were accurate in the control condition. Each coefficient compares the average for a different combination of fixed factor levels against this intercept, and *p* values are derived from the normal approximation method (Barr et al., 2013, see **Table 2** and **3**).

< Insert Tables 2 and 3 about here >

Both PPM and FPM yielded significantly lower accuracy compared with the past perfect control condition **(Fig. 2** and **Table 2**, *FF1* and *FF2*). There was no fixed effect of Native language (*FF3*), indicating that English and Chinese participants had similar performance for baseline control sentences. No interaction was found (*I1* and *I2*), suggesting that bilingual Chinese-English participants were similarly *inaccurate* in detecting either type of RTM.

< Insert Figure 2 about here >

As for RTs, English monolingual participants showed no processing time costs for RTM conditions compared with the intercept condition, and Chinese-English bilinguals did not differ from English monolinguals at baseline (**Table 2**, *FF1, FF2, FF3*). However, a significant interaction in the analysis showed that Chinese bilinguals had longer RTs in the PPM condition relative to the additive contribution to the model of their RTs in the control condition, and the RTs of the English natives in the control and PPM conditions (*I1*). No such interaction emerged for the FPM, however (*I2*).

Semantic Violation: Accuracy and reaction times (RT) were modeled as a function of native language (English, Mandarin Chinese) as between-group factor, and semantic violation (control, semantic violation) as within-participant factor. In all other respects, our models were implemented similarly to those described in the previous section: Intercept values comprised the average likelihood that English monolingual participants were accurate in the baseline control condition. For accuracy data (see **Fig. 3**), all participants performed at ceiling (> 97% accuracy on average).

< Insert Figure 3 about here >

Thus, no significant differences emerged in the model (**Table 3**, *FF1, FF2 and I1*). For RT data, English monolinguals took the same amount of time to respond to the control as the semantic violation sentences (*FF1*), and Chinese-English bilinguals did not differ from English controls at baseline (*FF2*). However, a significant interaction emerged, such that Chinese-English bilinguals were significantly slower responding to semantic violations than the additive contribution to the model of their own performance on correct trials and English monolinguals' performance in correct and semantic violation trials (*I1*).

2.2. Electrophysiological data

2.2.1. Reference Time Misalignment

ERP mean amplitudes were analysed with repeated-measures analyses of variance (ANOVA). Mean amplitudes were modelled as a function of native language (English, Mandarin Chinese) and condition (past perfect control, PPM and FPM). The N400 effect was examined after the onset of the critical verb (henceforth N4-1) and after the onset of the next word in line (henceforth N4-2).

N4-1: There was a significant main effect of native language ($F(1, 44) = 7.35$, $p = 0.01$, $\eta_p^2 = 0.14$) on N4-1 mean amplitude, and a significant interaction between native language and condition ($F(2, 88) = 4.84$, $p = 0.01$, $\eta_p^2 = 0.1$). The condition main effect was not significant ($F(2, 88) = 0.97$, $p = 0.38$, $\eta_p^2 = 0.02$). In English controls, N400 negativity was significantly greater in the PPM than in the baseline condition ($t(18) = 1.387$, $p = 0.09$; one-tailed t-test). In Chinese-English bilinguals, however, the difference between PPM and baseline condition did not attain statistical significance ($t(26) = -0.55$; $p = 0.29$; the one-tailed t-test; **Fig. 4a; Supplementary Fig. 1**). As for the FPM versus baseline comparison, native speakers of English had significantly more negative N400 amplitudes in response to FPM ($t(18) = 2.637$, $p = 0.01$; one-tailed t-test, **Supplementary Fig. 2**) but we found no such difference in the

Chinese-English bilinguals (t (26) =-1.62, p=0. 06; one-tailed t-test). In fact, the difference tended to go in the opposite direction in Chinese participants **(Fig. 4b,**).

< Insert Figure 4 about here >

N4-2: There was a significant main effect of native language on N4-2 mean amplitudes (F(1,44)=42.50, p<0.001, η_p^2= 0.25), and a significant interaction between native language and condition (F(2,88)=3.48;p=0.035, η_p^2= 0.073). However, the main effect of condition was not significant (F(2,88)=2.66, p=0.08, η_p^2= 0.06). In English native controls, N400 mean amplitudes elicited by the post-critical word (N4-2) differed statistically between PPM and baseline conditions (t (18) =2.54; p=0.01; one-tailed t-test) but this difference was not significant in the Chinese-English bilingual group (t (26) =-0.3; p=0.38; one-tailed t-test; **Fig. 4a; Supplementary Fig. 1**). FPM and baseline conditions also difference significantly in native speakers of English (t (18) =2.74; p<0.01; one-tailed t-test; **Supplementary Fig. 2**) and again, this difference was not significant in the Chinese participants (t (26) =-0.29; p =0.39; one-tailed t-test; **Fig. 4b**).

2.2.2. Semantic Violation

The N400 elicited by the sentence-final word was analysed using a repeated-measures ANOVA with native language (Mandarin Chinese, English) as between-group factor and semantic violation (control condition, semantic condition) as within-subject factor. Only the main effect of semantic violation proved statistically different (F (1,44) =20.58, p < 0.001, η_p^2=0.32; **Fig. 5**); there was no significant effect of native language (F (1,44) = 0.75, p=0.75, η_p^2=0.002), nor was there an interaction between native language and semantic violation (F (1,44) = 3.43, p=0.07, η_p^2=0.07). Both native speakers of English and Chinese-English bilinguals showed greater negativity in the N400 range for semantic violations as compared to control sentences (English: t (18) =3.39, p=0.002; one-tailed t-test; Mandarin Chinese: t (26) =2.6, p=0.008; one-tailed t-test).

< Insert Figure 5 about here >

3. Discussion

Here we examined whether linguistic differences between Mandarin Chinese and English in the domain of temporal encoding influences speakers' sensitivity to temporal sequence violations. Despite showing a lack of metacognitive awareness regarding the semantic acceptability of sentences featuring a reference time misalignment, native speakers of English produced a significant N400 modulation in response to the verb in critical position as well as the post-critical word. As expected, N400 mean amplitude was increased for verbs inducing an RTM as compared to verbs that did not conflict with the reference time of the adjunct clause. Critically, no such effect was found in fluent Chinese-English bilinguals with a good command of English grammar: Not only were they indistinguishable from their monolingual peers in terms of acceptability judgement, but in addition, there was no measurable shift in N400 amplitude in any of the conditions. Nevertheless, lexical-semantic violations elicited a classical N400 modulation in both groups, suggesting good levels of overall sentence meaning integration.

Most previous research investigating tense processing in L2 learners, healthy L1 speakers, or language impaired L1 speakers have tended to use simple constructs in which a time adverbial and the tense form are incompatible (i.e., the tense violation paradigm; see Steinhauer & Ullman, 2002; White, Genesee, & Steinhauer, 2012; Newman, Ullman, Pancheva, Waligura, & Neville, 2007). These studies focused on ERP amplitude modulations of either the P600, the left anterior negativity (LAN), or a biphasic LAN-P600 complex. However, in the current study, we did not investigate tense violations occurring within a given clause, but rather reference time misalignment *between* two clauses, each of them being grammatically correct when considered independently. Tense, thus, had global rather than local temporal relevance in our study. It was therefore only when participants encountered the verb of the main clause that they were in a position to perceive a reference time misalignment, bearing in mind that the tense used in the main clause did not constitute a tense violation until they were able to recall the temporal information encoded in the first clause. And indeed, an explorative analysis revealed no significant modulation of P600 mean amplitudes in either group of participants and for any experimental contrast.

We expected that native speakers of English would identify RTMs or –at the very least– that they would detect them more often than their Chinese-English bilingual peers. This is because an RTM produces a semantic incongruence at the level of the entire statement and results in a content that effectively does not 'make sense'. The absence of RTM detection in the behavioural data suggests that the information conveyed by tense can be subtle, especially when the misalignment depends on long-range integration of information across two clauses. This may be explained by automatic repair mechanisms in reading, especially in the context of this experiment in which we used word-by-word presentation and given that RTM differences are rather difficult to identify in general. Indeed, word-by-word presentation (Kaiser, 2014; Marinis, 2010; VanPatten, 2014) is very unnatural (even though it is often imperative in ERP studies of reading) and it is likely to tap into working memory more than natural reading, which may have contributed to blurring the events' timeline. Also, the task used in the experiment likely biased the participants to make basic semantic adequacy judgements because of the presence of a clearly aberrant word in the semantic violation condition. In a recent study by Nieuwland (2015), participants were required to either explicitly assess stimulus plausibility or simply read the same statements for comprehension. In both case, participants displayed larger N400 amplitudes for stimuli which were inconsistent with real-world knowledge. In addition, our data is consistent with recent findings from the language comprehension literature, in which language processing is construed as "good enough" (characterized by underspecified grammatical representations). For the purposes of our offline task, participants may have been using a simple heuristic to interpret these sentences according to existing schemata; avoiding full linguistic computation since this was task-irrelevant (e.g. Ferreira & Karimi, 2015; Ferreira & Patson, 2007). Thus, the subtle between-clause violations in FPM and PPM conditions may have escaped participants' initial scrutiny in terms of the degree to which these sentences "made sense".

Critically, however, English speakers did process the tense configuration of the matrix clauses as indicated by a modulation of the N400 elicited by the post-critical word following the locus of a reference time misalignment in the case of PPMs, and both the critical verb and the following word in the case of FPMs. We interpret this result as showing that the temporal representation of events was successfully extracted on the basis of tense information by native speakers of English, even though this did not translate into behavioural effects. Note that the RTM resulted in an N400 modulation as early as the critical verb for FPM but only at the post-critical word in the case of PPM. Even though we did not predict such a difference, we could have anticipated this on the basis of the magnitude of the misalignment. Indeed, an FPM is arguably more salient than a PPM, due to the time gap being wider. In addition, recall that it is a matter of debate whether or not the future form in English qualifies as tense, due to the mandatory use of the auxiliary 'will'. In other words, it could be that the auxiliary produced a strong expectation for a shift of the reference time into the future, leading to more salient incongruence than in the case of the PPM.

It must be noted that although the reference time of the present perfect is the present, it is mostly used to describe events that have happened in the recent past, that is, in the pre-present zone (Declerck, Reed, & Cappell, 2006). We propose that this contributes to making the PPM condition relatively less contrastive than the FPM condition. In this case, one could reasonably expect N400 modulations to appear later for PPM than FPM conditions, an effect akin to a spill-over, which is commonly observed in eye-tracking studies (Kaiser, 2014; Keating, 2014; Reichle, Pollatsek, & Rayner, 2007). Beyond the fact that spill-over effects have previously been identified in studies of tense violation in L1 (e.g. Qiu & Zhou, 2012*)*, it is unsurprising that the N400 modulation elicited by RTM should carry over to the post-critical word because of the requirement for cross-clause integration in order to retrieve the temporal relationship between the two events described.

One may wonder if the language proficiency of our Chinese-English bilingual participants was high enough to detect RTMs. First, the native speakers of Mandarin Chinese involved in this study performed with a very high level of accuracy in the semantic violation conditions, on a par with their English native peers, indicating that their level of comprehension was indeed excellent. Second, most of the bilingual participants obtained a high score at the IELTS, a standard test of English proficiency. Although 5 participants did not provide a score, their English proficiency level was expected to be high since they received instruction exclusively through the medium of English from high-school onwards. We also conducted a split-group analysis excluding the participants without a numerical score to test for potential differences in RTM sensitivity in relation to IELTS score and found that the latter failed to relate to the former (see **Supplementary Analysis**).

Importantly, all bilingual participants involved in this study reported having high English proficiency (**Fig. 6**). Based on an extensive review of the literature (e.g. LeBlanc & Painchaud, 1985; Palmer & Bachman, 1981; Rea, 1981; von Elek, 1981, 1982), Blanche and Merino (1989) concluded that self-reports provide "good or very good" measures of proficiency, and such measures are often used in ERP experiments involving bilingual participants since they are very quick to obtain (e.g. Dowens, Guo, Guo, Barber, & Carreiras, 2011; Lehtonen, et al., 2012).

To further assess the role of proficiency in the results, we tested an additional group of 21 Chinese-English bilinguals closely matched in IELTS score with the participants tested here on an overt time alignment judgment task along with a new group of native English controls. This new group of Chinese-English bilingual performed similarly to their English native peers (see **Supplementary Analyses**). Therefore, we assume that low proficiency in English is not the reason why Chinese participants failed to detect PPMs and FPMs.

Note that Chinese-English bilinguals needed a longer time to judge whether PPM sentences were acceptable as compared to control ones. It may be that re-evaluation mechanisms taking place over the processing of the entire sentence were longer in this condition because the sentences were in fact perceived as 'strange', but this effect could simply stem from the ambiguity of the present perfect form itself: is it a past form or a present form? Independently of whether there is an RTM between clauses, the delay in the PPM condition would then be due to internal processing issues rather than RTM resolution.

It is thus likely that the lack of detection of RTMs in the bilingual participants relates in some way to cross-linguistic differences between Mandarin Chinese and English, and more particularly, to the way temporal information about events is conveyed by language. Although Mandarin Chinese, just like English, features the perfective aspect, it has no direct equivalent for tense. This means that Chinese-English bilinguals reading a perfect form in English will know that the particular event described in the adjunct clause is completed but will have difficulty figuring out when completion occurs: past, present, or post-present.

The relative inability of Chinese-English bilinguals to perceive RTMs in English may have implications well beyond the domain of second language sentence comprehension and indeed concern time conceptualisation more generally. Given that native speakers of Mandarin Chinese tend to culturally care about the past more than their Canadian counterparts (Ji, Guo, Zhang, & Messervey, 2009), difficulties in identifying temporal relations in English may lead to significant misunderstandings in everyday language use. In other words, Chinese-English bilinguals may be expected to experience a blurred relationship between past and present when interacting in English, which would stand in stark contrast with their experience of the same relationship in their native language.

In conclusion, Chinese-English bilinguals do not register reference time misalignments between clauses when event time is encoded by tense in English, and probably over-rely on the adverbial form "after" to figure out temporal order, since the same is used in their native language to specify temporal sequencing. Despite such strategy, they fail to accurately position two events in relation to one another on the timeline, which becomes blurred as a result. Future studies will determine whether such effects remain when participants are directly instructed to process temporal sequences.

4. Materials and Methods

4.1. Participants

Thirty Chinese-English bilinguals and 25 native speakers of English took part in this study. Data from three bilingual participants and 6 native speakers of English were discarded due to poor ERP data quality, such as heavy blinking and excessive alpha elicitation. Of the remaining 27 bilingual participants, 10 were males and 17 females, with a mean age of 22.3 (*SD=2.7*) and were all right-handed. In the English native group, 19 participants dataset were kept (8 males and 11 females; *Mean* age= *22.4, SD=9.3;* one left-hander and 18 right-handers). All participants were students at Bangor University, UK, and received either payment or course credits for their participation.
The average age at which Chinese-English bilinguals started to learn English was 6.9 years (*SD*=3.2), and all were living in the UK at the time of testing. The average IELTS score for the bilingual group was 6.9 (*SD*=0.5, from 6.5 to 8). Five further bilinguals did not provide IELTS scores, since they received English medium instruction since high school. **Table 4** summarizes the Chinese-English bilinguals' language background.
Bilingual participants also self-reported their proficiency in both English and Mandarin Chinese (see **Fig. 6**). All participants had normal or corrected to normal vision. The study was approved by the School of psychology, Bangor University ethics committee.

4.2. Stimuli

The materials consisted of 70 sentence sets, each containing 8 sentences. Four were experimental sentences featuring either a (i) correctly tensed verb, (ii) PPM, (iii) FPM, or (iv) semantic violation (see **Table 1**) and 4 sentences served as fillers. The locus of the reference time misalignment coincided always with the second word of the main clause. For the main analyses, we compared the control condition (i) with the two RTM conditions (ii) and (iii). An additional analysis comprised (i) and (iv), in order to ascertain that the Chinese-English bilinguals comprehended the overall meaning of the sentences.

In order to dilute the critical experimental manipulations, filler sentences were included, in which the matrix sentences used the simple future tense. There were two presentation lists, which alternated so as to present experimental items and fillers in a fully counterbalanced fashion. Each presentation list featured 4 blocks and a given sentence from a given condition was only presented once per block. Stimuli from the same set were never presented together in the same block. In addition, verb regularity was systematically manipulated such that half were regular and the other half irregular. There was no significant difference in lexical frequency between regular and irregular lists even though there was a trend for irregular verbs to be more frequent ($U = 451.5$; $p = 0.06$).

4.3. Procedure

Bilingual participants first filled out a language background questionnaire. All participants were seated 100 cm away from a 19-inch computer monitor and responded by pressing button on a reaction time box. The first clause of each sentence was presented at once and participants were instructed to press any button when they had finished reading. The rest of the sentence then comprised individually presented words, in the centre of the screen, for a duration of 300 ms (ISI 400 ms), in order to minimise eye movements. Once the whole sentence had been read, participants were required to judge whether or not it made sense (see **Fig. 7**).

4.4. Design and behavioural data analysis

In this experiment, we compared two groups (English native speakers, Chinese-English bilinguals) and, within-subject, three reference time alignment conditions (correct, PPM, FPM). In addition, participants understanding of the sentences was assessed by analysing effects of semantic violations in sentence completions (final word). Accuracy and reaction times (RT) were modeled as a function of one between-groups factor: Native language (English, Mandarin Chinese), and one within-subject factor: RTM (correct, PPM, FPM). Accuracy was analyzed using a binomial logistic regression. Reaction time data were log transformed and analyzed based on linear mixed effects modeling using R (R Development Core Team, 2008) and the lme4 library (Bates, Maechler, & Dai, 2008). β-values are reported and tested at $p < 0.05$. As recommended by Barr, Levy, Scheepers and Tily (2013), we modeled the maximal random effect values of participant and items intercepts and slopes across groups and condition in both models (when models successfully converged).

4.5. ERP recording and Analysis

Electrophysiological data were recorded at a rate of 1 kHz from 64 Ag/AgCl electrodes according to the extended 10-20 convention using an online (0.05 – 200 Hz) bandpass filter. Two additional electrodes were used to monitor eye movements, one below and one above the right eye. Electrode Cz was the reference electrode and impedances were kept below 5 kΩ. EEG data was filtered bandpass using zero-phrase shift digital filtering (0.1 Hz, 24 dB/oct- 20 Hz, 48 dB/oct). Periods of EEG instability corresponding to experiment pauses were removed manually as well as major artefacts through visual inspection of the data and then we adopted the procedure proposed by Gratton, Coles and Donchin (1983) to mathematically correct eye-blink artefacts. ERPs were computed from epochs ranging from -200 ms to 1500 ms after the onset of critical word, always in second position within the main clause. For the semantic violation condition, epochs ranged from -200 ms to 1500 ms, so as to coincide with onset of the sentence-final word. Epochs with any activity exceeding ±100 µV at any electrode site except electroocculogram channels were eliminated. More than 30 trials in each participant and condition were included in the averaging procedure. Baseline correction was performed in reference to pre-stimulus activity and individual averages were digitally re-referenced to the global average reference. All analyses were conducted again using the average of the mastoid electrodes as reference and all effects reported based on the global average reference were qualitatively replicated in this analysis.

For RTM analyses, we measured ERP amplitudes over 6 centroparietal electrodes, CP1, CPz, CP2, Cz, C1, C2 at which the N400 is usually maximal (Kutas & Hillyard, 1980a; Kutas & Hillyard, 1980b; Kutas & Hillyard, 1984). We identified two time-windows for analysis, the usual N400 time-window between 350–500 ms after the onset of the critical word (the verb in the main clause: e.g., *worked*) and a window between 1200–1350 ms corresponding to the N400 window of the post-critical word. For semantic violation analyses, N400 modulations were analysed between 350–500 ms after the onset of the final word.

References

Athanasopoulos, P. (2009). Cognitive representation of colour in bilinguals: The case of Greek blues. *Bilingualism-Language and Cognition*, 12, 83-95.

Athanasopoulos, P., & Bylund, E. (2013). Does Grammatical Aspect Affect Motion Event Cognition? A Cross-Linguistic Comparison of English and Swedish Speakers. *Cognitive Science*, 37, 286-309.

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, 255-278.

Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed effects models using s4 classes

Blanche, P., & Merino, B.J. (1989). Self-assessment of foreign-language skills implications for teachers and researchers. *Language Learning*, 39, 313-340.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43, 1-22.

Boutonnet, B., Athanasopoulos, P., & Thierry, G. (2012). Unconscious effects of grammatical gender during object categorisation. *Brain Res*, 1479, 72-79.

Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean:The influence of language-specific lexicalization patterns. *Cognition*, 41, 83-121.

Comrie, B. (1985). *Tense*. Cambridge: Cambridge University Press.

Cubelli, R., Paolieri, D., Lotto, L., & Job, R. (2011). The effect of grammatical gender on object categorization. *J Exp Psychol Learn Mem Cogn*, 37, 449-460.

Declerck, R., Reed, S., & Cappell, B. (2006). *The Grammar of the English Verb, I: The Grammar of the English Tense System: A Comprehensive Analysis*. Berlin: Mouton de Gruyter.

Dowens, M.G., Guo, T.M., Guo, J.J., Barber, H., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish - Evidence from Event Related Potentials. *Neuropsychologia*, 49, 1651-1659.

Duff, P., & Li, D. (2002). The acquisition and use of perfective aspect in Mandarin. In S. R & S. Y (Eds.), *The L2 acquisition of tense-aspect morphology* (pp. 417-453). Philadelphia, PA: John Benjamins.

Ellis, C., Kuipers, J.R., Thierry, G., Lovett, V., Turnbull, O., & Jones, M.W. (2015). Language and culture modulate online semantic processing. *Social Cognitive and Affective Neuroscience*, 10, 1392-1396.

Ferreira, F., & Karimi, H. (2015). Prosody, performance, and cognitive skill: Evidence from individual differences. In L.F.E. Gibson (Ed.), *Explicit and implicit prosody in sentence processing: Studies in honor of Janet Dean Fodor* (pp. 119-132). Switzerland: Springer International Publishing.

Ferreira, F., & Patson, N.D. (2007). The “good enough” approach to   language comprehension. *Language and Linguistics Compass*, 1, 71-83.

Flecken, M. (2011). What native speaker judgments tell us about the grammaticalization of a progressive aspectual marker in Dutch. *Linguistics*, 49, 479-524.

Flecken, M., Athanasopoulos, P., Kuipers, J.R., & Thierry, G. (2015). On the road to somewhere: Brain potentials reflect language effects on motion event perception. *Cognition*, 141, 41-51.

Flecken, M., von Stutterheim, C., & Carroll, M. (2014). Grammatical aspect influences motion event perception: Evidence from a cross-linguistic non-verbal recognition task. *Language and Cognition*, 6, 45-78.

Gentner, D., & Goldin-Meadow, S. (2003). Whither Whorf. In G. D & G.-M. S (Eds.), *Language in mind:Advances in the study of language and thought* (pp. 3-14). Cambridge, MA: MIT Press.

Gratton, G., Coles, M.G.H., & Donchin, E. (1983). A new method for off-line removal of ocular atficat *Electroencephalography and Clinical Neurophysiology*, 55, 468-484.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, 36,3-15.

Hornstein, N. (1990). *As Time Goes By: Tense and Universal Grammar*. MIT Press.

Ji, L.J., Guo, T.Y., Zhang, Z.Y., & Messervey, D. (2009). Looking Into the Past: Cultural Differences in Perception and Representation of Past Information. *Journal of Personality and Social Psychology*, 96, 761-769.

Kaiser, E. (2014). Experimental paradigms in psycholinguistics. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 135-168): Cambridge University Press.

Keating, G.D. (2014). Eye-tracking with text. In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 69-92). London: Taylor & Francis.

Kutas, M., & Federmeier, K.D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62, 621-647.

Kutas, M., & Hillyard, S.A. (1980a). Reading between the lines - event-related brain potentials during natural sentence processing *Brain and Language*, 11, 354-373.

Kutas, M., & Hillyard, S.A. (1980b). Reading senseless sentences - brain potentials reflect semantic incongruity. *Science*, 207, 203-205.

Kutas, M., & Hillyard, S.A. (1984). Event-related brain potentials (ERPs) elicited by novel stimuli during sentence processing *Annals of the New York Academy of Sciences*, 425, 236-241.

Lantz, D., & Stefflre, V. (1964). Language and cognition revisited. *Journal of Abnormal and Social Psychology*, 69, 472-481.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, 19, 673-687.

Lehtonen, M., Hulten, A., Rodriguez-Fornells, A., Cunillera, T., Tuomainen, J., & Laine, M. (2012). Differences in word recognition between early bilinguals and monolinguals: Behavioral and ERP evidence. *Neuropsychologia*, 50, 1362-1371.

Li, C.N., & Thompson, S.A. (1989). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Liang, X., Xiao, F., Wu, L., Chen, Q., Lei, Y., & Li, H. (2016). The Temporal Order of Word Presentation Modulates the Amplitudes of P2 and N400 during Recognition of Causal Relations. *Frontiers in Psychology*, 7, 1890

Lucy, J. (1992). *Language diversity and thought*. Cambridge: Cambridge University Press.

Lupyan, G., & Ward, E.J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 14196-14201.

Macnamara, J., (1967). The bilingual's linguistic performance. *Journal of Social Issues*, 23, 58-77.

Marinis, T. (2010). On-line sentence processing methods in typical and atypical populations. In S. Unsworth & E. Blom (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 139-162): John Benjamins.

Newman, A.J., Ullman, M.T., Pancheva, R., Waligura, D.L., & Neville, H.J. (2007). An ERP study of regular and irregular English past tense inflection. *Neuroimage*, 34, 435-445.

Nieuwland M. (2015). The truth before and after: brain potentials reveal automatic activation of event-knowledge during sentence comprehension. *J Cog Neurosci*. 27,2215–2228

Palmer, A.S., & Bachman, L.F. (1981). Basic concerns in test validation. In J.C.A.A. Hughe (Ed.), *ELTdocuments 111: Issues in language testing*. London, United Kingdom: The British

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, Pts 1 and 2*, 928-933.

Qiu, Y.C., & Zhou, X.L. (2012). Processing temporal agreement in a tenseless language: An ERP study of Mandarin Chinese. *Brain Research*, 1446, 91-108.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English Language*. London: Longman.

Rea, P.M. (1981). Formative assessment of student performance: The role of self-appraisal. *Indian Journal of Applied Linguistics,* 7, 66-88.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. New York: Macmillan & Co.

Reichle, E.D., Pollatsek, A., & Rayner, K. (2007). Modeling the effects of lexical ambiguity on eye movements during reading. In R. van Gompel (Ed.), *Eye Movements: A Window on Mind and Brain* (pp. 271-292). Oxford: Elsevier Science.

Smith, C. (1991). *The Parameter of Aspect*. Dordrecht: Kluwer Academic Publishers.

Smith, C. (2008). Time with and without tense. In J. Guéron & J. Lacarme (Eds.), *Time and Modality* (pp. 227–249). Berlin: Springer.

Steinhauer, K., & Ullman, M.T. (2002). Consecutive ERP effects of morpho-phonology and morpho-syntax. *Brain and Language*, 83, 62-65.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 4567-4570.

Titone, R., (1972). *Bilinguismo pecoce ed educazione bilingue*. Amando (Ed.) Roma.

VanPatten, B. (2014). The psycholinguistics of SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction*: Routledge.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word-frequency in event-related brain potentials. *Memory & Cognition*, 18, 380-393.

von Elek, T. (1981). *Self-assessment ofSwedish as a second language(Sja1vbedomning av fardigheter i svenska som andra sprak).* Gijteborg, Sweden: University of Goteborg, Language Teaching Research Center.

von Elek, T. (1982). *Test of Swedish as a second language: An experiment in self-assessment.* Goteborg, Sweden: University of Goteborg, Language Teaching Research Center.

von Stutterheim, C., & Carroll, M. (2006). The impact of grammatical temporal categories on ultimate attainment in L2 learning. In H. Byrnes., H. Weger-Guntharp., & K. Sprang. (Eds.), *Educating for Advanced Foreign Language Capacities* (pp. 40–53). Georgetown: GUP.

White, E.J., Genesee, F., & Steinhauer, K. (2012). Brain Responses before and after Intensive Second Language Learning: Proficiency Based Changes and First Language Background Effects in Adult Learners. *Plos One*, 7.

Whorf, B. (1956). *Language, thought, and reality*. Cambridge, MA: MIT Press.

Figure Captions

Figure 1. Timeline representation of the critical stimuli. E: Event time; R: Reference time; S: Speech time. Index number identifies the clause (1 for the adjunct clause and 2 for the main clause).

Figure 2. Mean accuracy and Reaction time of semantic judgement for control condition, Present-Past time clash and Future-past time clash conditions for both English native participants and Chinese-English bilingual participants. Error bars represent s.e.m.

Figure 3. Mean accuracy and Reaction time of semantic judgement for control condition and Semantic Violation condition for both English native participants and Chinese English bilingual participants. Error bars represent SE of the mean.

Figure 4. ERPs elicited by reference time alignment manipulations. (a). ERPs elicited by the critical word and post-critical word in the past perfect control condition (black lines) and the PPM condition (blue line); (b). ERPs elicited by the critical word and post-critical word in the past perfect control condition (black lines) and the FPM condition (red line). Left, English natives; Right, Chinese-English bilinguals. ERP graphs depict variations of a linear derivation of channels C1, C2, Cz, CP1, CP2, and CPz.

Figure 5. ERPs elicited by the last word in the past perfect control condition (black line) and the semantic violation condition (yellow line). Left, English native speakers; Right, Chinese English bilinguals. ERP graphs depict variations of a linear derivation of channels C1, C2, Cz, CP1, CP2, and CPz.

Figure 6. Chinese-English bilingual participants' self-estimation of their Chinese and English level (10 point- scale). Error bar represents stand error.

Figure 7. Structure of an experimental trial.

Figures

Figure 1

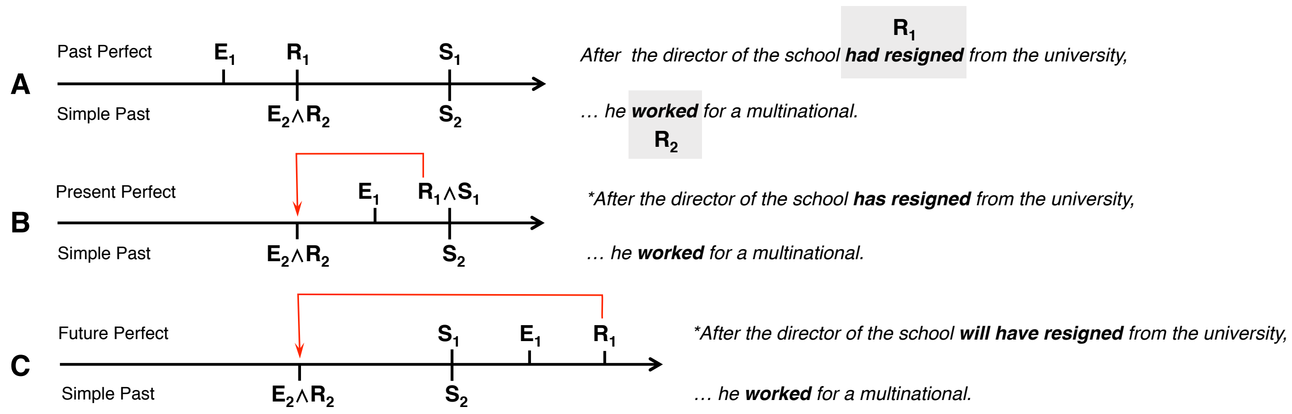
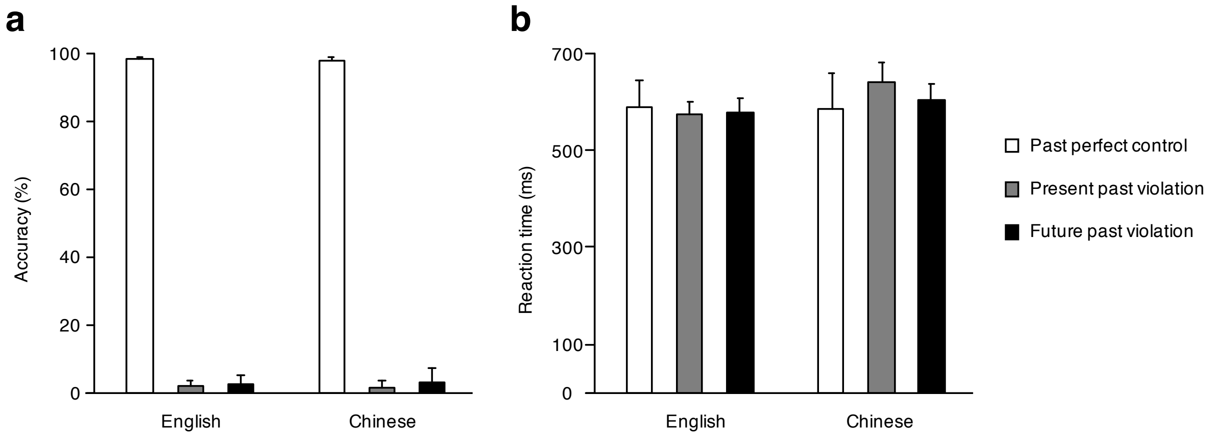
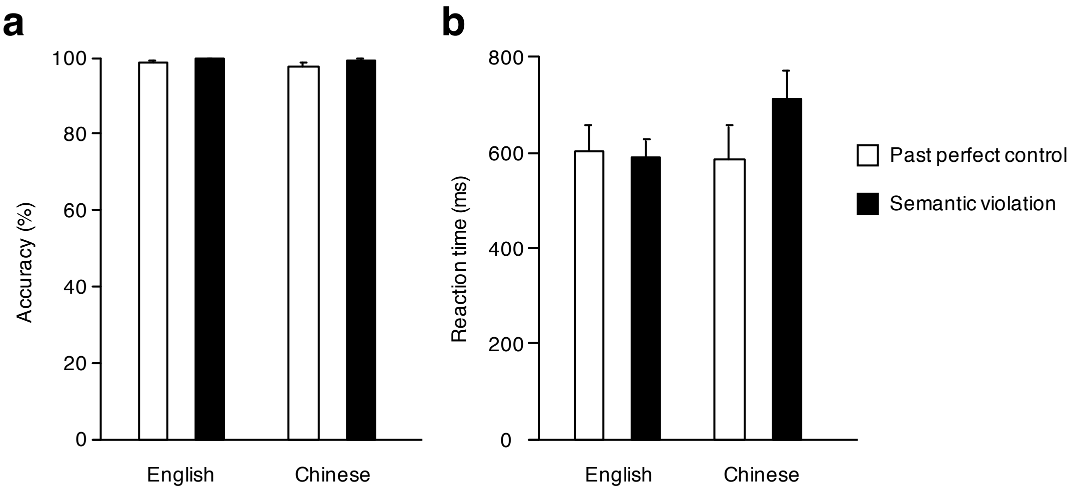


Figure 2



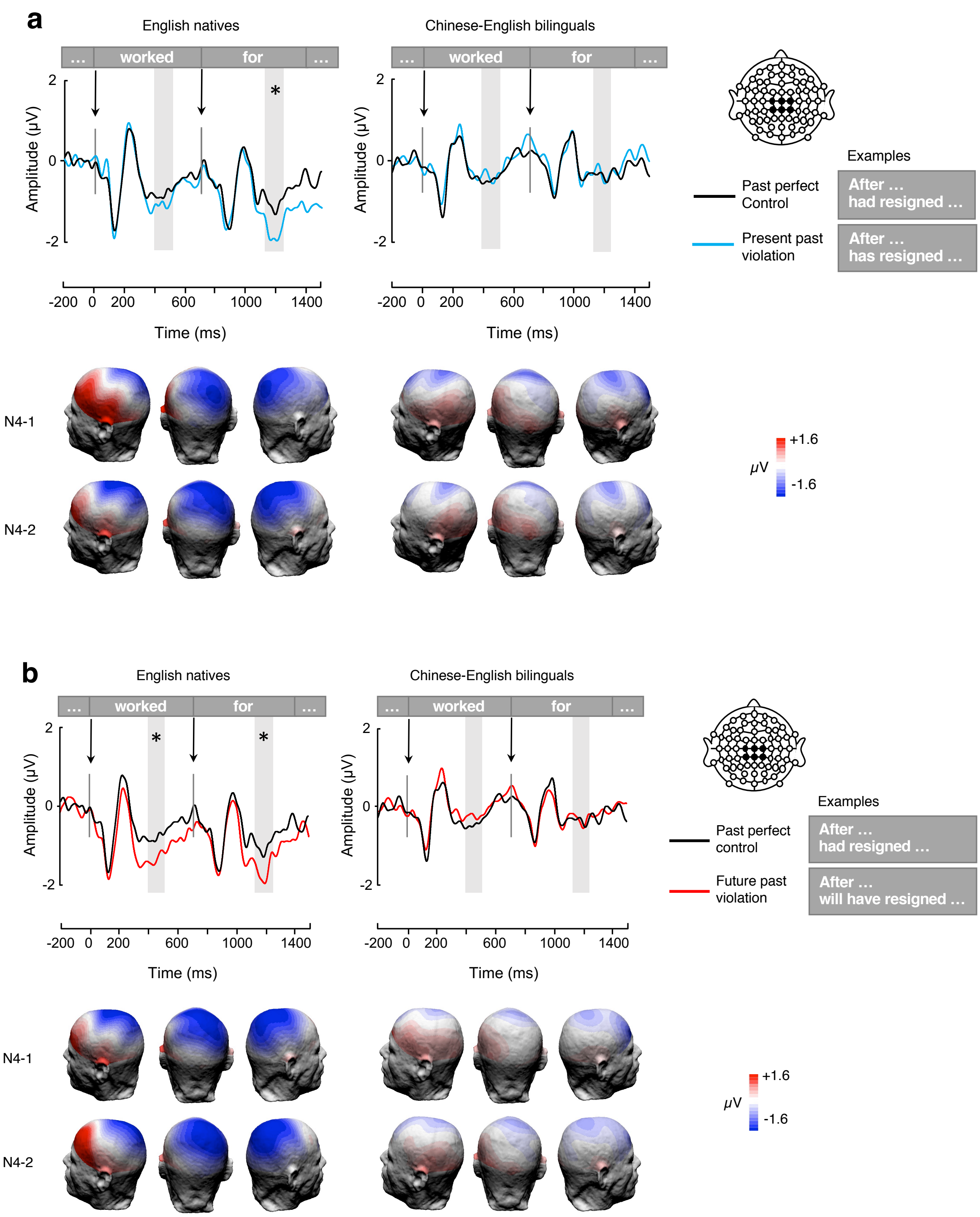
a
Accuracy (%)
100
80
60
40
20
0
English
Chinese
b
Reaction time (ms)
700
500
300
100
0
English
Chinese
Past perfect control
Present past violation
Future past violation

Figure 3



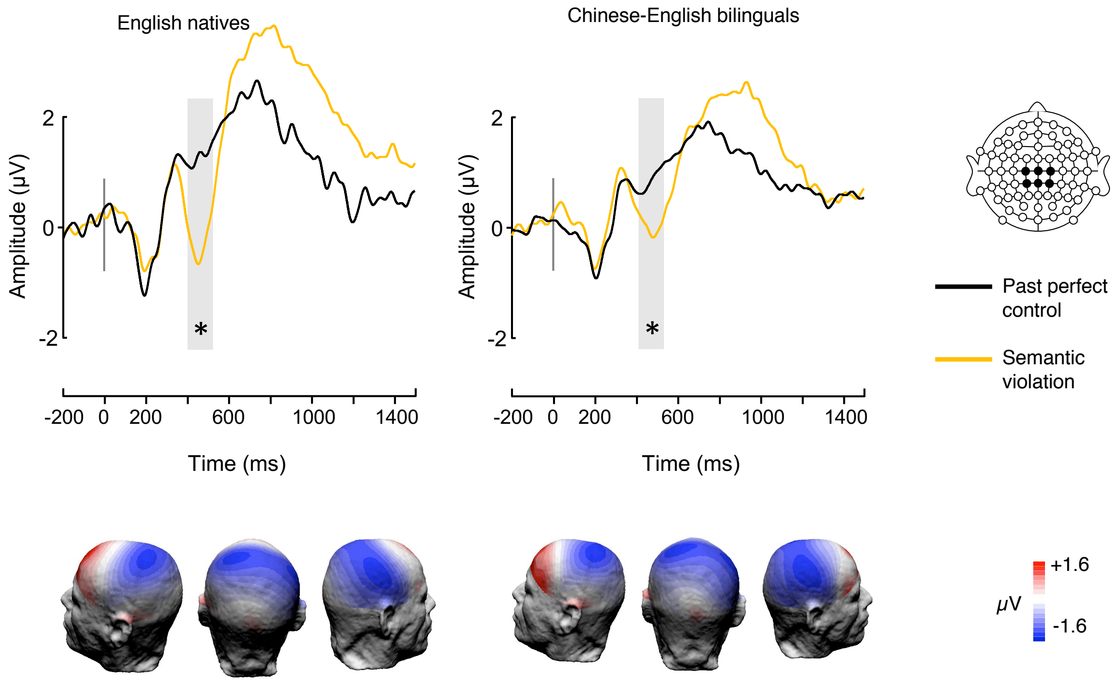
a
100
80
60
40
20
0
Accuracy (%)
English
Chinese
b
800
600
400
200
0
Reaction time (ms)
English
Chinese
Past perfect control
Semantic violation

Figure 4



a
English natives
Chinese-English bilinguals
… worked for …
2
0
-2
Amplitude (µV)
-200 0 200 600 1000 1400
Time (ms)
*
Past perfect
Control
Present past
violation
Examples
After …
had resigned …
After …
has resigned …
N4-1
N4-2
µV
+1.6
-1.6
b
English natives
Chinese-English bilinguals
Past perfect
control
Future past
violation
After …
will have resigned …

Figure 5



English natives
Chinese-English bilinguals
Amplitude (μV)
2
0
-2
*
-200
0
200
600
1000
1400
Time (ms)
Past perfect control
Semantic violation
μV
+1.6
-1.6

Figure 6

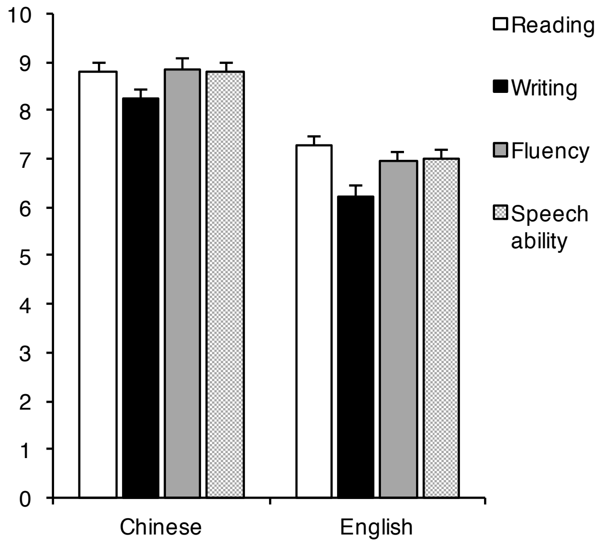
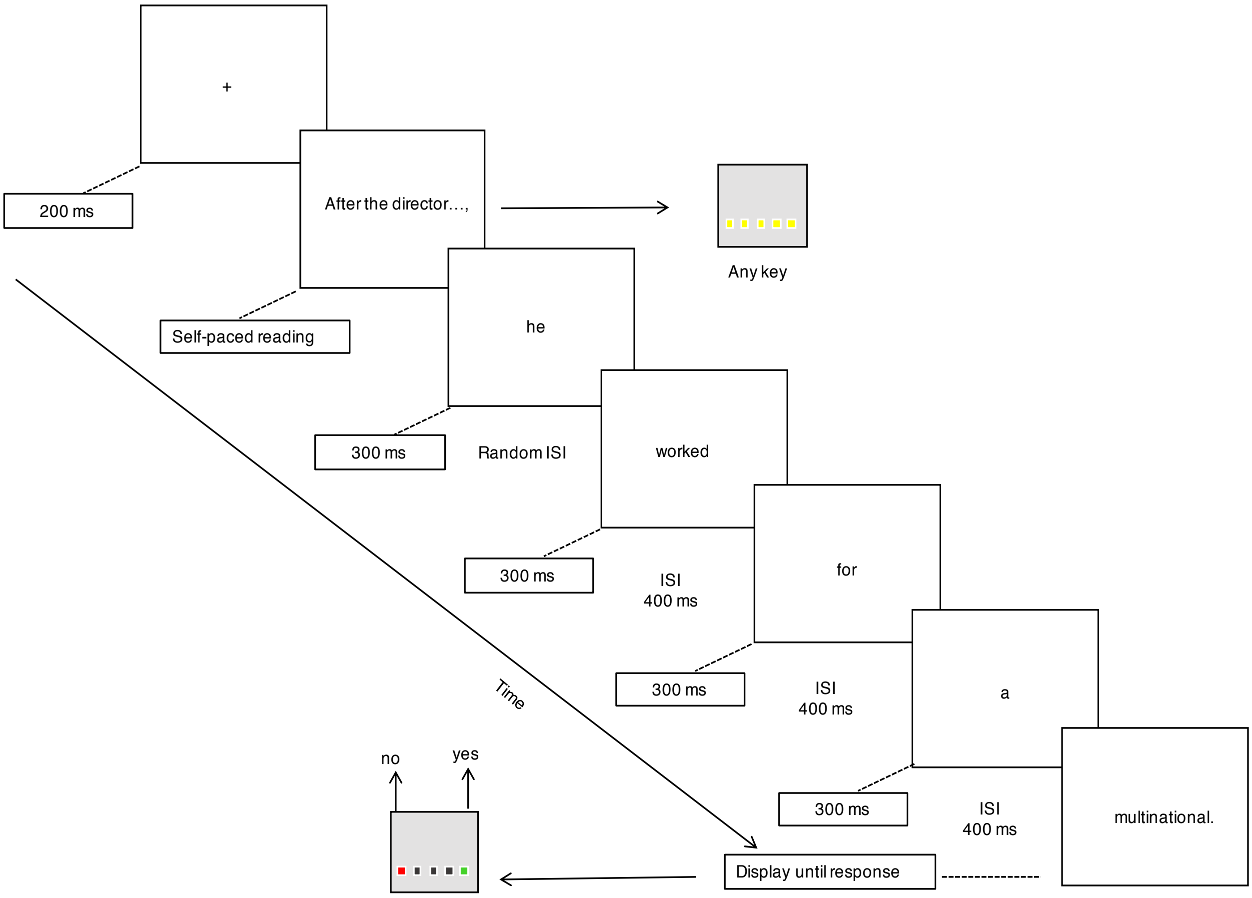

10
9
8
7
6
5
4
3
2
1
0
Chinese
English
Reading
Writing
Fluency
Speech ability

Figure 7


+
200 ms
After the director…,
Any key
Self-paced reading
he
300 ms
Random ISI
worked
300 ms
ISI
400 ms
for
300 ms
ISI
400 ms
a
Time
no
yes
300 ms
ISI
400 ms
multinational.
Display until response

Tables

Table 1. Examples of sentences in each of the four experimental conditions.

<table>
<tr><td>Correct</td><td>After the director of the school had resigned from the university, he worked for a multinational.</td><td rowspan="3">院长从大学辞职后,
他去了一家跨国公司工作

yuàn cháng cóng dà xué cí zhí hòu,
tā qù-le yī jiā kuà guó gōng sī gōng zuò</td></tr>
<tr><td>Present-Past Misalignment</td><td>*After the director of the school has resigned from the university, he worked for a multinational.</td></tr>
<tr><td>Future-Past Misalignment</td><td>*After the director of the school will have resigned from the university, he worked for a multinational.</td></tr>
<tr><td>Semantic violation</td><td>*After the director of the school had resigned from the university, he worked for a meter.</td><td>院长从大学辞职后,
他去了一家米工作

yuàn cháng cóng dà xué cí zhí hòu,
tā qù-le yī jiā mǐ gōng zuò</td></tr>
<tr><td>Filler incorrect</td><td>*After the director of the school had resigned from the university, he will work for a multinational.</td><td rowspan="3">院长从大学辞职后,
他将要去一家跨国公司工作

yuàn cháng cóng dà xué cí zhí hòu,
tā jiāng yào qù yī jiā kuà guó gōng sī gōng zuò</td></tr>
<tr><td>Filler correct</td><td>After the director of the school has resigned from the university, he will work for a multinational.</td></tr>
<tr><td>Filler correct</td><td>After the director of the school will have resigned from the university, he will work for a multinational.</td></tr>
<tr><td>Filler semantic violation</td><td>*After the director of the school had resigned from the university, he will work for a meter.</td><td>院长从大学辞职后,
他将要去一家米工作

yuàn cháng cóng dà xué cí zhí hòu,
tā jiāng yào qù yī jiā mǐ gōng zuò</td></tr>
</table>

Table 2. LMM analyses for Reference Time Misalignment behavioral data

Model 1 (English at int.)	**Accuracy**			**Reaction Time**			**Model 2 (Chinese at int.)**	**Accuracy**			**Reaction Time**		
	b	*SE*	*z*	*b*	*SE*	*t*		*b*	*SE*	*z*	*b*	*SE*	*t*
Int. English / Control	4.59	0.51	8.99***	6.40	0.09	71.17	**Int.** Chinese / Control	3.91	0.38	10.41***	6.37	0.08	80.54***
FF1. English / PPM	-8.64	0.74	-11.65***	-0.02	0.04	-0.52	**FF1.** Chinese / PPM	-8.07	0.62	-12.92***	0.09	0.04	2.36*
FF2. English / FPM	-8.27	0.74	-11.12***	0.03	0.04	-0.59	**FF2.** Chinese/ FPM	-7.33	0.59	-12.52***	0.03	0.04	0.9
FF3. Chinese / Control	-0.60	0.57	-1.05	-0.03	0.11	-0.27	**FF3.** English/ Control	0.78	0.58	1.34	0.02	0.11	0.27
I1. Chinese /PPM	0.46	0.86	0.54	0.11	0.06	2.00*	**I1.** English /PPM	-0.71	0.86	-0.83	-0.11	0.06	-2*
I2. Chinese /FPM	0.84	0.88	0.96	0.06	0.06	1.04	**I2.** English /FPM	-1.06	0.87	-1.22	-0.06	0.06	-1.04

* $p < .05$; ** $p < .01$; *** $p < .001$; FIXED EFFECTS (FF): Int. is the intercept (baseline) condition against which all other contrasts are compared. FF1 & FF2 consider only the data from the group used for intercept, and examine whether performance differs between the specified conditions and the intercept. FF3 examines differences between groups in the control condition. INTERACTION EFFECTS: I1 & I2 examine whether comparison group participants' performance on the PPM condition (I1) or FPM condition (I2) differs from their own performance in the control condition, and how the magnitude of this difference compares to that shown by the English group. Thus, whilst a significant interaction signals group differences as a function of condition, no significant interaction shows that both groups behaved similarly across conditions.

Table 3. LMM analyses for Semantic Violation behavioral data

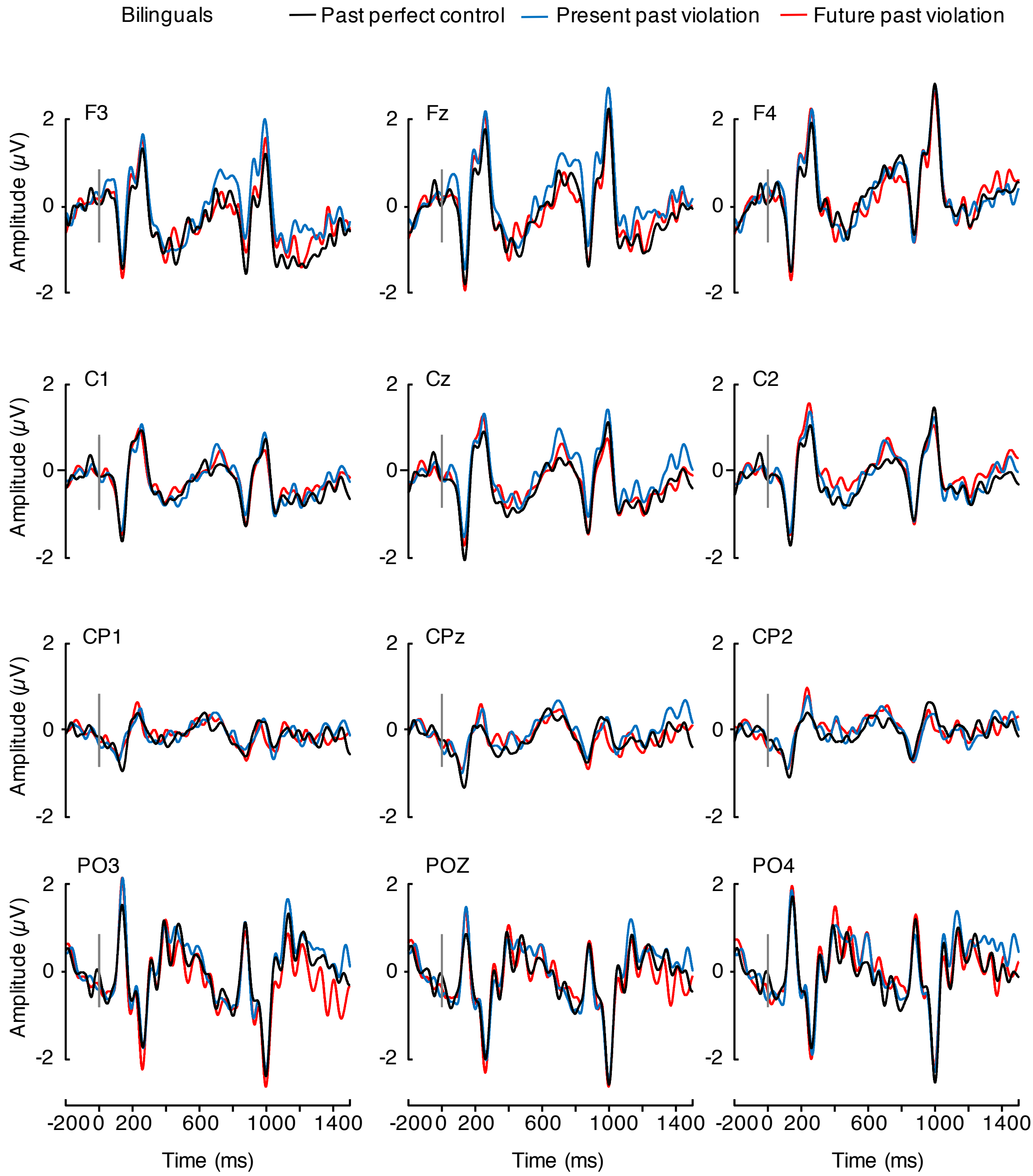
Model 1 (English at int.)	Accuracy			Reaction Time			Model 2 (Chinese at int.)	Accuracy			Reaction Time		
	b	*SE*	*z*	*b*	*SE*	*t*		*b*	*SE*	*z*	*b*	*SE*	*t*
Int. English / Control	4.45	0.52	8.52***	6.40	0.09	70.87***	**Int.** Chinese / Control	3.84	0.37	10.36***	6.37	0.08	80.35***
FF1. English / PSV	0.67	0.97	0.69	-0.03	0.06	-0.48	**FF1.** Chinese / PSV	1.11	0.73	1.53	0.20	0.06	3.18**
FF2. Chinese / Control	-0.63	0.59	-1.07	-0.03	0.11	-0.26	**FF2.** English/ Control	0.63	0.59	1.06	0.03	0.11	0.27
I1. Chinese /PSV	0.65	1.14	0.57	0.23	0.09	2.63**	**I1.** English /PSV	-0.04	1.33	-0.03	-0.23	0.08	-2.71**

* *$p < .01$; *** $p < .001$; FIXED EFFECTS (FF): Int. is the intercept (baseline) condition against which all other contrasts are compared. FF1 consider only English participants' data, and examine whether performance differs on the PSV vs. Control condition (FF1). FF2 examines differences between Chinese and English participants on the Past Perfect control condition. INTERACTION EFFECTS: I1 examine whether Chinese participants' performance on the PSV condition (I1) differs from their own performance on the control condition, and how the magnitude of this difference compares to that shown by the English group. Thus, whilst a significant interaction signals group differences as a function of condition, no significant interaction infers that both groups behave similarly across conditions.

Table 4. Chinese-English bilingual participants' language background

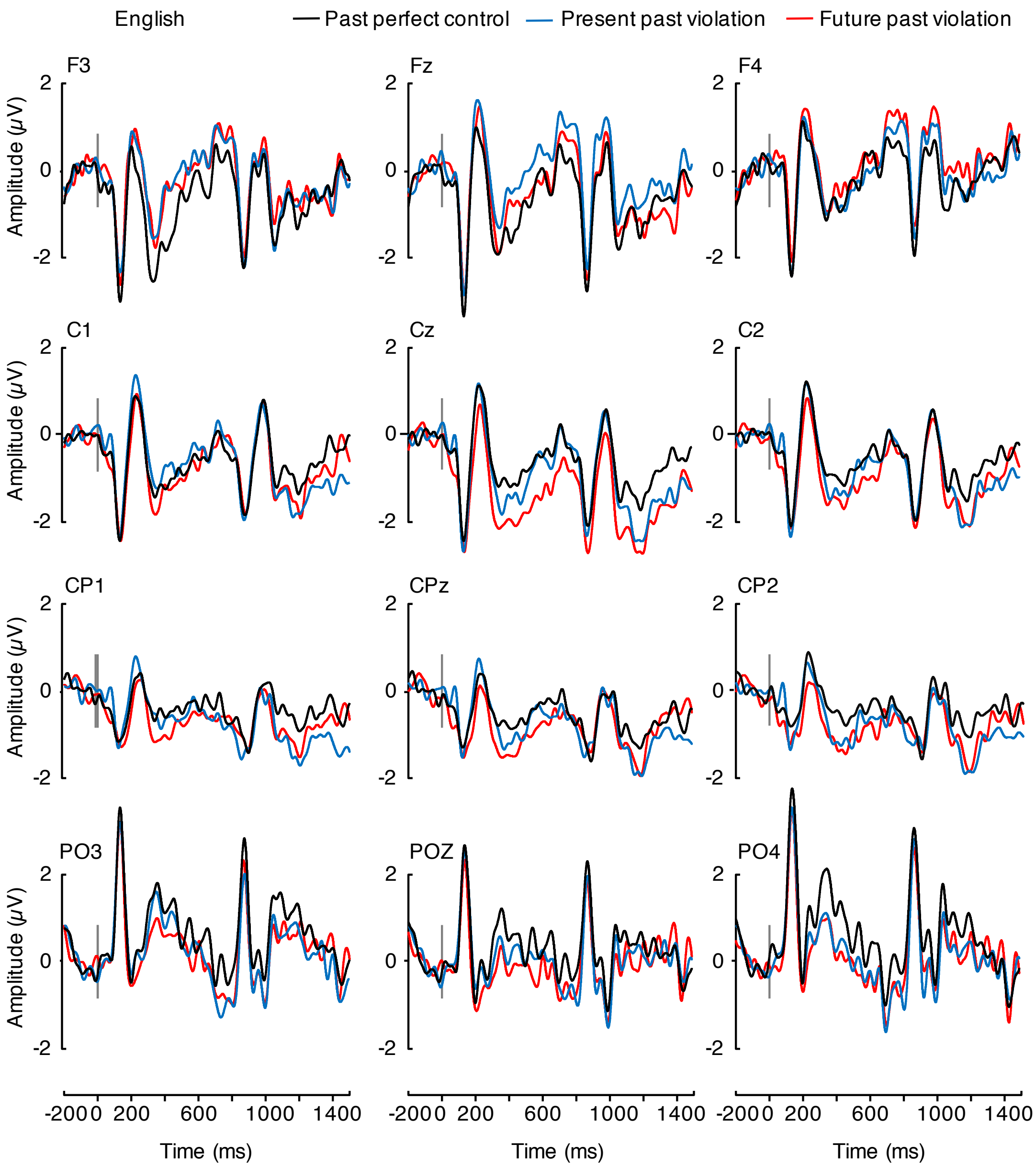
Measure	**Mean**	**(SD)**
Age of L2 acquisition	6.9	(3.2)
Length of L2 learning (years)	15.2	(3.5)
Length of staying in an English Speaking country (Months)	44.7	(61.4)
Daily Chinese usage (%)	50.7	(20.4)
Daily English usage (%)	49.3	(20.4)

Supplementary Figure 1



ERPs elicited by the critical verb and the post-critical word in the three experimental conditions in Chinese-English bilingual participants at 12 electrode sites.

Supplementary Figure 2



ERPs elicited by the critical verb and the post-critical word in the three experimental conditions in Native speakers of English at 12 electrode sites.

Supplementary Analyses

Split-group analyses. Yao and Chen (2016) investigated how late Chinese-English bilinguals are affected by cross-linguistic differences, using tense and progressive aspect violation tasks. They divided their participants into two groups based on proficiency in L2 English: (a) a high proficiency group including participants having obtained a score of 7 or above in the IELTS, having passed the TEM-8 test, or having achieved 100 or above in the TOEFL; and (b) a low proficiency group having passed the CET-4 test (equivalent to a score of 6 in the IELTS). They found that Chinese-English bilinguals with a score of 7 or above in the IELTS were able to detect tense violations.

Accordingly, we divided our group of 27 Chinese-English bilingual participants into two proficiency subgroups: (1) A high proficiency group including bilingual participants with an IELTS score of 7 or above (n = 14; Mean = 7.2),; (2) A lower proficiency group including bilingual participants with an IELTS score of 6.5 (n = 8) and we dismissed five participants who received instruction exclusively through the medium of English since high school and were thus high functioning bilinguals although we did not have a quantitative measurement of proficiency for them.

N400 mean amplitudes were analysed by means of repeated measures analyses of variance (ANOVA) with group (high / low proficiency) as between-subject factor and conditions (correct control, PPM, FPM) as within subject factor. There was no significant effect of group ($F(1,20) = 1.36$, $p = 0.26$, $\eta_p^2 = 0.06$) or condition ($F(2, 40) = 0.27$, $p = 0.76$, $\eta_p^2 = 0.01$) on N4-1 mean amplitude and the interaction between group and condition was not significant either ($F(2,40) = 2.37$, $p = 0.11$, $\eta_p^2 = 0.11$). There was no significant effect of group ($F(1,20) = 2.07$, $p = 0.17$, $\eta_p^2 = 0.09$) or condition ($F(2,40) = 0.2$, $p = 0.82$, $\eta_p^2 = 0.01$) on the mean amplitude of the N4-2 either and the interaction between group and condition was not significant ($F(2,40) = 1.53$, $p = 0.23$, $\eta_p^2 = 0.07$).

Correlation analyses. We found no significant correlations between participants' performance in an offline judgement task and N400 modulations in the N4-1 (PPM: $r = -0.18$, $p = 0.43$, $n = 22$; *FPM:* $r = -0.06$, $p = 0.78$, $n = 22$) or the N4-2 (PPM: $r = -0.31$, $p = 0.17$, $n = 22$; *FPM:* $r = -0.05$, $p = 0.83$, $n = 22$) analysis windows.

Additional proficiency test. In addition, we collected overt time alignment judgement accuracy using the most difficult misalignment condition PPM and the correct control condition in a new group of Chinese-English bilingual participants with proficiency closely matched to the group tested in the main experiment (n=21, $M_{\text{IELTS score}} = 6.86$, $SD = 0.5$, range 6.5–8).

These participants with a very similar proficiency to those tested in the ERP experiment did not significantly differ in accuracy from a control group of English native speakers tested on the same task, $F(1,37) = 0.95$, $p=0.34$ (Mean Accuracy PPM, English: 77%, Bilingual: 64%; Mean Accuracy control, English: 93%, Bilingual: 96%), making it very unlikely for the lack of differences between participants in the ERP experiment to hide an overall difference in proficiency.